

TESIS DOCTORAL

**Metodología de Caracterización Conceptual por
Condicionamientos Sucesivos. Una aplicación a
sistemas medioambientales.**

PRESENTADA POR: ALEJANDRA A. PÉREZ BONILLA

DIRIGIDA POR: KARINA GIBERT

Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

OCTUBRE DE 2009

AGRADECIMIENTOS

No quisiera llegar al final de este arduo y largo camino sin recordar y otorgar mi reconocimiento a todos aquellos y aquellas que en algún momento fueron parte de él. En primer lugar, al programa de becas para estudios de doctorado en el extranjero de la Comisión Nacional de Investigación Científica y Tecnológica *CONICYT* (Chile), “Beca Gobierno de Chile - Bid”, por haber depositado su confianza en mí y haber sido el sostén inestimable para salir y para llegar. A la doctora Karina Gibert, por haber dirigido esta tesis con agilidad y eficacia, por su dedicación y siempre pronta disposición.

Muy especialmente al Dr. Darko Vrecko, investigador del departamento de sistemas y control del Instituto de investigación Jozef Stefan; por haberme facilitado tan desinteresadamente la base de datos de la planta eslovena y por su apoyo en la interpretación de los *clusterings*; por su interés en mi trabajo y, sobre todo, por su gran hospitalidad durante mi estancia en Ljubljana. *Najlepsa hvala Darko!*

Al Laboratorio de Ingeniería Química y Ambiental de la Universidad de Girona, por la base de datos de la planta catalana y en especial a los Doctores Ignasi Rodríguez-Roda y Joaquim Comas por el conocimiento experto aportado en la validación de los resultados. También al proyecto de investigación GESCONDA de la Universidad Politécnica de Cataluña y a todos los miembros de la *Tribu KLASS*.

A los revisores externos por los comentarios enriquecedores realizados a esta tesis.

No podría olvidar a todos mis alumnos catalanes de estos últimos 4 años, gracias por las anécdotas compartidas y por valorar siempre lo que os entregaba.

Sense la confiança, suport i omnipresent afecte dels meus amics “d'aquest costat”; Nuria, Cristina i Marc, no podría de cap manera arribar al final del camí.

En momentos vitales María Teresa, María Alejandra, Fabiola, Freddie, Norma, Francisca, Evelyn, Javier, Marly, Miguel y Pablo; con la preciosista cadencia de sus palabras me regalaron la fascinación de sentirme, junto a ellos, mis amigos “del otro lado”.

La desprendida emoción de Andrea alimentó el valor necesario para dar el último paso. *Danke sher Andrea!*

No podría olvidar a mi familia chilena en Barcelona, a los que están y a los que pasaron, a todos aquellos que un día aparecieron en mi vida para quedarse para siempre. En especial a Graciela, por sus oídos maternales; a Luis por su cariñosa honestidad y a Pablo, por nunca dejar de quererme. A todos los compañeros y amigos latinoamericanos, unidos por un pasado común, siempre dispuestos, siempre generosos, siempre presentes.

Difícil resulta encontrar las palabras justas para expresar todo lo que aprendí en esta experiencia, a nivel personal y profesional. En definitiva a todos los que desinteresadamente han sido parte de esto.

A Alicia, mi madre
Por empujarme del nido ...
y enseñarme a volar.
Por estar ahí ...
cuando mis alas se cansaron.
Por dejarme creer ...
que todo es posible.
Y por hacer que mi vuelo ...
fuese siempre libre.

A Jaime, mi padre
Por mostrarme desde el cielo...
la luz al final del camino.

Gracias por creer en mi.

Abstract

In automatic classification we seek profiles underlying the structure of a domain for better understanding and providing support to decision making. Therefore understanding the meaning of classes is essential. On the other hand the validation of a *cluster* still is an open problem as an objective criteria for determining the quality of a set of classes has not yet been found in the context of the clustering (Hand 1996), which applies in situations where there is not a good, knowledge of the structure of the domain. In 1985 Volle illustrates that the concept of validity is not absolute, but relative to the conditions of the context and the usefulness of the clusters. Although it is difficult to objectivate interpretation, constitutes a fundamental phase of the process and remains still today one of the most commonly used criteria to validate the cluster. For this reason the validation is directly linked to a clear interpretation for clustering or partition.

Thus, it is now necessary to introduce tools to assist the user in the task of interpreting a partition on a set of objects in order to establish the meaning of the resulting classes. If the classes obtained don't make sense to the experts, the results of the classification are not considered valid, nor could be used, or support any subsequent decision. All validation techniques and algorithms focus on the structure of the partition, but having classes well-structured doesn't guarantee that an expert will be able to associate each of these groups with a semantic entity.

This thesis wants to make a contribution to this process, fundamental for understanding the meaning of the obtained classes and to give effective support to the subsequent decision-making.

The alternative which seem to be most promising to solve these limitations is to lighten the work of the expert, by developing techniques based on empirical evidence to identify the most important variables and formulate concepts that express the specifici of each class and are expressed in a conceptual representation able for automatic generation and directly understandable to the expert.

To incorporate procedures that translate the results of analysis (in this case of clustering) to a explicit knowledge representation is in line with what Fayyad in 1996 suggests for systems of *Knowledge Discovery from Data (KDD)* where the phase of post-process of the results to generate knowledge is almost as important as the analysis itself. Perhaps due to its semantic nature the automatic generation of interpretations of a classification has not been formally treated by statistics, but to resolve it is essential.

This thesis proposes an approximate solution to the problem of building a system of concepts $\mathcal{A}_{\mathcal{P}_\xi} = \{A_1, A_2, A_3, \dots, A_\xi\}$ describing the classes so that, given a partition in ξ classes, $\mathcal{P}_\xi = \{C_1, \dots, C_\xi\}$, over a set object $i \in \mathcal{I}$:

- $A, A' \in \mathcal{A}_{\mathcal{P}_\xi} \Rightarrow A \neq A'$
- $\forall i \in \mathcal{I}, \quad A_C(i) = True, \text{ if } C = C(i, \mathcal{P}_\xi), \quad A_C \in \mathcal{A}_{\mathcal{P}_\xi}$
- $\forall i \in \mathcal{I}, \quad A_C(i) = false, \text{ if } C \neq C(i, \mathcal{P}_\xi), \quad A_C \in \mathcal{A}_{\mathcal{P}_\xi}$

on $A_C(i)$ is a concept $\in \mathcal{A}_{\mathcal{P}_\xi}$, is true if object i satisfies the concept $A_C(i)$.

Considering that there is some uncertainty in the model, it is suggested to use more general rules of the type $r : A_C(i) \xrightarrow{p} C$ where $p \in [0, 1]$ is the probability of satisfying r . Thus the rules incorporate uncertainty in a probabilistic approach.

The methodology proposed tries to approximate in a formal model the natural process that follows an expert in its phase of interpretation of results by making an iterative approximation based on a hierarchical clustering. This methodology:

- Provides a systematizing of the process of interpretation of classes from a hierarchical cluster and represents a significant advance to the current situation in which the interpretation is done manually and more or less crafted.
- Likewise, it helps to systematize and objectify the mechanisms of interpretation used by human experts.
- The results generated by the methodology allow the expert to more easily understand the main characteristics of the classification obtained by generating explicit knowledge directly from the classes.

While the methodology proposed is general, application focuses on WasteWater Treatment Plant (WWTP) because this is one of the domains where conventional approaches work worse and belong to the lines under research developed in the group.

From a theoretical point of view, the main focus of this thesis has been to present a hybrid methodology that combines tools and techniques of statistics and Artificial Intelligence in a cooperative way, using a transversal and multidisciplinary approach combining elements of the induction of concepts from Artificial Intelligence, propositional logic and probability theory. Thus, this thesis contributes to the generic design of *KDD system*, which should include modules that support the definition of the problem (including knowledge), data collection, cleaning and preprocessing, data reduction, selection of data mining technique, interpretation and production of a posteriori discovered knowledge. It also contributes to objectivate procedures for the validation of results, as the fact that clustering has a clear interpretation is related to the usefulness of a classification (currently used as a criterion for validation) and usable to decide whether it is correct or not, evaluating the usefulness requires an posteriori mechanism of understanding the meaning of classes.

The methodology of Conceptual Characterization by Embedded Conditioning (*CCEC*) benefits from the hierarchical structure of the target classification to induce concepts iterating with binary divisions from dendrogram, so that, based on the variables that describe the objects belonging to a certain domain, the specifics of each class can be found, thus contributing to the automatic interpretation of conceptual description of clustering.

Finally an approximate solution of the initial problem has been operatively built: $\mathcal{A}_{\mathcal{P}_\xi} = \{C : \mathcal{A}_C \ \forall C \in \mathcal{P}_\xi\}$, where \mathcal{A}_C are concepts that allow to understand and distinguish the classes from a hierarchical construction of a set of rules $\mathcal{R}(\mathcal{P}_\xi) = \{r \ tq \ r : A \xrightarrow{p(r)} C \ \forall C \in \mathcal{P}_\xi\}$.

Resumen

En clasificación automática se buscan perfiles subyacentes a la estructura de un dominio que ayude a comprenderlo y permita una mejor toma de decisiones. Por ello comprender el significado de las clases resulta fundamental. Por otro lado la validación de un *cluster* sigue siendo un problema abierto por no haberse encontrado aún un criterio objetivo para determinar la calidad de un conjunto de clases en el contexto del *clustering* (Hand 1996), que se aplica en situaciones en las que no hay un buen conocimiento de la estructura del dominio. Volle en 1985 hace toda una disertación ilustrando que el concepto de validez no es absoluto, sino relativo a las condiciones del contexto y a la utilidad de las mismas. A pesar de que este es un extremo poco objetivable, la interpretación se convierte así en una fase fundamental del proceso y sigue siendo, aún hoy, uno de los criterios más utilizados en la práctica para validar el *cluster*. Por esta razón la validación queda directamente ligada a la existencia de una interpretación clara para el *clustering* o partición.

Así, actualmente es necesario introducir herramientas para asistir al usuario en las tareas de interpretación de una partición sobre un conjunto de objetos, con el fin de establecer el significado de las clases resultantes. Si las clases obtenidas no tienen sentido para el/los expertos, los resultados de la clasificación no son considerados válidos, ni tampoco se podrán utilizar, ni darán apoyo a ninguna decisión posterior. Todas las técnicas y algoritmos de validación van orientados a la vertiente estructural de la partición, pero disponer de clases bien formadas estructuralmente no ofrece garantía de que un experto vaya a ser capaz de asociar cada uno de esos grupos a una entidad semántica.

Esta tesis pretende contribuir a la mejora de este proceso, fundamental para comprender el significado de las clases obtenidas y dar soporte efectivo a la posterior toma de decisiones.

La alternativa que parece más prometedora para resolver estas limitaciones es aligerar al experto de este trabajo, mediante el desarrollo de técnicas que a partir de la evidencia empírica, identifiquen las variables más relevantes y formulen conceptos que expresen las particularidades de cada clase y se expresen en una forma de representación conceptual generable automáticamente y directamente comprensible para el experto.

Incorporar procedimientos que trasladen los resultados del análisis (en este caso del *clustering*) a una representación explícita del conocimiento obtenido, se sitúa en la línea de lo que Fayyad en 1996 propone para los sistemas de *Knowledge Discovery from Data (KDD)*, donde la fase de post-proceso de los resultados para generar conocimiento es casi tan importante como el análisis en si mismo. Quizás por su naturaleza más semántica la generación automática de interpretaciones de una clasificación no se ha tratado formalmente desde el ámbito estadístico, aunque resolverlo es fundamental.

En esta tesis se propone una solución aproximada al problema planteado de construir un sistema de conceptos $\mathcal{A}_{\mathcal{P}_\xi} = \{A_1, A_2, A_3, \dots, A_\xi\}$ que describen las clases de tal forma que dada una partición en ξ clases, $\mathcal{P}_\xi = \{C_1, \dots, C_\xi\}$, sobre un conjunto de objetos $i \in \mathcal{I}$:

- $A, A' \in \mathcal{A}_{\mathcal{P}_\xi} \Rightarrow A \neq A'$
- $\forall i \in \mathcal{I}, \quad A_C(i) = \text{verdadero}, \text{ si } C = C(i, \mathcal{P}_\xi), \quad A_C \in \mathcal{A}_{\mathcal{P}_\xi}$

- $\forall i \in \mathcal{I}, A_C(i) = \text{falso}, \text{ si } C \neq C(i, \mathcal{P}_\xi), A_C \in \mathcal{A}_{\mathcal{P}_\xi}$

donde $A_C(i)$ es un concepto tal que $A_C(i) \in \mathcal{A}_{\mathcal{P}_\xi}$, y es verdadero si el objeto i evaluado sobre $A_C(i)$, lo satisface.

Teniendo en cuenta que existirá cierta incertezza en el modelo, se propone tratar con reglas más genéricas de la forma $r : A_C(i) \xrightarrow{p} C$ donde $p \in [0, 1]$ es la probabilidad con que se cumple r . De este modo las reglas incorporan incertezza bajo una aproximación probabilística.

La metodología que se propone trata de aproximar en un modelo formal el proceso natural que sigue un experto en su fase de interpretación de resultados realizando una aproximación iterativa basada en la clasificación jerárquica. La propuesta que se presenta:

- Aporta una sistematización al proceso de interpretación de clases procedentes de un *cluster* jerárquico y supone un avance significativo respecto al estado actual en que la interpretación se realiza de forma manual y más o menos artesanal.
- Asimismo, contribuye a sistematizar y objetivar los mecanismos de interpretación que usan los expertos humanos.
- Los resultados que genera la metodología permiten que el experto pueda comprender más fácilmente las características principales de la clasificación obtenida ya que genera conocimiento explícito directamente a partir de las clases.

Si bien la metodología que se propone es general, se ha centrado la aplicación a Estaciones depuradoras de aguas residuales (*EDAR*) por ser éste uno de los dominios donde las aproximaciones clásicas funcionan peor y porque se encuadran en una de las líneas marco de investigación que se desarrolla en el grupo.

Desde un punto de vista teórico, el interés de esta tesis ha sido presentar una propuesta metodológica híbrida que combine herramientas y técnicas de Estadística e Inteligencia Artificial en forma cooperativa, siguiendo un enfoque transversal y multidisciplinar combinando elementos de la inducción de conceptos en Inteligencia Artificial, lógica proposicional y teoría de probabilidad. Es así como, ésta tesis, contribuye a la concepción genérica de sistema de *KDD*, que debe incluir módulos de soporte a la definición del problema (que incluya el conocimiento), recolección de datos, depuración y preproceso, reducción de datos, selección de la técnica de *data mining*, interpretación y producción del conocimiento descubierto a posteriori. También contribuye a objetivar los procedimientos de validación de resultados, ya que el hecho de que un *clustering* tenga una interpretación clara está relacionado con la utilidad de una clasificación (que actualmente se utiliza también como criterio de validación) y se puede utilizar para decidir si es correcta o no; evaluarla requiere un mecanismo a posteriori de comprensión del significado de las clases.

La metodología de Caracterización Conceptual por Condicionamientos Sucesivos (*CCCS*) aprovecha la estructura jerárquica de la clasificación objetivo para inducir conceptos iterando sobre las divisiones binarias que indica el dendrograma, de tal forma que, a partir de las variables que describen los objetos pertenecientes a cierto dominio, se puedan encontrar las particularidades de cada clase, contribuyendo así al proceso de interpretación conceptual automática de clases procedentes de un *cluster*.

Finalmente se ha construido operativamente una solución aproximada al problema inicial de la forma $\mathcal{A}_{\mathcal{P}_\xi} = \{C : \mathcal{A}_C \ \forall C \in \mathcal{P}_\xi\}$, donde \mathcal{A}_C son conceptos que permiten entender y distinguir las clases, a partir de una construcción jerárquica de un conjunto de reglas $\mathcal{R}(\mathcal{P}_\xi) = \{r \ tq \ r : A \xrightarrow{p(r)} C \ \forall C \in \mathcal{P}_\xi\}$.

Resum

En classificació automàtica es busquen perfils subjacents a l'estructura d'un domini que ajudi a comprendre'l i permeti una millor presa de decisions. Per això comprendre el significat de les classes resulta fonamental. D'altra banda la validació d'un *cluster* segueix sent un problema obert per no haver-se trobat encara un criteri objectiu per a determinar la qualitat d'un conjunt de classes en el context del *clustering* (Hand 1996), que s'aplica en situacions en les quals no hi ha un bon coneixement de l'estructura del domini. Volle al 1985 fa tota una dissertació il·lustrant que el concepte de validesa no és absolut, sinó relatiu a les condicions del context i a la utilitat de les mateixes. Tot i que aquest és un extrem poc objetivable, la interpretació es converteix així en una fase fonamental del procés i segueix sent, encara avui, un dels criteris més utilitzats en la pràctica per a validar el *cluster*. Per aquesta raó la validació queda directament lligada a l'existència d'una interpretació clara per al *clustering* o partició.

Així, actualment és necessari introduir eines per a assistir a l'usuari en les tasques d'interpretació d'una partició sobre un conjunt d'objectes, amb la finalitat d'establir el significat de les classes resultants. Si les classes obtingudes no tenen sentit per a l'expert(s), els resultats de la classificació no són considerats vàlids, ni tampoc es podran utilitzar, ni donaran suport a cap decisió posterior. Totes les tècniques i algoritmes de validació van orientats al vessant estructural de la partició, però disposar de classes ben formades estructuralment, no oferix garantia que un expert sigui capaç d'associar cadascun d'aquests grups a una entitat semàntica.

Aquesta tesi pretén contribuir a la millora d'aquest procés, fonamental per a comprendre el significat de les classes obtingudes i donar suport efectiu a la posterior presa de decisions.

L'alternativa que sembla més prometedora per resoldre aquestes limitacions és alleugerir a l'expert d'aquest treball, mitjançant el desenvolupament de tècniques que a partir de l'evidència empírica, identifiquin les variables més rellevants i formulin conceptes que expressin les particularitats de cada classe i s'expressin en una forma de representació conceptual generable automàticament i directament comprensible per a l'expert.

Incorporar procediments que traslladin els resultats de l'anàlisi (en aquest cas del *clustering*) a una representació explícita del coneixement obtingut, se situa en la línia del que Fayyad al 1996 proposa per als sistemes de *Knowledge Discovery from Data (KDD)*, on la fase de post-procés dels resultats per generar coneixement és gairebé tan important com l'anàlisi en si mateix. Potser per la seva naturalesa més semàntica la generació automàtica d'interpretacions d'una classificació no s'ha tractat formalment des de l'àmbit estadístic, encara que resoldre'l és fonamental.

En aquesta tesi es proposa una solució aproximada al problema plantejat de construir un sistema de conceptes $\mathcal{A}_{\mathcal{P}_\xi} = \{A_1, A_2, A_3, \dots, A_\xi\}$ que descriuen les classes de tal manera que, donada una partició en ξ classes, $\mathcal{P}_\xi = \{C_1, \dots, C_\xi\}$, sobre un conjunt d'objectes $i \in \mathcal{I}$:

- $A, A' \in \mathcal{A}_{\mathcal{P}_\xi} \Rightarrow A \neq A'$
- $\forall i \in \mathcal{I}, \quad A_C(i) = \text{veritable, si } C = C(i, \mathcal{P}_\xi), \quad A_C \in \mathcal{A}_{\mathcal{P}_\xi}$

- $\forall i \in \mathcal{I}, A_C(i) = \text{false}$, si $C \neq C(i, \mathcal{P}_\xi)$, $A_C \in \mathcal{A}_{\mathcal{P}_\xi}$

On $A_C(i)$ és un concepte tal que $A_C(i) \in \mathcal{A}_{\mathcal{P}_\xi}$, i és veritable si l'objecte i evaluat sobre $A_C(i)$, ho satisfà.

Tenint en compte que existirà certa incertesa en el model, es proposa tractar amb regles més genèriques de la forma $r : A_C(i) \xrightarrow{p} C$ on $p \in [0, 1]$ és la probabilitat amb que es compleix r . D'aquesta manera les regles incorporen incertesa sota una aproximació probabilística.

La metodologia que es proposa tracta d'aproximar en un model formal el procés natural que segueix un expert en la seva fase d'interpretació de resultats realitzant una aproximació iterativa basada en la classificació jeràrquica. La proposta que es presenta:

- Aporta una sistematització al procés d'interpretació de classes procedents d'un *cluster* jeràrquic i suposa un avanç significatiu respecte a l'estat actual que la interpretació es realitza de forma manual i més o menys artesanal.
- Així mateix, contribueixi a sistematitzar i objectivar els mecanismes d'interpretació que usen els experts humans.
- Els resultats que genera la metodologia permeten que l'expert pugui comprendre més fàcilment les característiques principals de la classificació obtinguda ja que genera coneixement explícit directament a partir de les classes.

Si bé la metodologia que es proposa és general, s'ha centrat l'aplicació a Estacions depuradores d'aigües residuals (*EDAR*) per ser aquest un dels dominis on les aproximacions clàssiques funcionen pitjor i perquè s'enquadren en una de les línies marc d'investigació que es desenvolupa en el grup.

Des d'un punt de vista teòric, l'interès d'aquesta tesi ha estat presentar una proposta metodològica híbrida que combini eines i tècniques d'Estadística i d'Intel·ligència Artificial en forma cooperativa, seguint un enfocament transversal i multidisciplinar combinant elements de la inducció de conceptes en Intel·ligència Artificial, lògica proposicional i teoria de probabilitat. És així com, aquesta tesi, contribueix a la concepció genèrica de sistema de *KDD*, que ha d'incloure mòduls de suport a la definició del problema (que inclogui el coneixement), recollida de dades, depuració i preprocés, reducció de dades, selecció de la tècnica de *data mining*, interpretació i producció del coneixement descobert a posteriori. També contribueix a objectivar els procediments de validació de resultats, ja que el fet que un *clustering* tingui una interpretació clara està relacionat amb la utilitat d'una classificació (que actualment també es fa servir com criteri de validació) i es pot emprar per a decidir si és correcta o no; avaluar-la requereix un mecanisme a posteriori de comprensió del significat de les classes.

La metodologia de Caracterització Conceptual per Condicionaments Successius (*CCCS*) aprofita l'estructura jeràrquica de la classificació objectiu per a induir conceptes iterant sobre les divisions binàries que indica el dendrogram, de tal manera que, a partir de les variables que descriuen els objectes pertanyents a cert domini, es puguin trobar les particularitats de cada classe, contribuint així al procés d'interpretació conceptual automàtica de classes procedents d'un *cluster*.

Finalment s'ha construït operativament una solució aproximada al problema inicial de la forma $\mathcal{A}_{\mathcal{P}_\xi} = \{C : \mathcal{A}_C \ \forall C \in \mathcal{P}_\xi\}$, on \mathcal{A}_C són conceptes que permeten entendre i distingir les classes, a partir d'una construcció jeràrquica d'un conjunt de regles $\mathcal{R}(\mathcal{P}_\xi) = \{r \text{ tq } r : A \xrightarrow{p(r)} C \ \forall C \in \mathcal{P}_\xi\}$.

Indice de Contenidos

Agradecimientos	i
Dedicatoria	iii
Abstract	v
Resumen	vii
Resum	ix
I Introducción y Fundamentos Teóricos	1
1 Introducción y motivación	3
1.1 Formulación del Problema de tesis	6
1.2 Descripción del Proyecto Marco	8
1.3 Estructura del documento	9
2 Objetivos	11
3 Estado del Arte	13
3.1 Introducción	13
3.2 Sistemas de soporte a la toma de decisiones	13
3.3 Knowledge Discovery from Data	17
3.4 Inteligencia Artificial y Estadística (<i>AI& Stats</i>)	21
3.4.1 Estadística	21
3.4.2 Inteligencia Artificial	23
3.5 Clustering	26
3.5.1 Clustering en la estadística	28
3.5.2 Clustering en la Inteligencia Artificial	30
3.5.3 Clustering en AI& Stats	33
3.5.4 Validación de un clustering	36
4 Antecedentes	43
4.1 Antecedentes	43
5 Conceptos previos	45
5.1 Notación general	45
5.2 Algoritmo genérico de clasificación ascendente jerárquica	46
5.3 Dendrograma	47
5.4 Índice de nivel de la jerarquía	49

5.5	Jerarquías indexadas (τ) y ultramétricas	49
5.6	Boxplot simple	50
5.7	Boxplot múltiple	50
5.8	Variables y valores caracterizadores de una clase	51
5.8.1	Alcance	52
5.9	Boxplot based Discretization (BbD)	53
5.10	Boxplot based Induction Rules (BbIR)	56
5.11	Uso de los condicionamientos sucesivos en la interpretación de clases	59
5.12	Lógica proposicional	59
II	Marco Teórico y Desarrollo de la Metodología	63
6	Introducción	65
7	Marco teórico de la metodología	67
7.1	Tipificación de los sistemas de reglas inducidos por el BbIR	67
7.1.1	Sistema de reglas completo $\mathcal{R}(X_k, \mathcal{P}_\xi)$	68
7.1.2	Sistema de reglas completo global, $\mathcal{R}(\mathcal{P}_\xi)$	68
7.1.3	Sistema de reglas reducido, $\mathcal{R}^*(X_k, \mathcal{P}_\xi)$	68
7.1.4	Sistema de reglas reducido global, $\mathcal{R}^*(\mathcal{P}_\xi)$	68
7.1.5	Sistemas de reglas de nivel η	69
7.1.6	Sistema de reglas efectivas $\mathcal{Re}(X_k, \mathcal{P}_\xi)$	69
7.1.7	Sistema de reglas efectivas global, $\mathcal{Re}(\mathcal{P}_\xi)$	69
7.1.8	Sistema de reglas efectivas reducido $\mathcal{Re}^*(X_k, \mathcal{P}_\xi)$	69
7.1.9	Sistema de reglas efectivas reducido global, $\mathcal{Re}^*(\mathcal{P}_\xi)$	70
7.1.10	Sistema de reglas Seguras $\mathcal{S}(X_k, \mathcal{P}_\xi)$	70
7.1.11	Sistema de reglas seguras global, $\mathcal{S}(\mathcal{P}_\xi)$	70
7.2	Tipos de valores	70
7.3	Cardinales y propiedades de los Sistemas de reglas	73
7.3.1	Relación entre los tipos de valores de una variable y los tipos de reglas que genera	74
7.3.2	Variables procedentes de la discretización por el <i>BbD</i> de una variable continua	76
7.4	Particiones binarias	77
7.5	Revisión del Boxplot based Discretization (BbD)	79
7.5.1	Análisis de los sistemas de intervalos inducidos a partir de una partición en 2 clases. Análisis por caso.	80
7.5.2	Propuesta de corrección	82
7.5.3	Propiedades	88
7.5.4	Comparación de las propuestas	89
7.6	Resumen del capítulo	93
8	Ventajas de la jerarquía indexada	97
8.1	La hipótesis de mundo cerrado y la forma de f	100
8.2	Propiedades	101

9 Criterios de selección de reglas	103
9.1 Criterios de evaluación para una única regla	103
9.1.1 Soporte ($Sup(r)$)	103
9.1.2 Grado de certeza ($p(r)$)	103
9.1.3 Cobertura relativa ($CovR(r)$)	104
9.2 Criterios para la evaluación de un sistema de reglas	105
9.2.1 Soporte total ($Sup_T(\mathcal{R})$)	105
9.2.2 Certeza o confianza media ($\bar{p}(\mathcal{R})$)	105
9.2.3 Cobertura global ($CovG_{lobal}(\mathcal{R})$)	105
9.2.4 Evaluación de sistemas de reglas frente a una partición de referencia .	106
9.3 Resumen del capítulo	107
10 Integración del conocimiento	109
10.1 Introducción	109
10.2 Best Global concept and Close-World Assumption	110
10.3 Best local concept and no Close-World Assumption	112
10.4 Best local concept and Close-World Assumption	113
10.5 Best local concept and partial Close-World Assumption	115
10.6 Best local-global concept and Close-World Assumption	116
10.7 Conclusión	118
11 Metodología CCCS	121
11.1 Introducción	121
11.2 Propuesta metodológica	122
III Aplicación y Resultados	127
12 Introducción	129
12.1 Acerca de las bases de datos	129
12.2 Acerca del Clustering jerárquico	130
12.3 Acerca de la aplicación de la metodología CCCS	131
13 Estaciones depuradoras de aguas residuales	133
13.1 Tratamiento de aguas residuales	133
13.1.1 El agua residual: composición	134
13.1.2 El problema de las aguas residuales	135
13.1.3 Descripción general del proceso de depuración	136
14 Caso de Estudio 1: Planta catalana	139
14.1 Descripción general	139
14.2 Seguimiento de la Planta	141
14.3 Control de la Planta	142
14.4 Legislación	142
14.5 Presentación de los datos	143
14.6 Descripción de las variables	143

15 Análisis descriptivo	147
15.1 Introducción	147
15.2 Análisis univariante	147
15.2.1 Q-E. Caudal de entrada (m ³ /d) Inflow wastewater	148
15.3 Análisis Bivariante	149
15.3.1 Sólidos en suspensión (SS)	149
16 Clustering planta catalana	151
16.1 Introducción	151
16.2 Base de conocimiento para la clasificación basada en reglas	151
16.3 Clustering	152
16.3.1 Familia de ficheros de clasificación	152
16.3.2 Clasificación basada en reglas	152
16.3.3 Secuencia de particiones	153
16.4 Interpretación validada por el experto	160
17 Aplicación, planta catalana	165
17.1 Interpretación de \mathcal{P}_4 con BG &CWA	165
17.1.1 Interpretación final:	175
17.1.2 Evaluación de la propuesta	176
17.2 Interpretación de \mathcal{P}_4 con BL+G &CWA	178
17.2.1 Interpretación final:	190
17.2.2 Evaluación de la propuesta	191
18 Análisis y resultados	193
18.1 Interpretación validada por el experto	193
18.2 Interpretaciones generadas por cada propuesta	194
18.2.1 Best global concept and Close-World Assumption:	194
18.2.2 Best local-global concept and Close-World Assumption:	194
18.3 Análisis de los resultados	195
18.3.1 Análisis cualitativo	195
18.3.2 Análisis cuantitativo	197
18.4 Conclusiones de la aplicación	199
18.5 Resumen	199
19 Caso de Estudio 2: Planta eslovena	201
19.1 Descripción general	201
19.2 Descripción detallada del sistema	203
19.2.1 Sistema de alcantarillas	203
19.2.2 Características de la EDAR	204
19.2.3 Estructuras de control actualmente disponibles	207
19.3 Planta piloto considerada en SMAC	209
19.3.1 Variables del proceso manipuladas	209
19.3.2 Configuración de la planta piloto	210
19.3.3 Experimentación con el “DO set-point control”	212
19.3.4 Inhibición del caudal de entrada a la planta piloto	214
19.3.5 Datos actualmente disponibles en linea para la planta piloto	214
19.3.6 Objetivos del SMAC	214
19.3.7 Legislación	214
19.3.8 Presentación de los datos	217

19.3.9 Descripción de las variables	217
20 Análisis descriptivo	219
20.1 Análisis univariante y bivariante	219
20.1.1 Introducción	219
20.1.2 Análisis univariante	219
20.1.3 Análisis bivariante	221
20.2 Time series plot	222
21 Clustering planta eslovena	225
21.1 Introducción	225
21.2 Base de conocimiento para la clasificación basada en reglas	225
21.3 Clustering	226
21.3.1 Familia de ficheros de clasificación	226
21.3.2 Clasificación basada en reglas	227
21.3.3 Secuencia de particiones	228
21.4 Interpretación validada por el experto	233
22 Aplicación, planta eslovena	237
22.1 Interpretación de \mathcal{P}_4 con BG &CWA	237
22.1.1 Interpretación final:	243
22.1.2 Evaluación de la propuesta	243
22.2 Interpretación de \mathcal{P}_4 con BL & _{no} CWA	245
22.2.1 Interpretación final:	252
22.2.2 Evaluación de la propuesta	252
22.3 Interpretación de \mathcal{P}_4 con BL &CWA	254
22.3.1 Interpretación final:	262
22.3.2 Evaluación de la propuesta	262
22.4 Interpretación de \mathcal{P}_4 con BL & _{partial} CWA	264
22.4.1 Interpretación final	272
22.4.2 Evaluación de la propuesta	272
22.5 Interpretación de \mathcal{P}_4 con BL+G &CWA	274
22.5.1 Interpretación final	282
22.5.2 Evaluación de la propuesta	282
23 Análisis y resultados	285
23.1 Interpretación validada por el experto	285
23.2 Interpretaciones generadas por cada propuesta	285
23.2.1 Best global concept and Close-World Assumption:	286
23.2.2 Best local concept and no Close-World Assumption:	286
23.2.3 Best local concept and Close-World Assumption:	286
23.2.4 Best local concept and partial Close-World Assumption:	286
23.2.5 Best local-global concept and Close-World Assumption:	287
23.3 Análisis de los resultados	287
23.3.1 Análisis cualitativo	287
23.3.2 Análisis cuantitativo	288
23.4 Conclusiones de la aplicación	291
23.5 Resumen	292

IV Conclusiones y líneas futuras	293
24 Conclusiones, trabajo futuro y líneas abiertas	295
Bibliografía	313
V Anexos	327
A KLASS	329
A.1 Introducción a KLASS	329
A.1.1 Vecinos recíprocos encadenados	330
A.1.2 Criterio de Agregación	331
A.1.3 El representante de la clase	332
A.1.4 Objetos compactos y objetos extendidos	332
A.2 Evolución de la plataforma KLASS	333
A.2.1 Satélites de KLASS	336
A.3 Aplicaciones Clustering basado en reglas	336
B Análisis de casos, BbD revisado	339
B.1 Introducción	339
B.2 Análisis por caso	339
B.2.1 Caso 2, $M_{C_i}^k < m_{C_j}^k$	339
B.2.2 Caso 3, $m_{C_i}^k = m_{C_j}^k \wedge M_{C_i}^k = M_{C_j}^k$	341
B.2.3 Caso 4, $m_{C_i}^k > m_{C_j}^k \wedge M_{C_i}^k > M_{C_j}^k \wedge m_{C_i}^k < M_{C_j}^k$	343
B.2.4 Caso 5, $m_{C_i}^k < m_{C_j}^k \wedge M_{C_i}^k < M_{C_j}^k \wedge m_{C_i}^k < M_{C_j}^k$	345
B.2.5 Caso 6, $m_{C_i}^k < m_{C_j}^k \wedge M_{C_i}^k > M_{C_j}^k$	346
B.2.6 Caso 7, $m_{C_i}^k > m_{C_j}^k \wedge M_{C_i}^k < M_{C_j}^k$	347
B.2.7 Caso 8, $m_{C_i}^k = m_{C_j}^k \wedge M_{C_i}^k < M_{C_j}^k$	348
B.2.8 Caso 9, $m_{C_i}^k = m_{C_j}^k \wedge M_{C_i}^k > M_{C_j}^k$	349
B.2.9 Caso 10, $m_{C_i}^k > m_{C_j}^k \wedge M_{C_i}^k = M_{C_j}^k$	351
B.2.10 Caso 11, $m_{C_i}^k < m_{C_j}^k \wedge M_{C_i}^k = M_{C_j}^k$	352
B.2.11 Caso 12, $M_{C_i}^k = m_{C_j}^k$	353
B.2.12 Caso 13, $M_{C_j}^k = m_{C_i}^k$	355
C Análisis descriptivo de los datos, planta catalana	359
C.1 Análisis Univariante	359
C.1.1 Variables de entrada	359
C.1.2 Variables después de la decantación	366
C.1.3 Variables del tratamiento biológico	371
C.1.4 Variables de salida	379
C.2 Análisis Bivariante	385
C.2.1 Caudales	385
C.2.2 Sólidos en suspensión (SS)	386
C.2.3 Sólidos volátiles en suspensión (SSV).	387
C.2.4 Relación entre SS y SSV.	388
C.2.5 Materia orgánica biodegradable (DBO)	389
C.2.6 Materia orgánica degradable (DQO)	389
C.2.7 Materia orgánica biodegradable vs Materia orgánica degradable	389

C.2.8 Comportamiento del PH	390
C.2.9 Licor Mezcla. MLSS-B y MLVSS-B.	391
D Análisis descriptivo por clases, planta catalana	393
D.1 Clustering Based on Rules planta catalunya	393
D.1.1 Class-variable: 2-classes [$P2_{Gi1,R1}^{En,G}$]	393
D.1.2 Class variable: 3-classes [$P3_{Gi1}^{EnW,G}$]	396
D.1.3 Class variable: 4-classes [$P4_{Gi1}^{EnW,G}$]	399
E BbD, BbIR y $\mathcal{R}(\mathcal{P}_\xi^*)$ planta catalana	403
E.1 BbD para $\mathcal{P}_2^* \equiv \mathcal{P}2_{Gi1,R1}^{EnW,G}$	403
E.2 $\mathcal{R}(\mathcal{P}_2^*)$	406
E.3 BbD para $\mathcal{P}_3^* \subseteq \mathcal{P}3_{Gi1,R1}^{EnW,G}$	410
E.4 $\mathcal{R}(\mathcal{P}_3^*)$	413
E.5 BbD para $\mathcal{P}_4^* \subseteq \mathcal{P}4_{Gi1,R1}^{EnW,G}$	417
E.6 $\mathcal{R}(\mathcal{P}_4^*)$	420
F Análisis descriptivo de los datos, planta eslovena	425
F.1 Análisis univariate	425
F.1.1 Variable NH4-influent	425
F.1.2 Variable NH4-2aerobic	426
F.1.3 Variable O2-1aerobic	428
F.1.4 Variable O2-2aerobic	430
F.1.5 Variable Valve-air	431
F.1.6 Variable Q-air	433
F.1.7 Variable h-ww	434
F.1.8 Variable Q-influent	436
F.1.9 Variable FR1-DOTOK	438
F.1.10 Variable Freq-rec	440
F.1.11 Variable TN-influent	441
F.1.12 Variable TN-effluent	443
F.1.13 Variable TOC-influent	444
F.1.14 Variable Nitritox-influent	446
F.1.15 Variable TOC-effluent	448
F.2 Análisis bivariate	450
G Análisis descriptivo por clases, planta eslovena	453
G.1 Clustering Based on Rules (Family 3) planta eslovena	454
G.1.1 Class variable: 2-classes [$P2_{Lj3,R2}^{EnW,G}$]	454
G.1.2 Class variable: 3-classes [$P3_{Lj3,R2}^{EnW,G}$]	457
G.1.3 Class variable: 4-classes [$P4_{Lj3,R2}^{EnW,G}$]	460
H BbD, BbIR y $\mathcal{R}_e(\mathcal{P}_\xi^*)$ planta eslovena	463
H.1 BbD para $\mathcal{P}_2^* \equiv \mathcal{P}2_{Lj3,R2}^{EnW,G}$	463
H.2 $\mathcal{R}(\mathcal{P}_2^*)$	466
H.3 BbD para $\mathcal{P}_3^* \subseteq \mathcal{P}3_{Lj3,R2}^{EnW,G}$	469
H.4 $\mathcal{R}(\mathcal{P}_3^*)$	471
H.5 BbD para $\mathcal{P}_4^* \subseteq \mathcal{P}4_{Lj3,R2}^{EnW,G}$	474
H.6 $\mathcal{R}(\mathcal{P}_4^*)$	476

Indice de Figuras

3.1	Clasificación de tipos de problemas propuesta por Simon en 1960.	15
3.2	Componentes de un IDSS.	16
3.3	Diagrama del proceso KDD.	18
5.1	Estructura de τ	47
5.2	Gráfico de inercia interna de las clases $[\tau_{Lj3,R2}^{EnW,G}]$	48
5.3	Boxplot de la variable MLSS-B.	50
5.4	Boxplot múltiple de la variable DBO-D <i>versus</i> $\mathcal{P}_2 = \{C_{391}, C_{392}\}$	50
5.5	Boxplot múltiple de la variable Q_E <i>vs</i> partición en 4 clases.	53
5.6	Boxplot múltiple de la variable Q_E <i>vs</i> partición en 4 clases, con ventanas de longitud variable para inducir la codificación I^k de Q_E	54
5.7	Grados de pertenencia de X_K a una partición \mathcal{P}_4 en 4 clases.	58
7.1	Caso 1: Boxplot based Discretization.	79
7.2	Caso 1: Boxplot based Discretization.	80
7.3	Ejemplo BbD revisado con patrón <i>centro abierto</i> para X_k vs \mathcal{P}_2	87
8.1	Árbol general de clasificación, corte en 2 clases.	97
8.2	Árbol general de clasificación, corte en 3 clases.	98
8.3	Dendrograma.	101
9.1	Ejemplo de inconsistencia.	106
13.1	Diagrama típico del proceso de tratamiento de aguas residuales.	136
13.2	Diagrama típico del proceso de lodos activos.	138
14.1	Vista aérea de la EDAR Catalana.	139
14.2	Línea de aguas de la planta.	140
14.3	Línea de lodos de la planta.	141
14.4	Caudal de entrada a la planta <i>versus</i> caudal de entrada al reactor biológico. .	144
15.1	Puntos de medición de variables.	147
15.2	Histograma de la variable Q-E.	148
15.3	Boxplot de la variable Q-E.	148
15.4	Diagrama bivariante para las variables SS-E y SS-D.	149
16.1	Árboles inducidos por las reglas de clasificación (P izq.-arriba, S der.-arriba y Q der.abajo).	154
16.2	CAJ. Árbol general de classificación con reglas. $[\tau_{Gi2,R1}^{En,G}]$	155
16.3	Análisis Descriptivo por clases para $[P2_{Gi1,R1}^{En,G}]$	156
16.4	Análisis Descriptivo por clases para $[P2_{Gi1,R1}^{En,G}]$	157

16.5 Análisis Descriptivo por clases para $[P3_{Gi1,R1}^{En,G}]$	158
16.6 Análisis Descriptivo por clases para $[P3_{Gi1,R1}^{En,G}]$	159
16.7 Análisis Descriptivo por clases para $[P4_{Gi1,R1}^{En,G}]$ -1.	161
16.8 Análisis Descriptivo por clases para $[P4_{Gi1,R1}^{En,G}]$ -2.	162
16.9 Análisis Descriptivo por clases para $[P4_{Gi1,R1}^{En,G}]$	163
16.10 Análisis Descriptivo por clases para $[P4_{Gi1,R1}^{En,G}]$	164
 19.1 Vista aérea de la planta (1999).	201
19.2 Otra vista area de la planta (2006).	202
19.3 Localización geográfica de la planta(WWTP).	202
19.4 Esquema simplificado del sistema de alcantarillas.	204
19.5 Diseño con las principales unidades de la planta.	206
19.6 Diseño en la planta del sistema SCADA.	207
19.7 Control de Oxígeno.	208
19.8 El diseño actual de la plantas piloto dentro de la EDAR eslovena.	209
19.9 Proceso.	211
19.10MBBR (<i>Moving Bed Biofilm Reactor</i>) planta piloto con sensores, actuadores y variables.	213
19.11Máquina de medición en la línea de las variables TOC y TN.	215
19.12Máquina de medición en la línea de las variables TOC y TN (zoom).	215
19.13Máquina de medición en la línea de la variable NH4-N (NH4-2aerobic)(punto superior).	216
19.14Máquina de medición en la línea de la variable NH4-N (NH4-2aerobic)(punto inferior).	216
19.15Máquina de medición en la línea de la variable <i>influent inhibition</i> (Nitritox). .	216
19.16Máquina de medición en la línea de la variable <i>influent inhibition</i> (Nitritox). .	217
 20.1 Serie temporal para la variable Temp-wastewater.	220
20.2 Histograma y Boxplot de la variable Temp-ww.	220
20.3 Diagrama bivariante para las variables TOC-influent and TOC-effluent. . .	222
20.4 Serie tiemporal para 4 variables.	222
20.5 Serie temporal para 3 variables.	223
 21.1 Árboles inducidos por las reglas de clasificación (Mmonia izq. y Nitrogen der.).	228
21.2 Gráfico de inercia interna de las clases $[\tau_{Lj3,R2}^{EnW,G}]$	229
21.3 Dendograma (árbol de clasificación) $[\tau_{Lj3,R2}^{EnW,G}]$	230
21.4 Análisis descriptivo por clases para $[P2_{Lj3,R2}^{EnW,G}]$	231
21.5 Análisis descriptivo por clases para $[P3_{Lj3,R2}^{EnW,G}]$	232
21.6 Análisis descriptivo por clases para $[P4_{Lj3,R2}^{EnW,G}]$	234
21.7 Análisis descriptivo por clases para $[P4_{Lj3,R2}^{EnW,G}]$	235
 A.1 El proceso de los vecinos recíprocos encadenados.	330
A.2 Cronología de KLASS.	335
 B.1 Caso 2: Boxplot based Discretization.	339
B.2 Caso 3: Boxplot based Discretization.	341
B.3 Caso 4: Boxplot based Discretization.	343
B.4 Caso 5: Boxplot based Discretization.	345
B.5 Caso 6: Boxplot based Discretization.	346

B.6 Caso 7: Boxplot based Discretization.	347
B.7 Caso 8: Boxplot based Discretization.	348
B.8 Caso 9: Boxplot based Discretization.	350
B.9 Caso 10: Boxplot based Discretization.	351
B.10 Caso 11: Boxplot based Discretization.	352
B.11 Caso 12: Boxplot based Discretization.	353
B.12 Caso 13: Boxplot based Discretization.	355
C.1 Histograma de la variable FE-E.	359
C.2 Boxplot de la variable FE-E.	359
C.3 Histograma de la variable PH-E.	360
C.4 Boxplot de la variable PH-E.	361
C.5 Histograma de la variable SS-E.	361
C.6 Boxplot de la variable SS-E.	362
C.7 Histograma de la variable SSV-E.	362
C.8 Boxplot de la variable SSV-E.	363
C.9 Histograma de la variable DQO-E.	364
C.10 Boxplot de la variable DQO-E.	364
C.11 Histograma de la variable DBO-E.	365
C.12 Boxplot de la variable DBO-E.	365
C.13 Histograma de la variable PH-D.	366
C.14 Boxplot de la variable PH-D.	366
C.15 Histograma de la variable SS-D.	367
C.16 Boxplot de la variable SS-D.	367
C.17 Histograma de la variable SSV-D.	368
C.18 Boxplot de la variable SSV-D.	368
C.19 Histograma de la variable DQO-D.	369
C.20 Boxplot de la variable DQO-D.	369
C.21 Histograma de la variable DBO-D.	370
C.22 Boxplot de la variable DBO-D.	370
C.23 Histograma de la variable V30-B.	371
C.24 Boxplot de la variable V30-B.	371
C.25 Histograma de la variable MLSS-B.	372
C.26 Boxplot de la variable MLSS-B.	372
C.27 Histograma de la variable MLVSS-B.	373
C.28 Boxplot de la variable MLVSS-B.	374
C.29 Histograma de la variable MCRT-B.	374
C.30 Boxplot de la variable MCRT-B.	375
C.31 Histograma de la variable QB-B.	375
C.32 Boxplot de la variable QB-B.	376
C.33 Histograma de la variable QR-G.	376
C.34 Boxplot de la variable QR-G.	377
C.35 Histograma de la variable QP-G.	377
C.36 Boxplot de la variable QP-G.	378
C.37 Histograma de la variable QA-G.	378
C.38 Boxplot de la variable QA-G.	379
C.39 Histograma de la variable PH-S.	380
C.40 Boxplot de la variable PH-S.	380
C.41 Histograma de la variable SS-S.	381
C.42 Boxplot de la variable SS-S.	381

C.43 Histograma de la variable SSV-S.	382
C.44 Boxplot de la variable SSV-S.	382
C.45 Histograma de la variable DQO-S.	383
C.46 Boxplot de la variable DQO-S.	383
C.47 Histograma de la variable DBO-S.	384
C.48 Boxplot de la variable DBO-S	384
C.49 Diagrama bivariante para las variables Q-E y QB-B.	385
C.50 Letter-plot del caudal de entrada por el caudal del reactor biológico.	385
C.51 Diagrama bivariante para las variables SS-D y SS-S	386
C.52 Diagrama bivariante para las variables SS-E y SS-S	386
C.53 Diagrama bivariante para las variables SSV-E y SSV-D.	387
C.54 Diagrama bivariante para las variables SSV-D y SSV-S.	387
C.55 Diagrama bivariante para las variables SSV-S y SSV-E.	388
C.56 Diagrama bivariante para las variables SS-E y SSV-E.	388
C.57 Diagrama bivariante para las variables DBO-E y DBO-D.	389
C.58 Diagrama bivariante para las variables DQO-E y DQO-D.	389
C.59 Diagrama bivariante para las variables DQO-E y DBO-E.	390
C.60 Diagrama bivariante para las variables DBO-D y DQO-D.	390
C.61 Diagrama bivariante para las variables PH-E y PH-S.	391
C.62 Diagrama bivariante para las variables PH-D y PH-S.	391
C.63 Diagrama bivariante para las variables QB-B y PH-S.	391
C.64 Diagrama bivariante para las variables MLSS-B y MLVSS-B.	392
F.1 Histograma y Boxplot de la variable NH4-influent.	425
F.2 Serie temporal para variable NH4-influent.	426
F.3 Serie temporal para variable NH4-2aerobic.	427
F.4 Histograma y Boxplot de la variable NH4-2aerobic.	428
F.5 Serie temporal para variable O2-1aerobic.	428
F.6 Histograma y Boxplot de la variable Variable O2-1aerobic.	429
F.7 Serie temporal para variable O2-2aerobic.	430
F.8 Histograma y Boxplot de la variable Variable O2-2aerobic.	430
F.9 Serie temporal para variable Valve-air.	431
F.10 Histograma y Boxplot de la variable Variable Valve-air.	432
F.11 Serie temporal para variable Q-air.	433
F.12 Histograma y Boxplot de la variable Variable Q-air.	433
F.13 Serie temporal para variable h-wastewater.	435
F.14 Histograma y Boxplot de la variable Variable h-ww.	435
F.15 Serie temporal para variable Q-influent.	436
F.16 Histograma y Boxplot de la variable Variable Q-influent.	437
F.17 Serie temporal para variable FR1-DOTOK-20s.	438
F.18 Histograma y Boxplot de la variable Variable FR1-DOTOK.	439
F.19 Serie temporal para variable Freq-rec.	440
F.20 Histograma y Boxplot de la variable Variable Freq-rec.	440
F.21 Serie temporal para variable TN-influent.	442
F.22 Histograma y Boxplot de la variable Variable TN-influent.	442
F.23 Serie temporal para variable TN-effluent.	443
F.24 Histograma y Boxplot de la variable Variable TN-effluent.	443
F.25 Serie temporal para variable TOC-influent.	444
F.26 Histograma y Boxplot de la variable Variable TOC-influent.	445
F.27 Serie temporal para variable Nitritox-influent.	446

F.28 Histograma y Boxplot de la variable Variable Nitritox-influent.	446
F.29 Serie temporal para variable TOC-effluent.	448
F.30 Histograma y Boxplot de la variable Variable TOC-effluent.	448
F.31 Diagrama bivariante para las variables TN-effluent and NH4-2aerobic.	450
F.32 Diagrama bivariante para las variables NH4-influent and NH4-2aerobic.	450
F.33 Diagrama bivariante para las variables O2-1aerobic and O2-2aerobic.	451
F.34 Diagrama bivariante para las variables TN-influent and TN-effluent.	451

Indice de tablas

1.1	Matriz de datos \mathcal{X}	7
1.2	Matriz de datos y clases.	7
5.1	Matriz de datos \mathcal{X}	45
5.2	Valores caracterizados.	51
5.3	Variables caracterizadoras.	52
5.4	relación de valores caracterizados y una clase C.	52
5.5	Relación entre valores caracterizados y p_{sc}	58
5.6	Negación (\neg), Disyunción (\vee), Conjunción (\wedge), Condicional (\rightarrow) y Bicondicional (\leftrightarrow).	60
7.1	Valores caracterizados.	71
7.2	Relación entre valores caracterizados y p_{sc}	72
7.3	Relación de valores y cardinalidad de los Sistema de reglas para el Caso 1. . .	82
7.4	Descripción de los casos estudiados, ver detalles en Anexo B.	84
7.5	Patrón y descripción de los casos estudiados, ver detalles en Anexo B.	85
7.6	Comparación entre la propuesta original y la revisada para la construcción de los intervalos, ver detalles en Anexo B.	86
7.7	Nueva propuesta para los sistemas de reglas efectivas.	88
7.8	Tipos de valores.	89
7.9	Comparación para los sistemas de reglas completos.	91
7.10	Cardinalidad del sistema de reglas efectivas y número de intervalos propios. .	96
8.1	Tabla de contingencia de \mathcal{P}_ξ vs $\mathcal{P}_{\xi+1}$	99
10.1	Cobertura relativa de $\mathcal{S}(\mathcal{P}_\xi)$	110
14.1	Cotas permitidas por la directiva del Consejo 91/271/CEE.	143
17.1	Cobertura relativa de $\mathcal{S}(\mathcal{P}_2)$	167
17.2	Cobertura relativa de $\mathcal{S}(\mathcal{P}_3^*)$	170
17.3	Cobertura relativa de $\mathcal{S}(\mathcal{P}_4^*)$	173
17.4	Evaluación: Best global concept and Close-World Assumption.	176
17.5	Cobertura relativa de $\mathcal{S}_{Classer392}(\mathcal{P}_2)$	179
17.6	Cobertura relativa de $\mathcal{S}_{Classer393}(\mathcal{P}_2)$	180
17.7	Cobertura relativa de $\mathcal{S}_{Classer389}(\mathcal{P}_3^*)$	184
17.8	Cobertura relativa de $\mathcal{S}_{Classer391}(\mathcal{P}_3^*)$	185
17.9	Cobertura relativa de $\mathcal{S}_{Classer383}(\mathcal{P}_4^*)$	187
17.10	Cobertura relativa de $\mathcal{S}_{Classer390}(\mathcal{P}_4^*)$	188
17.11	Evaluación: Best local-global concept and Close-World Assumption.	191

18.1 Resumen de las interpretaciones obtenidas por las diferentes propuestas.	195
18.2 Cobertura relativa de DQO-D, SS-D y SSV-D en $\mathcal{S}(\mathcal{P}_3^*)$	196
18.3 Cobertura relativa de DQO-E y DBO-E en $\mathcal{S}(\mathcal{P}_4^*)$	196
18.4 Class panel Graph para las variables de diferencia.	196
18.5 Descriptiva estadística por clase.	197
18.6 Resumen resultados.	197
19.1 Tamaño de la planta y configuración del proceso.	203
19.2 Caudal de entrada: datos teóricos y datos reales.	204
19.3 Unidades de la planta y volúmenes.	205
19.4 Normas del efluente.	206
19.5 Afluentes y efluentes de aguas residuales característicos de la planta.	207
19.6 Costos de operación de la EDAR: Domzale-Kamnik wastewater.	207
22.1 Cobertura relativa de $\mathcal{S}(\mathcal{P}_2^*)$	238
22.2 Cobertura relativa de $\mathcal{S}(\mathcal{P}_3^*)$	240
22.3 Cobertura relativa de $\mathcal{S}(\mathcal{P}_4^*)$	242
22.4 Evaluación: Best global concept and Close-World Assumption.	244
22.5 Cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$	246
22.6 Cobertura relativa de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$	246
22.7 Cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$	248
22.8 Cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$	248
22.9 Cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$	250
22.10 Cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$	251
22.11 Evaluación: Best local concept and not Close-World Assumption.	253
22.12 Cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$	255
22.13 Cobertura relativa de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$	255
22.14 Cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$	257
22.15 Cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$	257
22.16 Cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$	260
22.17 Cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$	260
22.18 Evaluación: Best local concept and Close-World Assumption.	263
22.19 Cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$	265
22.20 Cobertura relativa de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$	265
22.21 Cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$	267
22.22 Cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$	267
22.23 Cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$	270
22.24 Cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$	270
22.25 Evaluación: Best local concept and partial Close-World Assumption.	273
22.26 Cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$	275
22.27 Cobertura relativa de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$	275
22.28 Cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$	277
22.29 Cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$	277
22.30 Cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$	280
22.31 Cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$	280
22.32 Evaluación: Best local-global concept and Close-World Assumption.	283
23.1 Resumen de las interpretaciones obtenidas por las diferentes propuestas.	287
23.2 Resumen resultados.	289
B.1 Caso 2: Relación entre Intervalos y Reglas (Propuesta original).	340

B.2 Caso 2: Relación entre Intervalos y Reglas (Propuesta revisada).	341
B.3 Caso 5: Relación entre Intervalos y Reglas (Propuesta original).	345
B.4 Caso 5: Relación entre Intervalos y Reglas (Propuesta revisada).	345
B.5 Caso 6: Relación entre Intervalos y Reglas (Propuesta original).	346
B.6 Caso 6: Relación entre Intervalos y Reglas (Propuesta revisada).	347
B.7 Caso 7: Relación entre Intervalos y Reglas (Propuesta original).	348
B.8 Caso 7: Relación entre Intervalos y Reglas (Propuesta revisada).	348
B.9 Caso 8: Relación entre Intervalos y Reglas (Propuesta original).	349
B.10 Caso 8: Relación entre Intervalos y Reglas (Propuesta revisada).	349
B.11 Caso 9: Relación entre Intervalos y Reglas (Propuesta original).	350
B.12 Caso 9: Relación entre Intervalos y Reglas (Propuesta revisada).	350
B.13 Caso 10: Relación entre Intervalos y Reglas (Propuesta original).	351
B.14 Caso 10: Relación entre Intervalos y Reglas (Propuesta revisada).	352
B.15 Caso 11: Relación entre Intervalos y Reglas (Propuesta original).	352
B.16 Caso 11: Relación entre Intervalos y Reglas (Propuesta revisada).	353

Parte I

Introducción y Fundamentos Teóricos

Capítulo 1

Introducción y motivación

Se habla en la actualidad de que nos hallamos en la era de la Sociedad del Conocimiento. La noción de Sociedad del Conocimiento fue utilizada por primera vez en 1969 por un autor austriaco relacionado con el *management* o gestión, llamado Peter Drucker (Drucker 1969) y en el decenio de 1990 se profundizó en una serie de estudios detallados publicados por investigadores como Robin Mansel (Mansell 2000) o Nico Stehr (Stehr 1994).

En 1974, Peter Drucker (Drucker 1974) reclamaba para el futuro una sociedad del conocimiento en la que el recurso básico sería el saber, donde la voluntad de aplicar conocimiento para generar más conocimiento debía basarse en un elevado esfuerzo de sistematización y organización y colocaba el conocimiento en el centro del motor que movería la sociedad, señalando que lo mas importante no era la cantidad de conocimiento en sí, sino su potencial.

Las sociedades de la información surgieron con el uso e innovaciones intensivas de las tecnologías de la información y las comunicaciones, donde el incremento en la transferencia de información, modificó en muchos sentidos la forma en que se desarrollan muchas actividades en la sociedad moderna. Sin embargo, la información no es lo mismo que el conocimiento, ya que la información es efectivamente un instrumento del conocimiento, pero no es el conocimiento en sí, el conocimiento obedece a aquellos elementos que pueden ser comprendidos por cualquier mente humana razonable (Drucker 1974).

En las últimas décadas, el crecimiento explosivo de los avances científicos y sobre todo tecnológicos han generado sistemas complejos que han rebasado nuestra capacidad para analizarlos e interpretarlos, creando la necesidad de una nueva generación de métodos, técnicas y herramientas con la capacidad para asistir inteligente y automáticamente a los seres humanos en el análisis de estas bases de datos para extraer conocimiento útil que represente los dominios del mundo real cada vez más complejo y den soporte a la *toma de decisiones*, elemento fundamental de la sociedad de la información.

Gestionar estos dominios es una tarea enormemente compleja y tomar decisiones adecuadas requiere tener una percepción completa y ajustada de la realidad que se está manejando. Por ello en los últimos años se han invertido esfuerzos en proporcionar metodologías de toma de decisiones informadas y desarrollar herramientas de soporte a dicha labor (Gibert 1994). La dificultad crece si nos hallamos ante un *Dominio Poco Estructurado* (Ill-Structured Domain, ISD). Gibert en su tesis doctoral (Gibert 1994) aparecen caracterizados por primera vez. En (Gibert and Cortés 1998a) se definen como aquéllos donde:

- Los elementos del dominio vienen descritos por conjuntos *heterogéneos* de variables, siendo las variables numéricas y las categóricas igualmente relevantes en dicha descripción. Estas últimas acostumbran a tener mayor número de modalidades cuanto mayor es el grado de conocimiento (*expertise*) de quien origina los datos.
- Existe un *conocimiento a priori adicional* sobre la estructura del dominio. Suele ser

conocimiento declarativo relativo a la estructura global del mismo (relaciones entre variables, objetivos de clasificación que se persiguen, ...).

- La complejidad inherente al dominio hace que el conocimiento que de él se tiene sea *parcial* (en estos dominios existe gran cantidad de conocimiento implícito y grandes incógnitas) y *no homogéneo* (el grado de especificidad del conocimiento disponible es distinto para distintas partes del dominio).

En efecto, descubrir la estructura de fondo de un dominio permite organizar mejor el conocimiento del experto y facilita esta toma de decisiones especialmente en dominios poco estructurados. La realidad es que en muchas aplicaciones reales la estructura del dominio es desconocida y precisamente lo que se pretende descubrir, con lo que la aproximación supervisada que se apoya en el principio de que la estructura conceptual del dominio es previamente conocida (Ripley 1996), (Cheeseman and Stutz 1996) deja de ser aplicable. En ese caso las técnicas de *clustering* nos permiten identificar las clases existentes en el dominio a partir de la descripción de un conjunto de casos. Por ello ante realidades desconocidas, o con lagunas, la utilización del clustering es una herramienta fundamental de extracción de conocimiento.

La comprensión de la naturaleza de los métodos que utilizamos los seres humanos para clasificar datos o conocimientos, es un problema de gran interés teórico y práctico ya que la acción de clasificar es uno de los procesos básicos de la mente humana y una de las etapas iniciales de los procesos de adquisición de conocimiento en cualquier campo científico (Gibert 2004). Así resulta importante también en el proceso cognitivo de toma de decisiones puesto que en un buen número de situaciones el decisor realiza procesos mentales de clasificación de la situación para *asignar* la acción o decisión correcta.

En la práctica, el desarrollo de sistemas automáticos de clasificación que pueden tratar grandes Bases de Datos provenientes de dominios poco estructurados es, hoy por hoy, una necesidad imperiosa de la sociedad actual, ya que en muchos procesos la cantidad de datos que se genera es tan grande, que resulta muy difícil manipularlos y transmitirlos sin el auxilio de esta clase de sistemas.

Como el clustering produce clases donde se desconoce la estructura real subyacente, la validación de los resultados obtenidos se hace muy difícil. De hecho la validación de un cluster sigue siendo un problema abierto por no haberse encontrado aún un criterio objetivo para determinar la *calidad* de un conjunto de clases en el contexto del clustering (Hand 1996), que se aplica en situaciones en las que *no hay* un buen conocimiento de la estructura del dominio que pueda servir de referencia (como se hace en el caso supervisado), y si lo hay, es sólo parcial. Uno de los criterios utilizados habitualmente hasta ahora para la validación de clases es su *utilidad* que pasa por que tengan *significado* claro para el experto desde el punto de vista conceptual. De hecho Volle (Volle 1985) hace toda una disertación ilustrando que el concepto de validez no es absoluto, sino relativo a las condiciones del contexto y a la utilidad de las mismas. A pesar de que este es un extremo poco objetivable, la *interpretación* se convierte así en una fase fundamental del proceso y sigue siendo, aún hoy, uno de los criterios más utilizados en la práctica para validar el cluster. Por esta razón la validación queda directamente ligada a la existencia de una interpretación clara para el clustering o partición. Actualmente es necesario introducir herramientas para asistir al usuario en las tareas de interpretación de una partición sobre un conjunto de objetos, para establecer el significado de las clases resultantes. Si las clases obtenidas no tienen sentido para el/los expertos, los resultados de la clasificación no son considerados válidos.

Sin embargo todas las técnicas y algoritmos de validación van orientados a validar la vertiente estructural de la partición, ver sección §3.5.4, pero disponer de clases bien formadas estructuralmente no ofrece la menor garantía de que un experto vaya a ser capaz de asociar

cada uno de esos grupos a una entidad semántica de su dominio a un concepto de su base de conocimiento personal sobre la base del cual él pueda tomar decisiones y razonar. Por ello es fundamental que, aparte de validar una clase estructuralmente se asiste al usuario a comprender cual fue el criterio de clasificación subyacente y a entender el significado de las clases. En relación a la generación de conceptos a partir de clases.

De hecho, la gran crítica que hacen algunos autores a las técnicas de clustering en todas sus vertientes, es que al generar solamente una descripción extensional de las clases, queda para el analista su interpretación y caracterización (McCarthy 1983), (Fisher 1993) no es inmediata. No basta con obtener automáticamente las clases, sino que se necesita ayuda para entender *por qué* se detectan unas ciertas clases y no otras. Asistir al experto en estas tareas con herramientas automáticas es otro de los tópicos importantes para un sistema de *Knowledge Discovery from Data (KDD)* y también para el soporte de la toma de decisiones (Fayyad 1996). Algunos paquetes estadísticos (como SPAD¹ y SPSS²) incluyen varias herramientas orientadas a la interpretación de una clasificación dada, como la posibilidad de calcular la contribución de cierta variable a la formación de una clase. Sin embargo, en la etapa final, la interpretación misma viene haciéndola el experto en una forma no-sistemática; usando su propia experiencia el experto examina los resultados del clustering con mayor o menor asistencia por parte del software (que puede proporcionar información gráfica o estadística de las clases más o menos sofisticada), pero es responsabilidad exclusiva del experto integrar todos esos elementos con su conocimiento del dominio para finalmente sintetizar el concepto asociado a cada clase y evaluar si se corresponde o no a una entidad semántica reconocida en el dominio de aplicación.

El proceso de interpretación es extremadamente complejo cuando el número de variables que describen los casos es grande y en aplicaciones relacionadas con *Data Mining* es frecuente que el número de clases descubiertas sea también grande, lo que convierte la interpretación de resultados en una tarea todavía difícil de afrontar.

De esta forma la alternativa que parece más prometedora para resolver estas limitaciones es aligerar al experto de este trabajo, mediante el desarrollo de técnicas que a partir de la evidencia empírica, identifiquen las variables más relevantes y formulen conceptos que expresen las particularidades de cada clase y se expresen en una forma de representación conceptual generable automáticamente y directamente comprensible para el experto. Además incorporar procedimientos que trasladen los resultados del análisis (en este caso del *clustering*) a una representación explícita del conocimiento obtenido, se sitúa en la línea de lo que Fayyad propone para los sistemas de KDD, donde la fase de post-proceso de los resultados para generar conocimiento es casi tan importante como el análisis en sí mismo. Quizás por su naturaleza más semántica la generación automática de interpretaciones de una clasificación no se ha tratado formalmente desde el ámbito estadístico, aunque resolverlo es fundamental. Éste, de hecho, es uno de los problemas objeto del aprendizaje automático, del cual ID3(Quinlan 1990) y sus sucesores son exponentes característicos.

Por lo anterior, el interés de esta tesis es presentar una propuesta metodológica, tal que, a partir de las variables que describen los objetos pertenecientes a cierto dominio, podamos caracterizar las situaciones características (clases resultantes) que se pueden encontrar en él, contribuyendo así al proceso de interpretación conceptual automática de clases procedentes de un cluster. De hecho esta tesis culmina un trabajo intenso y complejo desarrollado en los últimos 5 años para resolver un problema tan difícil como omnipresente en las aplicaciones reales. A lo largo de este documento se hace referencia, donde es oportuno, a las publicaciones que esta investigación ha dado lugar durante su desarrollo. Al final del documento se

¹Paquete estadístico orientado al análisis multivariante. (Lebart, Morineau, and Lambert 1994)

²Paquete estadístico de propósito general. (Visauta 1998)

encuentra una sección con la relación completa de las mismas. En este trabajo se presentan resultados que contribuyen a:

- Facilitar esta tarea en bases de datos con muchas variables y muchas clases.
- Aportar una sistematización al proceso de interpretación de clases procedentes de un cluster jerárquico, realizando hasta ahora de forma mas o menos artesanal y objetivizar los mecanismos de interpretación.
- Generar conocimiento explícito directamente a partir de las clases de forma que el experto pueda comprender más fácilmente las características principales de la clasificación obtenida.
- Objetivar los procesos de validación de los resultados de un proceso de cluster.
- Contribuir a la concepción de sistemas integrales de KDD donde el post-proceso se incluye en la metodología global para producir conocimiento explícito y directamente comprensible por el experto.

Como es ya característico de los problemas de KDD se requiere combinar técnicas y herramientas de diversos campos, en nuestro caso, de Estadística (análisis multivariante de datos, clustering, etc.), Inteligencia Artificial (aprendizaje automático, sistemas basados en el conocimiento), Sistemas de Información (análisis, diseño, implementación, etc.), etc. para construir *Metodologías Híbridas* que permitan interpretar situaciones típicas (o conceptos) en las bases de datos, para extraer conocimiento útil que represente la estructura de estos dominios y que den mejor desempeño que las técnicas tradicionales o las aproximaciones clásicas de sistemas basados en conocimiento.

Una vez identificadas e interpretadas estas situaciones típicas, el conocimiento generado puede ser usado posteriormente como herramientas de apoyo al proceso de toma de decisiones o a los sistemas de gestión. Incluso se ha llegado a decir que la *validación de una clasificación* consiste, precisamente, en *probar* que las clases tienen *sentido* o *utilidad* (Aluja 1996). En esta dirección, la propuesta pretende llegar un poco más lejos y establecer las bases de una metodología que facilite la generación automática de caracterizaciones e interpretaciones conceptuales en estos dominios poco estructurados. Obviamente las interpretaciones generadas estarán muy cerca de resolver el problema ulterior de generación automática de reglas de clasificación a utilizar con fines predictivos, fundamentales como apoyo a la toma de decisiones, lo que confiere un doble interés a esta investigación.

1.1 Formulación del Problema de tesis

En esta sección se hace una descripción formal del problema de tesis.

Sea $\mathcal{I} = \{i_1, \dots, i_n\}$ un conjunto de individuos u objetos de un universo de discurso, que está descrito por una serie de variables cualitativas y/o cuantitativas $X_1 \dots X_K$, cuyos valores para cada uno de los individuos $i \in \mathcal{I}$ se representan por una matriz rectangular \mathcal{X} de dimensión (n, K) , como se muestra en la Tabla 1.1:

Llamamos $\mathcal{D}^k = \{c_1^k, c_2^k, \dots, c_s^k\}$ al dominio de X_k , si ésta es categórica, es decir, al conjunto de valores posibles que puede tomar X_k y $D^k = r_k$ al dominio de X_k , si ésta es numérica, siendo $r_k = [\min X_k, \max X_k]$, el rango de los valores posibles de la variable X_k .

En la matriz \mathcal{X} , se tiene que x_{ik} con $1 \leq i \leq n$ y $1 \leq k \leq K$, es el valor que, el individuo i -ésimo toma para la k -ésima variable. Es decir, que las filas de la matriz de datos

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1K-1} & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K-1} & x_{2K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n-11} & x_{n-12} & \dots & x_{n-1K-1} & x_{n-1K} \\ x_{n1} & x_{n2} & \dots & x_{nK-1} & x_{nK} \end{pmatrix}$$

Tabla 1.1: Matriz de datos \mathcal{X} .

\mathcal{X} contienen información relativa a las características de los individuos, la cual se puede representar como un vector de variables de la forma:

$$x_i = (x_{i1} \ x_{i2} \ \dots \ x_{ik} \ \dots \ x_{nK})$$

y las columnas hacen referencia a las K variables X_k . Se supone desconocida la estructura subyacente al dominio, pero en esta tesis partiremos de la suposición que existe una partición de referencia de los elementos de \mathcal{I} en ξ clases, la que se denota por $\mathcal{P}_\xi = \{C_1, C_2, \dots, C_\xi\}$, y que puede ser el resultado de un clustering previo, o la puede haber suministrado el experto anteriormente. La cardinalidad $card(\mathcal{P}_\xi) = \xi$ y \mathcal{P}_ξ satisface las siguientes propiedades:

- $\cup_{C \in \mathcal{P}_\xi} C = \mathcal{I}$,
- $\forall C, C' \in \mathcal{P}_\xi, C \cap C' = \emptyset$

Siendo $C(i, \mathcal{P}_\xi)$ la clase a la que pertenece el individuo i en la partición \mathcal{P}_ξ , se puede construir una matriz extendida como se muestra en la Tabla 1.2.

$$\mathcal{X}_{\mathcal{P}} = \left(\begin{array}{ccccc|c} x_{11} & x_{12} & \dots & x_{1K-1} & x_{1K} & C(1, \mathcal{P}_\xi) \\ x_{21} & x_{22} & \dots & x_{2K-1} & x_{2K} & C(2, \mathcal{P}_\xi) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n-11} & x_{n-12} & \dots & x_{n-1K-1} & x_{n-1K} & C(n-1K, \mathcal{P}_\xi) \\ x_{n1} & x_{n2} & \dots & x_{nK-1} & x_{nK} & C(n, \mathcal{P}) \end{array} \right)$$

Tabla 1.2: Matriz de datos y clases.

Se quiere construir un sistema de conceptos $\mathcal{A}_{\mathcal{P}_\xi} = \{A_1, A_2, A_3, \dots, A_\xi\}$ que describen las clases de tal forma que:

- $A, A' \in \mathcal{A}_{\mathcal{P}_\xi} \Rightarrow A \neq A'$
- $\forall i \in \mathcal{I}, A_C(i) = \text{true} , \text{ si } C = C(i, \mathcal{P}_\xi), A_C \in \mathcal{A}_{\mathcal{P}_\xi}$
- $\forall i \in \mathcal{I}, A_C(i) = \text{false}, \text{ si } C \neq C(i, \mathcal{P}_\xi), A_C \in \mathcal{A}_{\mathcal{P}_\xi}$

Siendo $A_C \in \mathcal{A}_{\mathcal{P}_\xi}$ una expresión booleana que identifica de forma única los elementos de C , A_C función exclusiva de las variables X_1, X_2, \dots, X_K y los valores que éstas toman en cada clase y $A_C(i)$ es la evaluación de A_C sobre las coordenadas x_{ik} del individuo i .

Así,

$$A = [(\langle X \rangle \langle opr \rangle \langle v \rangle) \langle opl \rangle]^*$$

donde;

- $\langle X \rangle = \{X_1, X_2, \dots, X_K\}$

- $\langle opr \rangle = \{=, \leq, \geq, >, <, \in\}$
- $\langle opl \rangle = \{\vee, \wedge\}$
- $\langle v \rangle \subseteq \begin{cases} D^k & \text{si } X_k \text{ es categórica.} \\ D^k & \text{si } X_k \text{ es numérica.} \end{cases}$

1.2 Descripción del Proyecto Marco

Esta tesis se integra en un proyecto de investigación marco que dirige la Dra. Karina Gibert y que tiene por objetivo el desarrollo de metodologías híbridas de soporte al *Knowledge Discovery and Data Mining* en dominios poco estructurados (Gibert and Cortés 1994) para resolver problemas de soporte a la toma de decisiones en dominios médicos y medioambientales. La línea de investigación inicia en 1995 con el objetivo principal de estudiar este tipo de dominios desarrollando nuevas metodologías híbridas que combinen técnicas Estadísticas con las de Inteligencia Artificial para superar las limitaciones de las técnicas clásicas en las distintas fases del análisis de este tipo de dominios. Considerando las características especiales de este tipo de dominios, se han desarrollado métodos mixtos de análisis que combinan técnicas estadísticas con técnicas de inteligencia artificial para resolver los problemas que se plantean en este contexto (Gibert and Aluja 1998), (Gibert 2004).

La primera propuesta constituye la tesis de licenciatura (Gibert 1991) y después la tesis doctoral de Karina Gibert (Gibert 1994) que cristalizó en la formulación de la *metodología de clasificación basada en reglas* y una primera versión del sistema informático que la implementa, denominado *KLASS* (Gibert 1994) y que se ha utilizado en diversas aplicaciones reales (Gibert and Cortés 1993a; Gibert and Cortés 1993b; Gibert, Hernández, and Cortés 1996; Gibert and Cortés 1998a; Gibert and Sonicki 1999; Gibert and Roda 2000), para más detalles sobre la evolución de *KLASS* ver Anexo A.

Todas las metodologías que se desarrollan en el seno del proyecto marco se acaban integrando a lo que podríamos llamar *herramienta master*, que actualmente es *java.KLASS* (Gibert and Nonell 2008), y que aglutina herramientas de muy distinta naturaleza, ofreciendo la interfaz necesaria para que, los distintos módulos, puedan comunicarse entre ellos y transferir la información necesaria en cada momento del análisis.

Esta herramienta informática ha venido evolucionando de forma continua desde su origen en la medida en que se ha avanzado en la investigación y experimentación de la línea de investigación antes mencionada.

Desde su inicio *KLASS* ha venido vehiculando los métodos desarrollados con el tiempo y se han producido diversas ampliaciones del sistema que le han permitido evolucionar y abrir nuevas posibilidades de investigación. En el anexo A se aporta una breve cronología que destaca las etapas más relevantes del desarrollo de *KLASS*, que se orienta hacia una plataforma integrada de soporte al análisis inteligente de *dominios poco estructurados*, incluyendo distintos tipos de herramientas, desde las más básicas de análisis descriptivo hasta las más sofisticadas como la *clasificación basada en reglas* y herramientas de *apoyo a la interpretación de resultados*, relacionadas con la minería de datos y el proceso KDD (Gibert, Aluja, and Cortés 1998).

En el seno de este proyecto marco se han desarrollado distintas tesis doctorales, tesis de master, proyectos de fin de carrera (PFCs) tanto de la Diplomatura en Estadística como de la Licenciatura en técnicas Estadísticas o Ingeniería en Informática en todos sus niveles (superior y técnicas de sistemas y de gestión). Actualmente existe un grupo de personas investigando y trabajando en equipo, entre los que se están desarrollando 4 tesis doctorales en los programas de doctorado; Inteligencia Artificial y Aplicaciones Técnicas e Informáticas

de la Estadística, la Investigación Operativa y la Optimización de la Universidad Politécnica de Cataluña, para mas detalles ver Anexo A.

Las líneas principales de trabajo actualmente se centran en el desarrollo de herramientas y metodologías de soporte a la interpretación de clusterings y a la modelización y conceptualización de sistemas dinámicos en dominios médicos y medioambientales.

1.3 Estructura del documento

Este documento esta estructurado en cuatro partes y una sección de anexos:

Primera Parte: Introducción y Fundamentos Teóricos. En la primera parte de este trabajo se ofrece una rápida visión de lo que constituye la clasificación, interpretación y validación de bases de datos en dominios poco estructurados, comenzando con el capítulo §1 en el que se presenta la introducción y la motivación de esta tesis, incluyendo la formulación del problema, así como la descripción del proyecto de investigación en que se enmarca ésta y finalmente la estructura del presente documento. A continuación, en el capítulo §2, se enuncian los objetivos planteados para el desarrollo de esta investigación. Posteriormente en el capítulo §3 se describe el estado del arte, que permitirá contextualizar el la investigación desarrollada en esta tesis, abordando temas como los Sistemas de soporte a la toma de decisiones; *Knowledge Discovery from Data*; las áreas de conocimiento de Estadística, Inteligencia Artificial y *AI& Stats*; y por último el *Clustering* desde éstas 3 últimas áreas de conocimiento, así como lo que se ha hecho en validación de *Clustering*.

Finalmente, en el capítulo §4, los antecedentes de la línea de investigación consolidada en donde se inserta esta investigación y, en el capítulo §5, los conceptos necesarios para facilitar la compresión de los métodos y herramientas utilizados en el desarrollo de la tesis.

Segunda Parte: Marco Teórico y Desarrollo de la Metodología. En la segunda parte de este trabajo se presenta, en el capítulo §11, la metodología de caracterización conceptual por condicionamientos sucesivos (CCCS), resultado final de esta investigación. Alcanzar la metodología que se presenta ha requerido a su vez del desarrollo de un marco formal de apoyo sobre el que se construye posteriormente la propuesta final, es así como, el capítulo §7 presenta las definiciones de conceptos nuevos, métodos y propiedades sobre los que se construye la metodología.

Esta metodología aprovecha la estructura jerárquica de la clasificación objetivo para inducir conceptos iterando sobre las divisiones binarias de la secuencia de clases que indica el dendrograma y por ello en el capítulo §8 se presentan las ventajas de trabajar con una jerarquía indexada de clases lo que permitirá abordar el problema de la interpretación de forma recursiva. Seguidamente se definen, en el capítulo §9, los criterios con los que se selecciona el o los mejores conceptos (reglas probablizadas) asociado a las clases que se quieren interpretar. Por último, en el capítulo §10, se proponen 5 formas de combinar el conocimiento, extraído en forma de conceptos (reglas probablizadas), en cada iteración.

Tercera Parte: Aplicación. En esta parte se presenta la aplicación de la metodología CCCS a la interpretación de situaciones características en estaciones depuradoras de aguas residuales. En primer lugar, en el capítulo §13, se encuentra el dominio de aplicación; entorno en el cual la extracción automática de conocimiento tiene un gran

interés como apoyo a la toma de decisiones en los procesos de control y supervisión de la planta.

Es así como, se presentan 2 casos de estudio introducidos en los capítulos §14 y §19.

En primer lugar; una base de datos que proviene de una Planta Depuradora situada en la costa catalana. Se analiza una muestra de 396 observaciones con datos que fueron obtenidos en un período de un año y un mes; desde el 1 de Setiembre de 1995 al 30 de Septiembre de 1996, correspondiendo a mediciones medias de cada día. Cada observación incluye mediciones de las 25 variables que son relevantes (según la opinión del experto) para el funcionamiento de la planta. En los capítulos §15 y §16, se encuentra el análisis estadístico de los datos y de la clasificación jerárquica de la partición estudiada. Finalmente, se encuentra la aplicación de la propuesta metodológica en el capítulo §17 y la presentación y el análisis de los resultado obtenidos con la aplicación en el capítulo §18.

En segundo lugar; una base de datos proveniente de una planta depuradora de aguas residuales Eslovena. Los datos obedecen a la colaboración existente con el Department of Systems and Control Jozef Stefan Institute de Ljubljana Esovenia y en particular con el doctor Darko Vrecko, quien realiza investigación con esta Planta depuradora de aguas residuales. Se analiza una muestra de 365 observaciones (medias diarias) que fueron tomadas desde el 1 de junio de 2005 de el 31 de mayo de 2006. Cada observación incluye mediciones de las 16 variables que son relevantes (según la opinión del experto) para el funcionamiento de la planta piloto. La descriptiva estadística tanto univariante como bivariante se encuentra en el capítulo §20 y la clasificación jerárquica de la partición estudiada en el capítulo §21. Finalmente, se encuentra la aplicación de la propuesta metodológica en el capítulo §22 y la presentación y el análisis de los resultado obtenidos con la aplicación en el capítulo §23.

Cuarta Parte: Conclusiones, trabajo futuro y líneas abiertas. Para acabar en esta cuarta parte se presenta el apartado de conclusiones y las directrices del trabajo que queda abierto como continuación de éste estudio y se presenta en el capítulo §24.

Anexos. Por último los anexos asociados al análisis por casos de la revisión del *BbD* en el capítulo §B; a la aplicación del la Planta de la costa catalana en los capítulos §C, §D y §E; y los anexos asociados a la aplicación del la Planta eslovena en los capítulos §F, §G y §H.

Capítulo 2

Objetivos

El objetivo general de este trabajo es aportar en la interpretación automática de clasificaciones y avanzar en el ámbito de la construcción de sistemas integrales de Descubrimiento de Conocimiento en Bases de Datos (*Knowledge Discovery from Data, KDD*) que cubran todas las fases del proceso, en el sentido que las define Fayyad, incluida la generación automática de interpretaciones conceptuales y que combinen herramientas de Estadística e Inteligencia Artificial en forma cooperativa, de esta forma se quiere:

1. Establecer una metodología formal para la generación automática de descripciones conceptuales (representadas por predicados de primer orden CP_1) que interpreten un conjunto de clases construidas en dominios complejos, enmarcados en dominios poco estructurados de tal forma que se pueda aportar una sistematización al proceso de interpretación de clases procedentes de un cluster jerárquico, realizando hasta ahora de forma más o menos artesanal.
2. Realizar una propuesta bajo una aproximación metodológica híbrida de Estadística e Inteligencia Artificial que, de forma colaborativa, permita generar conocimiento explícito directamente a partir de las clases de manera que el experto pueda comprender más fácilmente las características principales del conjunto de clases obtenido.
3. Obtener un modelo conceptual que permita hacer contribuciones a la validación de clases de un clustering, en relación a su representación formal y la calidad de éstas, considerando que la calidad va ligada al grado de interpretabilidad (y/o utilidad) de las clases formadas y de ésta forma objetivizar los mecanismos de interpretación de clases procedentes de un cluster jerárquico.
4. Focalizar en una primera fase en variables numéricas puesto que la extensión a variables categóricas resultaría trivial en términos de su generalización a partir de la obtención del conjunto \mathcal{D}^k .
5. Aplicar la propuesta metodológica a dos casos de estudio en el contexto de plantas depuradoras de aguas residuales, para generar de forma automática caracterizaciones conceptuales en estos dominios poco estructurados, donde a partir de una base de datos y una partición de referencia previa (obtenida necesariamente por un clustering de estructura jerárquica) de los mismos, se genera una interpretación de las clases usando las variables recomendadas por el experto. Este sistema permitirá, para un nuevo objeto (día), establecer la clase (situación típica de la planta) que le corresponde y generar las caracterización e interpretación conceptual correspondiente a esa clase.

Capítulo 3

Estado del Arte

3.1 Introducción

Ya en la introducción se ha puesto de manifiesto que esta tesis se enmarca en un contexto marcadamente multidisciplinar y para centrarla es necesario introducir distintas áreas de investigación, algunas de las cuales integran varias áreas de conocimiento. Así, estructurar este estado del arte no ha sido simple por ser relevante introducir temas de amplio alcance y que algunas veces interseccionan entre si.

Se ha tratado de dar una visión global que permita situar el trabajo que aquí se presenta. Puesto que esta tesis trata de desarrollar una metodología de Knowledge Discovery from Data híbrida de Inteligencia Artificial y Estadística (*AI& Stats*) para generar interpretaciones automáticas de un cluster que apoyen un Sistema de Soporte a las Decisión (DSS, por sus siglas en inglés *Decision support system*), presentaremos primero las grandes áreas de Sistemas de Soporte a la Decisión (DSS), Knowledge Discovery from Data (KDD), *AI& Stats* y Clustering, para adentrarnos después en detalles complementarios de interés en el desarrollo de esta tesis.

3.2 Sistemas de soporte a la toma de decisiones

Debido a que hay muchos enfoques para la toma de decisiones y debido a la amplia gama de ámbitos en los cuales se toman las decisiones, el concepto de sistema de apoyo a las decisiones (DSS por sus siglas en inglés *Decision support system*) es muy amplio. Un DSS puede adoptar muchas formas diferentes. En general, podemos decir que un DSS es un sistema informático utilizado para servir de apoyo, más que automatizar, al proceso de toma de decisiones. La decisión es una elección entre alternativas basadas en estimaciones de los valores en términos de costos, riesgos, estabilidad o algún otro criterio pertinente de esas alternativas. El apoyo a una decisión significa ayudar a las personas que trabajan solas o en grupo a reunir inteligencia, generar alternativas y tomar decisiones. Apoyar el proceso de toma de decisión implica el apoyo a la estimación, la evaluación y/o la comparación de alternativas. En la práctica, las referencias a DSS suelen ser referencias a aplicaciones informáticas que realizan una función de apoyo, (Alter 1980).

El término sistema de apoyo a la decisión se ha utilizado de formas muy diversas y se ha definido de diferentes maneras dependiendo del punto de vista del autor (Druzdzel and Flynn 1999). En realidad no hay una definición universalmente aceptada de lo que es un DSS (Power 2002). Según (Keen 1978), un DSS *combina recursos intelectuales individuales con las capacidades de un ordenador para mejorar la calidad de las decisiones (son un apoyo informático para los encargados de tomar decisiones sobre problemas semiestructurados)*. Más

concretamente, los sistemas inteligentes de soporte a la decisión debe contribuir a la reducción de la incertidumbre que enfrentan los administradores cuando es necesario tomar las decisiones relativas a las futuras opciones en problemas donde la complejidad impide una decisión individual y es necesaria la conceptualización (Sánchez-Marré and et.al. 2008). Para Keen, el concepto de apoyo a las decisiones ha evolucionado desde dos áreas principales de investigación: los estudios teóricos de organización de la toma de decisiones, hechos en el Carnegie Institute of Technology a finales de 1950 y comienzos de 1960, y el trabajo técnico sobre sistemas informáticos interactivos, principalmente llevadas a cabo en el Instituto Tecnológico de Massachusetts en la década de 1960. Se considera que el concepto de DSS se convirtió en un espacio de investigación como tal a mediados de la década de 1970, antes de ganar en intensidad durante el decenio de 1980. A mediados y finales de 1980, los sistemas de información ejecutiva (EIS), los sistemas de apoyo a la decisión en grupo (GDSS) y los sistemas organizacionales de apoyo a la decisión (ODSS) evolucionaron desde el usuario individual y el DSS orientados a modelos. A partir de 1990 aproximadamente, los almacenes de datos y el procesamiento analítico en línea (OLAP) comenzó a ampliar el ámbito de los DSS. Como el cambio de milenio, se introdujeron nuevas aplicaciones analíticas basadas en la web.

Desde la aparición del concepto de Sistemas de Soporte a la Decisión (DSS) aunque de forma académica durante los años 60's, este concepto continuó su desarrollo hasta los años 80's donde ya se involucró el diseño de los sistemas de información para ejecutivos así como los sistemas de soporte a la decisión organizacionales (ODSS), para la década de los 90's las tecnologías de procesamiento analítico en línea y data warehousing son ya componentes de los DSS. Además que algunas técnicas basadas en aplicaciones web fueron adicionadas hacia finales de los 90's.

Es evidente que los DSS pertenecen a un entorno con fundamentos multidisciplinarios, incluyendo (pero no exclusivamente) la investigación en bases de datos, Inteligencia Artificial, interacción hombre-máquina, métodos de simulación, ingeniería de software y telecomunicaciones. Los DSS también tienen una débil conexión con el paradigma de la interfaz de usuario de hipertexto. Tanto el sistema PROMIS (para la toma de decisiones médicas)(Goldberg 1988) de la Universidad de Vermont, como el sistema ZOG/KMS (para la toma de decisiones militares y de negocios)(Robertson, McCracken, and Newell 1979) de la Universidad Carnegie Mellon fueron dos sistemas de apoyo a las decisiones que constituyeron grandes avances en la investigación de interfaz de usuario. Por otra parte, aunque las investigaciones en hipertexto, por lo general, se hayan entrado en la sobrecarga de información, algunos investigadores, en particular, Douglas Engelbart, se han centrado en la toma de decisiones en particular.

En la actualidad el uso de los DSS se ha extendido debido a su capacidad de analizar grandes volúmenes de datos procedentes de dominios muy complejos y a la forma que tienen de presentar en resumen esta información.

La información que típicamente puede recopilar y mostrar una aplicación de soporte a la decisión incluye todos los datos almacenados en la empresa u organización y que van desde sistemas heredados, hasta bases de datos relacionales, data warehouses, etc.

La información resultado se presenta de tal manera que sean de fácil comprensión aun para los usuarios que no están muy familiarizados con sistemas computacionales.

Una de las funciones más importantes de este tipo de sistemas es la proyección a futuro del comportamiento de algunos factores de negocio, basada en un modelo de sistema que ha sido diseñado de acuerdo con los expertos del área en la que se desarrolla la aplicación.

La tecnología de la información está inmersa en cambios cada vez más rápidos y los DSS no son la excepción, a pesar de su excelente funcionalidad. Uno de los principales problemas que presentan en la actualidad los DSS está en su diseño, ya que estos requieren de una considerable experiencia sobre cuestiones de explotación de datos y de un complejo análisis humano para optimizar sus tiempos de operación. La experiencia ha puesto de manifiesto

que existe un tipo de dominios de estructura particularmente compleja donde la construcción de DSS es especialmente difícil. En realidad, ya en los años 60, Simon (Simon 1960) clasifica los problemas según su nivel de dificultad (ver figura 3.1) y el hecho es que cuanto más desestructurado es un dominio más difícil es conseguir un buen DSS.

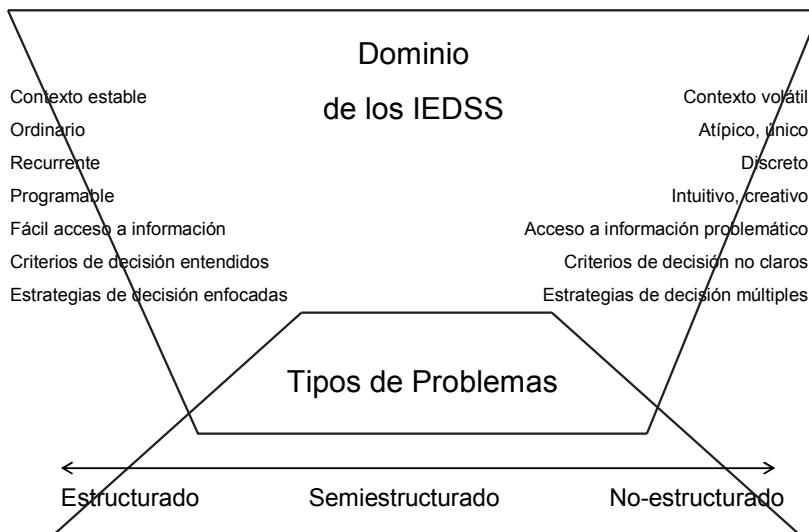


Figura 3.1: Clasificación de tipos de problemas propuesta por Simon en 1960.

En efecto, en las últimas décadas, los modelos matemáticos-estadísticos, numéricos y algoritmos de simulaciones se han utilizado como los medios adecuados para conocer el dominio objetivo, los problemas de gestión, etc. y proporcionar información útil para la toma de decisiones. Con este fin, un amplio conjunto de técnicas científicas se han aplicado a este tipo de problemas durante un largo tiempo y con buenos resultados. Sin embargo, la mayoría de estos esfuerzos se centraron en los problemas que podrían ser asignados al primer nivel de complejidad de Simon, Figura 3.1 (es decir, simple, sistemas de baja incertidumbre donde lo que se está tratando es de ámbito limitado o acotado)(Sàncchez-Marrè and et.al. 2008).

Cuando se trata de dominios de las más altas complejidades parece que el único modo de construir DSS solventes, es incorporar en el sistema *conocimiento* específico acerca del dominio concreto que se pretende apoyar. Según (Fox and Das 2000), un sistema de apoyo a la decisión es un *sistema que asiste a la toma de decisiones en la elección de alternativas entre las creencias o acciones mediante la aplicación de conocimiento acerca de la decisión sobre el dominio para llegar a recomendaciones para las distintas opciones*. Incorporar un procedimiento de decisión explícita basado en un conjunto de principios teóricos que justifique la *racionalidad* de este procedimiento es algo que se hace necesario y de este modo integrar la inteligencia a este proceso es el siguiente paso.

Así, un sistema inteligente de soporte a la decisión (Intelligence Decision Support System (IDSS))(Sojda 2002) utiliza una combinación de modelos, técnicas analíticas, de recuperación de la información y el conocimiento sobre el dominio necesario para ayudar a desarrollar y evaluar alternativas adecuadas (Adelman 1992), estos sistemas se centran en decisiones estratégicas y no operacionales y están orientados a reducir el tiempo necesario para tomar las decisiones en un dominio, y mejorar la coherencia y la calidad de las decisiones (Haagsma and Johanns 1994).

Como ya se ha dicho, para los dominios que se sitúan en el segundo y tercer nivel de complejidad, es necesario definir un marco donde tenga cabida el conocimiento sobre el dominio y así en (Sàncchez-Marrè and et.al. 2008) se propone la construcción de un IDSS integrando varios aspectos: métodos de Inteligencia Artificial, componentes de información geográfica

del sistema, técnicas matemáticas o estadísticas, ontologías, y algunos pequeñas componentes económicas (véase Figura 3.2 (Sàncchez-Marrè and et.al. 2006)).

En estos tipos de dominio crítico en los que las decisiones de gestión equivocadas pueden tener consecuencias desastrosas de tipo sociales, económicas y ecológicas como podrían ser los dominios médicos y medioambientales, el proceso de toma de decisiones asistida por IDSSs debe ser de colaboración, no contradictorio, y la toma de decisiones debe informar e involucrar a aquéllos que deben vivir con las decisiones y sus consecuencias.

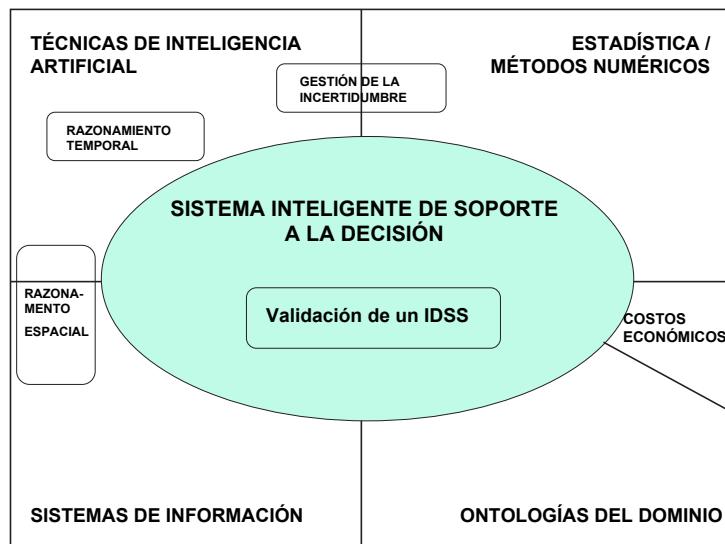


Figura 3.2: Componentes de un IDSS.

En (Cortés, Sàncchez-Marrè, Ceccaroni, R-Roda, and Poch 2000) se propone una estructura de 5 niveles o capas para los IDSS:

1. El primer nivel, incluye las tareas implicadas en la recolección de datos y su almacenamiento en bases de datos. Los datos originales son a menudo *defectuosos*, lo que hace necesario una serie de procedimientos de pre-procesamiento de los datos antes de que puedan ser almacenados de forma comprensible e interpretados. Los datos faltantes y la incertidumbre son factores que deben ser, también, consideradas en este nivel. Las técnicas específicas de análisis de datos corresponden también a éste nivel.
2. El nivel de diagnóstico, incluye los modelos de razonamiento que se utilizan para inferir la *situación del proceso*, de modo que se tenga una propuesta de acción razonable. Esto se logra con la ayuda de modelos estadísticos, numéricos y de Inteligencia Artificial que utilizarán el conocimiento adquirido en el nivel anterior.
3. El nivel de apoyo a la decisión, implica la recopilación y la fusión de las conclusiones derivadas de los modelos de conocimiento IA y de los modelos estadísticos. Este nivel también plantea la interacción de los usuarios con el ordenador a través de un sistema interactivo. Cuando no se puede alcanzar una conclusión clara y única deberá presentarse al usuario un conjunto ordenado de decisiones de acuerdo a su probabilidad de éxito o grado de certeza o algún otro criterio relevante (utilidad, corte del error, etc.).
4. En el cuarto nivel, se formulan planes y se presenta al usuario una lista general de acciones o estrategias sugeridas para resolver un problema específico.
5. El conjunto de acciones que se llevan a cabo para resolver el(es) problema(s) en el dominio que se consideran es el quinto. El sistema recomienda no sólo la acción, o

secuencia de acciones (un plan), si no que también una *solución* que ha de ser aceptada por el encargado de tomar las decisiones.

Un IDSS contribuye no sólo como un eficiente mecanismo para encontrar una solución óptima o sub-óptima, dado cualquier conjunto de preferencias, sino también como un mecanismo para hacer todo el proceso más abierto y transparente. En este contexto, un IDSS puede desempeñar un papel clave en la interacción de los seres humanos y los sistemas, ya que son herramientas diseñadas para hacer frente a la naturaleza multidisciplinaria y de alta complejidad de los problemas. Desde un punto de vista funcional, y teniendo en cuenta el tipo de problema que la IDSS resuelve, hay dos tipos de IDSS que podían distinguirse:

1. Los IDSS de *control-supervisión* de un proceso en tiempo real (o casi en tiempo real). Debe garantizar la robustez contra el ruido, la falta de datos, los errores tipográficos frente a cualquier combinación de datos de entrada. En general, el usuario final es responsable de aceptar-refinar-rechazar las soluciones propuestas por el IDSS. Esto puede disminuir la responsabilidad del usuario (por lo tanto, hay un aumento de la confianza en el IDSS) a lo largo del tiempo en la medida en que el sistema se enfrenta a situaciones que se han resuelto con éxito en el pasado (validación).
2. Los IDSS que dan apoyo a la toma de decisiones puntuales. Se utilizan principalmente para justificar decisiones multi-criterios (formular políticas transparentes para los usuarios) más que para tomar las decisiones en el día a día. Es interesante para el usuario final ya que da la posibilidad de jugar con los escenarios posibles, para explorar la respuestas y la estabilidad de la solución (cuán sensible es nuestra decisión a los pequeños variaciones del peso y del valor de las variables), etc. La confianza no aumenta de acuerdo a los resultados frente a situaciones similares, porque estos IDSS son muy específicos y, a veces, sólo se construyen para tomar (justificar) una decisión.

3.3 Knowledge Discovery from Data

Hace más de una década se estimó que la cantidad de información en el mundo se dobla cada 20 meses (Fayyad, Piatetsky-Shapiro, and Smyth 1996). Hoy se sabe que realmente la cantidad de información almacenada sólo en la red crece de forma continua y desmesuradamente (Baeza-Yates and Pino 2006). Según un informe de la consultora internacional IDC y el fabricante de sistemas de almacenamiento EMC (Gantz and et.al. 2007), en el 2007, la cantidad de información creada estuvo a punto de sobrepasar, por primera vez, la capacidad física de almacenamiento disponible. El estudio recuerda que, en 2006, la cantidad de información digitalizada fue 3 millones de veces mayor que la de todos los libros escritos sobre papel. En 2006, la cantidad de información digital creada, capturada y replicada fue de 1,288 x 1018 bits, esto es 161 exabytes o 161 billones de gigabytes; esto es más de lo generado en los 5000 años anteriores. Esto significa que, científicos, gobierno y sistemas de información corporativos están siendo inundados por una gran cantidad de datos que son generados y almacenados rutinariamente, los cuales aumentan las bases de datos a una velocidad vertiginosa. Estos volúmenes de datos rebasan los métodos manuales tradicionales de análisis de datos. De ahí que existe una necesidad imperativa y urgente de desarrollar una nueva generación de técnicas y herramientas con la capacidad de asistir *inteligente* y *automáticamente* a las personas en el análisis de grandes cantidades de datos para obtener conocimiento útil. Estas técnicas y herramientas son temas de un campo emergente *descubrimiento del conocimiento en base de datos* (KDD)¹(Fayyad 1996).

¹Del Inglés Knowledge Discovery from Data.

Así en 1989 se celebra en el seno del *IJCAI* (la *International Joint Conference on Artificial Intelligence*) el primer *Workshop on Knowledge Discovery of Data*. Siete años después Fayyad hace una de las definiciones más famosas de lo que se entiende por *Knowledge Discovery and Data Mining*:

“The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.”

(Fayyad, Piatetsky-Shapiro, and Smyth 1996).

Y la minería de datos se consolida rápidamente como un área de investigación también interdisciplinar donde se hace necesario combinar técnicas avanzadas de la Inteligencia Artificial, la Estadística, Sistemas de Información y Visualización para afrontar la obtención de conocimiento de bases de datos de dimensiones inimaginables antes del boom Internet (Gibert 2004). Según Fayyad, el término *Knowledge Discovery of Data* se acuña en 1989 para referirse a las aplicaciones de alto nivel que incluyen métodos particulares de *Data Mining* (Fayyad 1996). Es decir, KDD se situaría en un plano superior, combinando los métodos de *Data Mining* con otras herramientas para extraer *conocimiento* de los datos.

Obtener conocimiento de conjuntos de datos grandes o pequeños—y además, poco estructurados es una tarea muy difícil. La combinación de técnicas de análisis multivariante de datos (ej. clustering), aprendizaje inductivo (ej. sistemas basados en conocimiento), gestión de bases de datos y representación gráfica multidimensional, deberán producir beneficios en esta dirección y a corto plazo. La Figura 3.3 muestra un diagrama del proceso KDD.

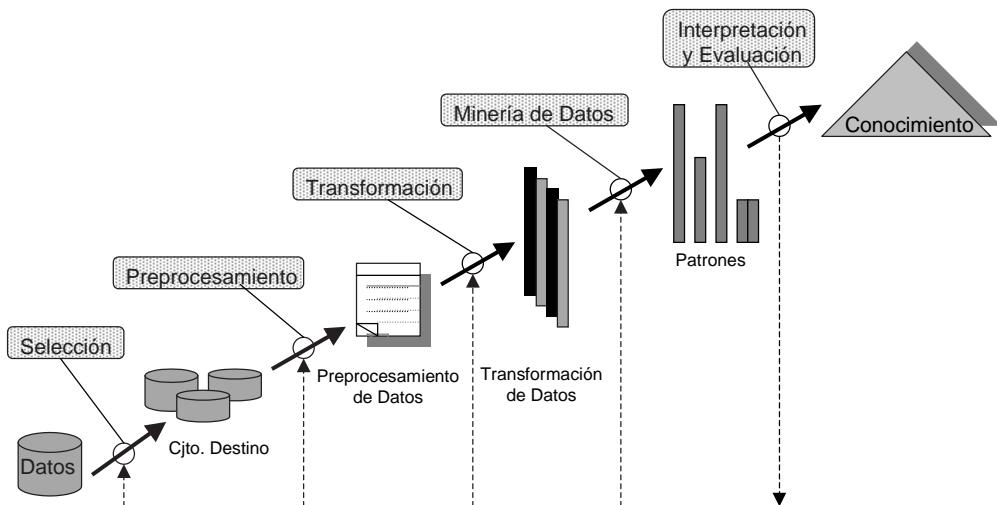


Figura 3.3: Diagrama del proceso KDD.

A continuación se resume cada una de las etapas (para detalles véase (Fayyad, Piatetsky-Shapiro, and Smyth 1996), (Fayyad 1996)):

1. Comprensión del dominio de aplicación, el conocimiento *a priori* relevante, y las metas del usuario final.
2. Creación de un conjunto de datos objetivo. Seleccionar un conjunto de datos, o seleccionar un subconjunto de atributos o muestra de datos, sobre los cuales se realizará el análisis.

3. Preparación y preprocesamiento de datos. Operaciones básicas, si fueran necesarias, como la eliminación de ruido, el tratamiento de datos atípicos (outliers) o faltantes, recabar la información necesaria para modelar el ruido, decidir sobre estrategias para manejar los outliers, etc.
4. Reducción y proyección de datos. Encontrar características relevantes para representar los datos depende de las metas del proceso. Usar técnicas de reducción de la dimensionalidad o métodos de transformación de variables para reducir el número de atributos bajo consideración o para encontrar representaciones invariantes para los datos.
5. Seleccionar el método concreto de minería de datos, con el que realizar el análisis propiamente dicho. Según el objetivo del proceso KDD será adecuado tratar los datos con técnicas de clasificación, regresión, clustering, optimización, razonamiento inductivo, etc.
6. Seleccionar el o los algoritmo(s) de minería de datos. Seleccionar la técnica que se emplearán en la investigación. Esto incluye decidir qué modelos y parámetros son los apropiados y escoger un método de minería de datos compatible con el criterio del proceso de KDD.
7. La minería de datos. Descubrimiento de conocimiento de patrones en una representación formal o un conjunto de representaciones como: reglas de clasificación o árboles, regresión, clustering y así sucesivamente. El usuario puede apoyar el método de minería de datos realizando correctamente los pasos previos.
8. Interpretación de los resultados obtenidos, posible retorno a cualquiera de los pasos previos del 1–7 para iteraciones posteriores.
9. Consolidación del conocimiento descubierto. Incorporación de este conocimiento al sistema, o simplemente documentarlo y reportarlo a las partes interesadas.

El proceso KDD es interactivo e iterativo y algunos autores hacen especial énfasis en la naturaleza interactiva del proceso (Brachman and Anand 1996). Involucra decisiones y elecciones complejas entre paso y paso del proceso.

También Fayyad (Fayyad 1996) señala que el proceso de KDD puede incluir *iteraciones* significativas y contener ciclos entre cualesquiera dos pasos; así en cada etapa el *minero de datos* puede volver a la etapa que requiera para continuar su trabajo. La etapa donde se produce propiamente la explotación de datos y se realiza el núcleo del descubrimiento de conocimiento es la denominada Minería de datos.

Según el objetivo del proceso de KDD las técnicas de *Data Mining* pueden ser muy diferentes y variar entre la simple descripción del dominio con fines estructuradores y comprensivos hasta la modelización con fines predictivos en toda su complejidad. También debemos decir que existen diversas herramientas informáticas comerciales que tratan algunas de las situaciones mencionadas (ej. Clementine, Intelligent Manager, SPAD (Lebart, Morineau, and Lambert 1994), SPSS (Visauta 1998), WEKA (Ian H. Witten 1999), DAVIS (Huh and Song 2002) entre otras son algunas de las más famosas hoy en día), las cuales presentan principalmente una combinación de técnicas existentes, permitiendo comparación de resultados.

Fayyad con la propuesta que se ha señalado anteriormente marcó el comienzo de una nuevo paradigma en la investigación del KDD (Gilbert, Spate, Sàánchez-Marrè, Comas, and Athanasiadis 2008):

“La mayoría de los trabajos previos en KDD, se centraban en [...] la etapa de Minería de Datos. Sin embargo, los otros pasos son de considerable importancia para el éxito de las aplicaciones de KDD en la práctica.” (Fayyad 1996).

Fayyad incluye en su propuesta las tareas pretratamiento e interpretación de los datos, además de la propia aplicación de algoritmos de minería de datos. Esto, de hecho, requiere un gran esfuerzo cuando se trata de aplicaciones reales. La depuración y transformación de datos, la selección de técnicas de minería de datos apropiadas y la optimización de los parámetros de ejecución de dichas técnicas (si es necesario), a menudo consume mucho tiempo y son tareas complejas, principalmente porque los enfoques deben adaptarse a cada una de las aplicaciones, y la interacción humana es obligatoria. Una vez que esas tareas se han realizado, la aplicación de algoritmos de minería de datos se convierte en trivial y puede ser automatizada, lo que requiere sólo una pequeña proporción de tiempo dedicado a todo el proceso de KDD. La interpretación de los resultados también a menudo consume mucho tiempo y requiere de intensa orientación humana.

De hecho, el análisis previo y posterior al Data Mining requiere un gran esfuerzo cuando se trata de aplicaciones reales.

El análisis previo es fundamental, debido principalmente a dos razones:

- Los conjuntos de datos reales tienden a ser imperfectos, contienen errores, *outliers*, hay datos que faltan, ruido y hace falta desarrollar nuevas herramientas, ya sea para detectar o corregir estos problemas.
- La aplicación de una determinada técnica de minería de datos podrá exigir condiciones específicas del conjunto de datos (sólo variables binarias, datos centrados, normalidad, sólo variables cualitativas, etc.) En este caso, se hacen necesarias herramientas para verificar que esas condiciones existen, así como para transformar los datos en la forma apropiada para la técnica de Data Mining que se vaya a utilizar.

A parte de la importante función de preprocesamiento, junto con la correcta selección de la técnica de minería de datos, hay un importante trabajo por hacer en cuanto a utilizar los resultados obtenidos en la fase de Data Mining para apoyar la toma de decisiones, es decir, asistir al usuario final en la *comprensión de los resultados*. De hecho, la calidad de las decisiones dependerá no sólo de la calidad de los resultados en sí, sino de la capacidad del sistema para *comunicar* los resultados en un forma comprensible para la toma de decisiones.

Los softwares comerciales utilizados para la aplicación de las técnicas de Data Mining suelen producir largos listados repletos de todos aquellos resultados que pueden llegar a ser útiles en una aplicación cualquiera. Cuanto más nos acercamos a paquetes estadísticos, más compleja es la salida y más información de carácter numérico contiene. Sin embargo, dado un caso real particular, no toda la información que se proporciona es útil. Es más, la mayor parte suele ser irrelevante. Por lo tanto, es importante (Gibert, Izquierdo, Holmes, Athanasiadis, Comas, and Sàncchez-Marré 2008):

- Identificar la información pertinente de la salida del software dependiendo de los objetivos de cada análisis particular. Por ejemplo, en un análisis de regresión, puede ser irrelevante para los expertos ambientales conocer el valor exacto de los índices h_i , pero a partir de esta información, se debe poder listar el conjunto de observaciones influyentes que deben ser cuidadosamente analizados por parte del experto.
- Buscar la mejor forma de presentar los resultados *seleccionados* al usuario de tal manera que puedan ser directamente comprensibles, habida cuenta que el usuario final no conoce los detalles técnicos del método de Minería de Datos utilizado. Así que, probablemente, a partir de los resultados de una regresión logística, es más interesante ofrecer la interpretación de los coeficientes estimados en lugar de la propia ecuación logística.

Considerando el caso particular en que la técnica de Minería de Datos aplicada haya sido el clustering, la mayoría de los software proporcionan información sobre el número de grupos descubiertos y el conjunto de objetos pertenecientes a cada uno de los grupos. Tras estos resultados, surge de inmediato la necesidad comprender los criterios subyacentes de agrupación, así como el significado de las propias clases. En los contextos de Minería de datos donde el número de clases y variables es elevado, las herramientas que ayuden al usuario a la conceptualización de las clases resultan extremadamente útiles. En (Gibert, Spate, Sàncchez-Marrè, Comas, and Athanasiadis 2008) se analizan las ventajas de utilizar el KDD para extraer conocimiento en dominios medio ambientales.

3.4 Inteligencia Artificial y Estadística (*AI& Stats*)

Durante la década pasada, en una gran variedad de dominios de aplicación, investigadores de las áreas de aprendizaje automático, teoría del aprendizaje computacional, reconocimiento de patrones y análisis de datos han hecho un esfuerzo por establecer un puente de comunicación y cooperación. La *Inteligencia Artificial y Estadística* es un área de investigación interdisciplinar cuyo origen podríamos situar en la fundación de la *Artificial Intelligence and Statistics Society (AI&Stats)*² por Douglas H. Fisher en 1985, en el seno del *First International Workshop on Artificial Intelligence and Statistics* que impulsó Bill Gale de los AT&T Laboratories. Desde entonces la conferencia internacional de dicha sociedad se ha venido celebrando bianualmente de forma ininterrumpida.

El principal objetivo de la *AI&Stats Society* es promover la comunicación entre la comunidad Estadística y la de la Inteligencia Artificial. Hace ya 15 años, en su introducción al primer volumen de las actas de la conferencia, Cheeseman y Oldford escribían:

“Nos parece que hay un potencial de desarrollo enorme en la intersección de la IA, la Ciencia de la Computación y la Estadística.”

(Cheeseman and Oldford (eds.) 1994).

Efectivamente existen algunas familias de problemas que han sido objeto de la Estadística y paralelamente de la Inteligencia Artificial, proponiendo cada una de estas disciplinas soluciones distintas para alcanzar un mismo objetivo (Gibert 2004). Entre éstos, uno de los más conocidos es el de la clasificación (*clustering*), que básicamente consiste en encontrar las clases en que se estructura un dominio dado (Everitt 1981).

Es claro que los objetivos de la *AI& Stats Society* encajan perfectamente en la interdisciplinariedad enmarcada en el KDD.

Además precisamente por el interés de esta cooperación, las técnicas de Inteligencia Artificial y Estadística constituye módulos fundamentales de los sistemas inteligentes de soporte a la decisión (IDSS), también con el objetivo de que integrando ambos módulos el sistema mejora notablemente (Sàncchez-Marrè and et.al. 2006) sus recomendaciones.

3.4.1 Estadística

El término **Estadística** se deriva del latín *Status*, que se refiere a política y situación social, al Estado, empieza como una ciencia de recolección de datos económicos y demográficos. En su evolución, y aún hoy en día, se considera una ciencia relacionada con la recolección y el análisis de datos, para extraer información y presentarla en forma comprensible y sintética y podemos situar sus orígenes en los grandes imperios de la antigüedad. Como se ha dicho, se puede situar el origen de la Estadística en la necesidad de las más antiguas sociedades de

²AI& Stats. Asociación para la Inteligencia Artificial y la Estadística.

humanas de enumerar sus individuos o los bienes disponibles. Ya en el Pentateuco, el libro de los Números del Antiguo Testamento, habla del censo; en este caso como algo maldito. El censo del emperador Yao en la China del año 2238 a. J.C., documentos asirios, egipcios y griegos, preceden a los más cercanos del Imperio Romano, donde la preocupación por la actividad de los recuentos de los individuos y bienes del Estado, tenían una clara intención tributaria y/o militar.

Podemos establecer que el objetivo principal de la Estadística es desde sus orígenes: *presentar de forma sintética y comprensible la información contenida en cualquier colección de datos* (Rodas, Gramajo, and Gibert 2000).

A partir de la correspondencia entre Pierre de Fermat y Blaise Pascal (1654) acerca del problema planteado por el de Chevelier de Mérè se fundamenta la teoría de la probabilidad, que ofrece un formalismo para modelar los procesos en los que interviene el azar que en realidad preocupan al hombre desde sus orígenes (en los tiempos antiguos los fenómenos aleatorios se interpretaban como manifestaciones de la voluntad divina; oráculos, sacerdotes y pitonisas utilizaban la configuración resultante de los juegos de azar, como tirar los dados, para predecir el futuro; se han encontrado dados en tumbas egipcias que datan del año 3000 a. C. (Gibert 2008)). Christian Huygens (1657) da el primer tratamiento científico que se conoce a la teoría de la probabilidad. El Ars Conjectandi (póstumo, 1713) de Jakob Bernoulli y la Doctrina de Posibilidades (1718) de Abraham de Moivre estudiaron la materia como una rama de las matemáticas. En la era moderna, el trabajo de Kolmogórov originalmente publicado en 1933 en alemán como (Kolmogórov 1933) y posteriormente en otras ediciones en inglés como (Kolmogórov 1956), ha sido un pilar fundamental en la formulación del modelo de la Teoría de Probabilidades que terminó con las divergencias entre la escuela laplaciana y frecuentista.

La teoría de la probabilidad evoluciona de forma independiente a la estadística hasta que surge la necesidad de tratar con muestras y recurre a ella para poder estudiar la parcialidad. Los primeros trabajos arrancan de la escuela inglesa del s. XVII y culminaran entrando el s. XX con Fisher como se comentará más adelante.

No es hasta el s. XIX que la Estadística amplía sus miras y se extiende fuera del ámbito censal, estatal o ligado al cálculo de impuestos y seguros y lo hace con el corpus Darwiniano como fondo. El primero en resaltar la necesidad de acudir a métodos estadísticos para contrastar la teoría de Darwin fue Francis Galton (1822-1911). La lectura de la obra de Darwin supuso una transformación radical en la vida de Galton que, casi a los 40 años, pasó a dedicar sus esfuerzos al estudio de la herencia humana y en 1877 presentó sus primeros trabajos sobre *Análisis de regresión*. La importancia de Galton radica, en gran parte, en la influencia que ejerció sobre W.R.F. Weldon (1860-1906) y Karl Pearson (1895-1980), Weldon abandona el camino de estudios embriológicos y morfológicos como medio de contrastar las hipótesis de Darwin y comienza a investigar la aplicación de los métodos estadísticos a la Biología animal; la resolución de tales problemas requería el desarrollo de métodos estadísticos mas avanzados que los existentes en la época y Weldon busca para ello la colaboración del matemático y filósofo Pearson. La colaboración de estos 2 autores y el apoyo de Galton va a constituir el impulso generador de la corriente de contribuciones que fundamentará la Estadística actual. Así en 1901 Pearson (1895-1980) presentó, entre otros trabajos, una versión preliminar del *Análisis de Componentes Principales*. Su principal discípulo Fisher (1890–1962), cuyos trabajos son considerados la base de la *Estadística moderna*, logra sistematizar un importante conjunto de ideas en relación con la *estimación* de parámetros a partir de muestras sentando las bases formales de la Inferencia estadística tal como se entiende ahora. Aunque las cuestiones tratadas por Fisher habían sido ya planteadas de forma imprecisa con anterioridad. Sus trabajos culminan con la publicación de *Statistical Methods for Research Workers*, publicado en 1925. En esta concepción de la Estadística entra en juego, formalmente, por primera vez, la teoría de la probabilidad, que se utiliza para acotar el error de estimación cuando en

lugar de trabajar con datos poblacionales se utilizan muestras.

Junto con Mahalanobis (Mahalanobis 1936), presentaron los primeros trabajos acerca del *Análisis Discriminante* en el cual existe una variable respuesta, que indica la clase de todo objeto y encuentra la mejor combinación lineal de todos los atributos para distinguir la clase.

En esa misma época los problemas de las Ciencias Sociales impulsan el desarrollo de métodos para analizar muchas variables conjuntamente (métodos multivariantes). En principio, se trata de medir factores no directamente observables. A raíz de las necesidades planteadas en este ámbito, Spearman (1904) desarrolla el análisis factorial en factores comunes y específicos y Harold Hotelling (1895–1973) lo generaliza al análisis de componentes principales. En 1931 Whishart introduce la matriz de covarianza muestral. Un año más tarde, Hotelling generaliza al caso multivariante la *t-studend* concibiendo la distribución que lleva su nombre (Hotelling 1931) y Sokal y Sneath presentan los trabajos formales en cluster que se detallan más adelante, pero no es hasta que la informática toma cuerpo como herramienta potente de cálculo que todos estos métodos multivariantes empiezan a ser realmente utilizados en aplicaciones reales. A partir de 1950 podemos considerar que comienza la época moderna de la Estadística que, con la aparición del ordenador digital, culmina en la más reciente rama de esta disciplina, la Estadística Computacional que ha abierto la Estadística a otro tipo de modelos no analíticos que permiten afrontar realidades cada vez más complejas. Asimismo en la segunda mitad del S.XX aparecen alternativas a los modelos clásicos como la estadística basada en datos difusos (Zadeh 1965), (Zadeh 1993) o la estadística con objetos simbólicos (Bock and Diday 1999), (Gowda and Diday 1992) por citar algún enfoque que pretende contribuir a modelar realidades cada vez más complejas.

La Estadística, al igual que otras ciencias, se está enfrentando actualmente a las consecuencias de una revolución tecnológica importantísima, cuyas repercusiones se sienten ya incluso a nivel social. Las autopistas de la información parecen estar dando paso a lo que se está denominando la Sociedad de la Información, y la estadística tiene que evolucionar muy rápidamente para adaptarse a las nuevas necesidades de esta sociedad. Los datos surgen constantemente por todos lados y en cantidades ingentes. Se requiere la colaboración de otras ciencias como la Inteligencia Artificial o Sistemas de Información, para seleccionar y organizar el inmenso *corpus* y se hace necesario un análisis inteligente que seleccione resultados oportunos, no obsoletos, con mucha agilidad.

La Estadística se está integrando, como pilar fundamental del KDD ofreciendo soporte a las etapas de depuración y Data Mining donde se requieren modelos descriptivos o predictivos básicamente construidos sobre variables numéricas aunque también aporta en el análisis de variables categóricas y que evoluciona independientemente de la Estadística hasta bien entrado el s. XIX, así pues la Estadística es una ciencia con una larga historia, cuyo desarrollo se debe, en parte, a la interacción con otras ciencias experimentales, que actualmente sigue siendo objeto de una investigación intensa y, en algunas áreas, bastante fecunda (Gibert 2008).

3.4.2 Inteligencia Artificial

La Inteligencia Artificial (IA) es una disciplina formal que surge a mediados de los años 50's con el objetivo de emular el comportamiento inteligente de los humanos en la resolución de problemas difíciles.

Las ideas más básicas se remontan a los griegos, antes de Cristo. Aristóteles (384-322 a.C.) fue el primero en describir un conjunto de reglas que describen una parte del funcionamiento de la mente para obtener conclusiones racionales, y Ktesibios de Alejandría (250 a.C.) construyó la primera máquina autocontrolada, un regulador del flujo de agua (racional pero sin razonamiento). En 1315 d.C Ramón Lull, en su obra *Ars Magna* (mechanización del

razonamiento) tuvo la idea de que el razonamiento podía ser efectuado de manera artificial. En 1943 Warren McCulloch y Walter Pitts presentaron su modelo de neuronas artificiales, el cual se considera el primer trabajo de redes neuronales, aún cuando todavía no existía el término. Los primeros avances importantes en Inteligencia Artificial comenzaron a principios de los años 1950 con el trabajo de Alan Turing.

En 1956 John McCarthy organiza una reunión en la escuela de verano de Dartmouth a la que asisten Mavrin L. Minsky (Harvard University), Nathaniel Rochester (I.B.M. Corporation), Claude E. Shannon (Bell Telephone Laboratories), Ray Solomonoff, Oliver Selfridge, Trenchard More, Arthur Samuel, Herbert Simon y Allen Newell. La reunión se basa en la conjectura de que cada aspecto del aprendizaje y cada característica de la inteligencia podían ser tan precisamente descritos que se podían crear máquinas que las simularan. El encuentro, ahora conocido como la *conferencia de Dartmouth*, duró un mes y se llevó a cabo con tal éxito que se considera esta conferencia como la conferencia fundacional de la Inteligencia Artificial como disciplina. De Andrés (De Andrés Argente 2002) refiere con una visión muy crítica que en esta reunión se hicieron previsiones triunfalistas a diez años vista que jamás se cumplieron, lo que provocó el abandono casi total de las investigaciones durante quince años .

Al inicio la Inteligencia Artificial se situó bajo el paradigma de von Neumann que se basaba en el concepto de *programa almacenado* lo que permitía la lectura de un programa dentro de la memoria del ordenador, y después la ejecución de las instrucciones del mismo sin tener que volverlas a escribir (el primer ordenador en usar el citado concepto fue la llamada EDVAC (Electronic Discrete-Variable Automatic Computer) de von Neumann, y en las técnicas de computación secuencial, en los que su característica es usar procesos de búsqueda en un espacio de soluciones para construir máquinas que *piensen* (Baek 1997).

En 1961, Minsky (Minsky 1961) divide la IA en cinco tópicos: búsqueda, reconocimiento de patrones, aprendizaje, planificación e inducción. La mayoría de los trabajos de esta época (años 60) estuvieron relacionados con búsqueda heurística, pero mostraron serias limitaciones cuando se trató de hacer frente a problemas reales y complejos porque la mayoría de los problemas de los que se ocupa la Inteligencia Artificial son NP-completos. Los problemas NP (o no Polinómicos) son aquéllos problemas que no se pueden resolver en un tiempo acotado por una función polinómica; los problemas NP-completos descansan sobre espacios de búsqueda aún más complejos y tienen un tiempo de resolución combinatorio que se dispara de forma descomunal apenas las dimensiones del problema crecen ligeramente y así la búsqueda heurística se ocupaba de hallar formas adecuadas de acotar este espacio de búsqueda para alcanzar una solución apropiada (aunque no sea óptima) en un tiempo razonable (Garey and Johnson 1979).

Uno de los primeros éxitos de la aplicación de la Inteligencia Artificial se sitúa en el ámbito de los Sistemas Expertos, una rama del aprendizaje donde se pretende que un sistema razoné sobre un dominio como lo haría un experto. La mayor parte de los trabajos en esta área se centran en los sistemas orientados al diagnóstico (MYCIN-1976, diagnosis de infecciones pulmonares (Shortlife 1976)). Orbitan alrededor de la construcción de Sistemas Expertos áreas como *representación del conocimiento*, *aprendizaje automático*, *razonamiento automático*, *procesamiento de lenguaje natural*, etc. Sin embargo las representaciones simbólicas muestran limitaciones ante problemas reales y complejos principalmente porque a parte de descansar sobre los espacios de búsqueda NP-completos, transferir al sistema una descripción completa del dominio es sólo abordable en dominios acotados, muy estructurados y muy específicos. El gran cuello de botella es el conocimiento implícito del experto. Así, la Ingeniería del Conocimiento se estructura en desarrollar mecanismos que ayuden al experto a explicitar su conocimiento implícito. Sin embargo, entrados los años 80 la Inteligencia Artificial sufre una gran crisis al constatar que es prácticamente imposible capturar todo el *expertise* de un dominio y los métodos conexiónistas (como las redes neuronales) tratan

de emular el comportamiento inteligente ya sin pretender imitar el proceso cognitivo del ser humano.

El problema que aborda la Inteligencia Artificial es uno de los más complejos. Actualmente se está tan lejos de cumplir la prueba de Turing como cuando se formuló: *Existirá Inteligencia Artificial cuando no seamos capaces de distinguir entre un ser humano y un programa de computadora en una conversación a ciegas*.

Las principales críticas a la Inteligencia Artificial tienen que ver con su incapacidad de imitar por completo a un ser humano los Sistemas Expertos se basan en la descripción del dominio y el motor de inferencia.

En realidad, cuando las personas hablamos acerca de un sistema del mundo real, lo hacemos en tres etapas:

- Seleccionamos un conjunto de variables que caracterizan un conjunto de entidades bien diferenciadas. Tales variables pueden estar directamente vinculados a la experiencia sensorial—y entonces expresadas en una manera informal—o pueden estar determinadas por medio de procedimientos de mediciones más precisas.
- Establecemos las relaciones entre los atributos, ligando sus estados particulares. Esto en realidad se hace dando reglas como *Si (hecho A) entonces (hecho B)*, donde cada hecho describe un estado o un valor preciso de algún atributo particular.
- Finalmente, hay una tercera etapa donde los conjuntos de reglas se organizan para construir una teoría o un modelo que describe el sistema del mundo real bajo estudio.

El sistema no está bien construido cuando su teoría nos conduce a conclusiones contradictorias o a enunciados experimentalmente falsos acerca del sistema. En este contexto el término *inferencia* se aplica a cualquier algoritmo que se use para derivar consecuencias de hechos conocidos dentro del modelo. La inferencia en un amplio sentido puede aparecer en diferentes formas dependiendo del contexto considerado, desde la manipulación simbólica en una base de datos lógica hasta la evaluación de una función numérica o vectorial (López de Mántaras 1990). En el contexto de la Inteligencia Artificial es la inferencia lógica lo que se utiliza más frecuentemente.

En un sistema de inferencia lógica clásico se usan reglas de la forma *Si–Entonces* que evalúan a *cierto* o *falso*. Los operadores se definen por tablas de verdad.

La lógica como base para el razonamiento puede distinguirse por sus tres componentes principales (independientes del contexto): valores de verdad, vocabulario (operadores) y reglas de razonamiento (tautologías, silogismos). En la lógica de Boole, los valores de verdad son 0 (falso) o 1 (verdadero) y por medio de estos valores de verdad, se define el vocabulario vía las tablas de verdad.

La lógica es conocida como una de las ciencias más antiguas, tanto es así que se le atribuye a Aristóteles la paternidad de esta disciplina. En un principio se llamó Analítica, en virtud del título de las obras en que trató los problemas lógicos. Más tarde los escritos de Aristóteles relativos a estos eventos fueron recopilados por sus discípulos con el título de *Organon*, por considerar que la lógica era un instrumento para el conocimiento de la verdad. Aristóteles se planteó cómo es posible probar y demostrar que un conocimiento es verdadero, es decir, que tiene una validez universal. Aristóteles encuentra el fundamento de la demostración en la deducción, procedimiento que consiste en derivar un hecho particular de algo universal. La forma en que se afecta esa derivación es el silogismo, por cuya razón la silogística llega a ser el centro de la lógica aristotélica.

Normalmente la lógica clásica ante descripciones no completas de un fenómeno llega a aserciones que no son comunes del ser humano y ello impulsó una fértil área de investigación

relacionada con las lógicas formales extendidas (Mamdani, , and Gaines 1981), todas ellas de semántica y sintaxis bien definida, y una alta capacidad expresiva y con propuestas diferentes de derivar conclusiones que pueden incluso incorporar la incertidumbre en el proceso de razonamiento. En estos paradigmas se incluye una medida de la incerteza de las aserciones en el modelo y se define cómo propagar dicha incerteza a lo largo del razonamiento (razonamiento aproximado (Zadeh 1973)). Entre ellas podemos citar: la lógica modal (Hughes and Cresswell 1968) distingue entre verdad necesaria y posible; la lógica no monótona y la lógica temporal (McDermott 1982) que distingue entre enunciados que fueron verdaderos en el pasado y aquellos que serán verdaderos en el futuro; la lógica difusa que maneja un continuo de grados de certeza asociados a etiquetas lingüísticas (Zadeh 1965). La lógica epistémica (Brachman and Anand 1996) trata del conocimiento y las creencias, la lógica déontica (Ris 1971) con lo que debe hacerse y que permite ser verdadero.

Otra alternativa es la lógica probabilística donde la teoría de la probabilidad es el modelo de soporte para la incertidumbre (Nilsson 1986). Existen varias propuestas para *integrar* lógica y probabilidad incluyendo entre otras: Modelos relacionales probabilistas (Getoor, Friedman, Koller, Pfeffer, and Taskar 2007), Programas lógicos-probabilistas (Haddawy, Restificar, Geisler, and Miyamoto 2003), Lógica de alternativas independientes (Poole 1997), Redes bayesianas con nodos lógicos (Morales 2008).

Ésta es todavía un área activa de investigación y aún no se ha encontrado un marco formal definitivo para imitar el razonamiento humano.

A pesar de todo, parece claro que por más sofisticado que sea el motor de inferencia (el procedimiento de razonamiento formal) siempre hay parte del conocimiento del dominio que se quedó en el tintero y conduce al sistema a inducciones falsas por muy formalmente correctas que resulten. Actualmente (Geffner 1992) existe toda una corriente que trabaja en la búsqueda de un modelo de razonamiento que pueda asumir la existencia de conocimiento implícito y de bloques de razonamiento inconsciente como partes propias del modelo y se está resolviendo el diseño de mecanismos generalistas de resolución de problemas que sean independientes del dominio concreto de aplicación. La metáfora de los agentes racionales (Shoham 1993), (Jeiss. 1999), pretende construir modelos de dominios complejos a base de redes de agentes independientes que interactúan entre si y, conociendo cada uno parte del funcionamiento del dominio, de acuerdo con su propio rol en el mismo, la propia interacción entre ellos define la evolución del sistema.

Por otro lado, existen grupos que trabajan desde un enfoque más multidisciplinar tratando de construir metodológicas híbridas que combinen técnicas de razonamiento inductivo con otras áreas de conocimiento (como la estadística) con el fin de evitar la asunción de completitud en la descripción del dominio a cambio de realizar parte de la inducción por otros medios (Gibert 2004). Este enfoque ha mostrado gran potencial y resultados prometedores en diversas pruebas reales y en el ámbito donde se sitúa esta tesis (Gibert and García-Rudolph 2008).

3.5 Clustering

Sección aparte merece el clustering por las características particulares de esta tesis.

Clustering es un término usado para denotar un gran número de técnicas que intentan determinar grupos o *clusters* existentes en un conjunto de datos.

Se realiza a partir de un conjunto de casos pertenecientes al dominio en estudio del que a priori se desconoce la estructura y precisamente lo que se pretende es identificar las clases que la componen. En la mayoría de ocasiones, ni siquiera se conoce el número de clases. Básicamente, se traduce en encontrar agrupamientos o descubrir grupos y subgrupos que

revelean la naturaleza de la estructura del dominio.

En realidad, las técnicas de *clustering* son las más populares a la hora de separar datos en grupos y una de las técnicas de Minería de Datos más utilizadas(Fayyad 1996). Es más, la *clasificación* está entre las tres técnicas básicas (junto a la diferenciación de la experiencia en objetos particulares y sus atributos y la distinción entre un todo y sus partes) que dominan el pensamiento humano en su proceso de comprensión del mundo. De hecho, nosotros coincidimos (Gibert 2004) en que un buen número de aplicaciones reales en *KDD* o bien requieren un proceso de clasificación o son reducibles a él (Nakhaeizadeh 1996). Y precisamente por eso el clustering ha sido uno de los objetivos principales del grupo de investigación en que se inserta esta tesis.

Si bien es cierto que el hombre clasifica por naturaleza desde siempre, no es hasta bien entrado el siglo XX que aparece el primer tratado magistral, dentro del ámbito estadístico, que aborda la clasificación desde un punto de vista formal. En la *Principles of Numerical Taxonomy* de Sokal y Sneath (Sokal and Sneath 1963) se sientan, por primera vez, las bases algebraicas de las técnicas estadísticas de *clustering*, que parten de una matriz de datos donde todas las variables son, en principio, numéricas³. Probablemente por ser ésta una actividad inherente a la mente humana, desde sus principios también la Inteligencia Artificial se ha ocupado de mimetizar los procesos de clasificación. Así, Michalski (Michalski 1980) inicia con la *clasificación conceptual* una línea de métodos de clasificación basados en la generalización de conceptos, entendida en un sentido más o menos amplio, que parten de una matriz de datos donde todas las variables (atributos en el contexto de la Inteligencia Artificial) son categóricas (cuantitativas). En cuanto al hecho de clasificar matrices de datos mixtas, ya Anderberg (Anderberg 1973), propone tres estrategias principales:

- El *particionamiento* de variables, dividiéndolas por tipos y reduciendo el análisis al tipo dominante (si el es numérico, análisis de correspondencias seguido de un clustering sobre las componentes factoriales (Völle 1985), (Lebart 1990), lo que produce clases en un espacio ficticio de difícil interpretación). Entre otras cosas, esta aproximación, pierde la información los grupos no dominantes.
- Realizar la *conversión* de todas las variables a un único tipo, conservando el máximo de información posible. En Estadística, tradicionalmente, las variables numéricas se convierten a grupos de variables binarias, generando la *tabla de incidencia completa*. Con ella se puede realizar un clustering en la métrica de χ^2 (Dillon and Goldstein 1984). Las dimensiones de dicha tabla hacen la clasificación muy costosa. En Inteligencia Artificial, agrupar los valores de las variables numéricas en símbolos es lo más habitual (Roux 1985). Ello implica la pérdida relevante de información, así como la introducción de un sesgo que incide en los resultados, totalmente dirigidos por la forma como se realice la codificación (Gibert, Sonicki, and Martín 2001).
- El uso de medidas de *compatibilidad* que cubran las distintas combinaciones de tipos de variables; la idea es permitir el clustering en matrices heterogéneas sin necesidad de aplicar transformaciones previas sobre las propias variables. Pasa por la definición de distancias (o disimilaridades) entre individuos que utilicen expresiones diferentes según el tipo de la variable. En la literatura se hallan diferentes propuestas en esta dirección Gower en 1971 (Gower 1971), Gowda & Diday en 1992 (Gowda and Diday 1992), Gibert en 1991 (Gibert and Cortés 1992), (Gibert 1991), Ichino & Yaguchi en 1994 (Ichino and Yaguchi 1994), Ralambondrainy en 1995 (Ralambondrainy 1995b), Ruiz-Schulcloper (Ruiz-Shulcloper *et al.* 1996).

³Existen formas de cambiar de espacio métrico para trabajar con variables qualitativas.

La clave del asunto es elegir, de entre todas las clasificaciones posibles que se pueden construir sobre un conjunto de objetos la mejor en relación a un cierto criterio. En esencia un proceso de clustering se podría formular como la construcción de todas las clasificaciones posibles, la evaluación sobre cada una de un cierto criterio de calidad y la selección de aquélla partición que lo maximice. Es obvio que nos hallamos ante un problema NP-completo. Así, ya desde la Estadística o la Inteligencia Artificial los métodos de clustering consisten básicamente en la definición de distintos heurísticos para acotar la búsqueda, evitando así la construcción de todo el espacio de búsqueda de dimensiones prohibitivas. La naturaleza del heurístico y el criterio que permite comparar distintas clasificaciones es lo que cambia de un método a otro y según (Gibert 2004) es de naturaleza más algebraica en los métodos estadísticos y más lógica en los de la Inteligencia Artificial .

El principal problema para desarrollar métodos de clasificación automática es que el concepto de *cluster* no es fácil de definir. Algunas aproximaciones para definir un *cluster* pueden basarse en sus propiedades como: máxima cohesión interna y máximo aislamiento externo, propiedades propuestas por (Cormack 1971) y (Gordon 1980). Además, las clases pueden presentar formas y magnitudes muy diferentes y se puede entender la dificultad de que exista una definición general de *clusters* que los incluya a todos.

Como ya hemos dicho, la clasificación es uno de los problemas que han sido objeto de la Estadística y la Inteligencia Artificial simultáneamente. En efecto ambas disciplinas proporcionan diferentes métodos para descubrir cuáles son las clases subyacentes a un dominio. En las próximas secciones presentamos detalles de algunas de estas técnicas.

3.5.1 Clustering en la estadística

Los tratados de Sokal y Sneath ya determinan una formulación genérica de los métodos de clasificación donde el criterio de agregación está parametrizado y da lugar a los distintos métodos (distancias mínimas, distancias máximas, distancias medias, etc.)(Sokal and Sneath 1963) . Intentos de formular modelos matemáticos más precisos y rigurosos que den soporte teórico al análisis de clasificación son, entre otros, los trabajos de Johnson en 1967, Wolfe en 1970, Jardine & Sibson en 1971 (Jardine and Sibson 1971), Hartigan 1975 (Hartigan 1975), Diday 1979 (Diday 1979) y Benzecri en 1973 (Benzécri 1973).

Por descansar la mayoría de éstos métodos en la elección de una distancia o medida de (dis)similaridad, sobre el espacio de las variables gran parte de las investigaciones se han centrado en la definición de diferentes medidas de proximidad y desarrollo de algoritmos eficientes para utilizarlas. Podemos encontrar compilaciones de estos medidas en (Anderberg 1973), (Everitt 1981), (Romesburg 1990), (Jain, Dubes, and Chen 1987), (Dubes and Jain 1980).

Actualmente, y considerando la forma como se estructura el conjunto de individuos, podemos agrupar las heurísticas (Gibert 1994) existentes en:

- Métodos de particiones. Se busca la partición óptima del conjunto que se estudia en un número prefijado de classes k . Hay de dos tipos:
 - Métodos de particiones directas: Las clases que se forman serán disjuntas, y pueden ser aglomeradas o divisivas. Los primeros, parten de una nube de puntos inicial y van haciendo fusiones progresivas hasta obtener una única clase. Entre ellos se puede destacar el Método de los *centros móviles* (Forgy 1962), que es el caso particular del de *nubes dinámicas* de (Diday 1971) o el *k-means* de (MacQueen 1967). Otros, difieren en la elección de los centros de clases iniciales. Así, el *Isodata* de (Ball and Hall 1965) involucra un set de parámetros que pilotan la partición. Los métodos divisivos o descendentes, inicialmente escinden la nube

total de individuos generalmente de forma dicotómica y proceden a subdividir cada parte, hasta obtener un número suficiente de clases. Mucho menos utilizados que los aglomerativos y a menudo poco precisos (Volle 1985).

- Métodos de particiones en clases solapadas: Las clases pueden solaparse, es decir, un mismo objeto puede pertenecer simultáneamente a más de una clase. Estos métodos fueron introducidos por Jones y Needhan en la *Cambridge Language Research Unit* y también (Jardine and Sibson 1971) han intentado formalizar esta línea.
- Métodos de clasificación jerárquica. Se busca el árbol que refleja la estructura jerárquica de los datos. Según el nivel por el que se corte el árbol se obtendrá una partición más o menos precisa del conjunto objeto de estudio. Una ventaja respecto al anterior método es que no hace falta avanzar el número de clases que se quiere obtener al final si no que una vez construida la taxonomía se puede decir a qué nivel se *corta horizontalmente*. El usuario es quién decide cuantas clases hacer, pero el método le proporciona información de cómo son los datos, lo que le ayudará a tomar esta decisión.
- Métodos de clasificación piramidal. Introducidos por Diday, generalizan el concepto de jerarquía y permiten que un nodo pueda tener dos padres simultáneamente (Diday, Brito, and Mfoumou 1993).
- Métodos de clasificación con datos simbólicos. Propuesto por (Gowda and Diday 1992), que generaliza los métodos clásicos al uso de objetos simbólicos, los cuales describen las clases a nivel intensional y son fácilmente interpretables aunque, como se discutirá más adelante, nuestra propuesta presenta algunas ventajas respecto a este enfoque.

El Análisis de Datos Simbólicos parte de la necesidad de aprovechar en el análisis mismo la valiosa información cualitativa tan frecuentemente disponible. La respuesta de la estadística clásica era que no se podía cuantificar dicha información, y por lo tanto, no era analizable (Ruiz-Shulcloper, Chac-Kantún, and Martínez-Trinidad 1997). El análisis de Datos Simbólicos propone una teoría para superar esta barrera basada en el concepto de objeto simbólico que es en realidad una representación intensional de un conjunto de objetos y que sirve también como de modelo de incertezza o de imprecisión de los datos originales.

En Análisis de Datos Simbólicos se estudian conjuntos de datos de más alto nivel, donde ya no se focaliza sobre los individuos aislados sino sobre objetos simbólicos que representan agrupaciones de individuos, en definitiva, conceptos. Un objeto simbólico es una descripción intensional de una clase de objetos elementales que constituye su extensión. Entonces, el objetivo es reemplazar los individuos clásicos del análisis de datos tradicional por estos objetos simbólicos, más complejos y más aptos para representar conocimiento, porque están definidos en intención, utilizando el poder de la lógica. Los objetos simbólicos se expresan bajo forma de conjunción y generalizan las propiedades individuales de todas las instancias cubiertas por el objeto simbólico en forma maximal (utilizando todas las variables de la matriz de datos).

Cada variable puede tomar valores múltiples ya sean conjuntos discretos de modalidades o rangos numéricos para un mismo objeto simbólico en función del tipo de variable considerado. El objeto, {categoría = {obrero, empleado}, edad = [30, 40]}, tiene por extensión todos los objetos elementales en los cuales la categoría es ya sea obrero ya sea empleado y que esté entre 30 y 40 años de edad. Así, la Teoría del Análisis de Datos Simbólicos extiende el análisis de datos clásico al estudio de objetos simbólicos.

Por su estructura fundamentalmente conjuntiva un objeto simbólico es autointerpretable. En un contexto de clustering jerárquico el corte horizontal de un árbol daría lugar a un cierto conjunto de objetos simbólicos, uno para cada clase, que se debería poder interpretar directamente. Cada uno constituiría una descripción intensional del conjunto de elementos de la clase, descripción conjuntiva, con todas las variables y ello que puede solapar con la descripción de otras clases de la misma partición.

- Métodos de árboles aditivos. La distancia entre individuos se calculan por adición de las longitudes de las aristas que los unen. La nube de puntos queda estructurada en unos diagramas en forma de estrella que pueden dificultar la interpretación (Roux 1985).
- Métodos de clases latentes. En éstos métodos la variable que indica la clase de cada objeto se considera una variable latente. El método consiste en estimar la probabilidad con que cada elemento puede pertenecer a cada clase (McCutcheon 1987).

Clustering jerárquico

El clustering jerárquico es una herramienta exploratoria diseñada para revelar las agrupaciones naturales (o los conglomerados o clusters) dentro de un conjunto de datos que no sería de otra manera evidente. Es el más útil cuando usted desea agrupar un número pequeño (menos que algunos cientos) de objetos. Los objetos en análisis cluster jerárquico pueden ser casos o variables, dependiendo de si usted desea clasificar casos o examinar relaciones entre las variables.

El Clustering ascendente Jerárquico comienza separando cada objeto en un cluster por sí mismo. En cada etapa del análisis, el criterio por el que los objetos son separados se relaja en orden a enlazar los dos conglomerados más similares hasta que todos los objetos sean agrupados en un árbol de clasificación completo. El criterio básico para cualquier agrupación es la distancia. Los objetos que estén cerca uno del otro pertenecerán al mismo conglomerado o cluster, y los objetos que estén lejos uno del otro pertenecerán a distintos clusters. Para un conjunto de datos dado, los clusters que se construyen dependen de nuestra propia especificación de los siguientes parámetros:

1. El método cluster define las reglas para la formación del cluster. Por ejemplo, cuando calculamos la distancia entre dos clusters, podemos usar el par de objetos más cercano entre clusters o el par de objeto más alejados, o un compromiso entre estos métodos.
2. La medida define la fórmula para el cálculo de la distancia. Por ejemplo, la medida de distancia Euclídea calcula la distancia como una línea recta entre dos clusters. Las medidas de intervalo asumen que las variables están medidas en escala; las medidas de conteo asumen que son números discretos, y las medidas binarias asumen que toman dos valores.
3. La estandarización permite igualar el efecto de las variables medidas sobre diferentes escalas.

3.5.2 Clustering en la Inteligencia Artificial

A pesar de los poco más de cincuenta años de investigación y desarrollo en este campo, el problema general de *reconocimiento de patrones* en la Inteligencia Artificial con una orientación, ubicación y escalamiento no se ha resuelto, esto es, no se ha conseguido un diseño de un reconocedor de patrones automático de propósito general.

El diseño de un sistema de reconocimiento de patrones incluye los siguientes dos aspectos:

1. Adquisición de datos y preprocesamiento.
2. Representación de datos.

El dominio del problema sugiere la selección de los sensores, la técnica de preprocesamiento, el esquema de representación y el modelo de toma de decisiones. Generalmente un problema de reconocimiento bien definido y suficientemente delimitado (pocas variaciones intra clases y muchas variaciones inter clases) conducen a una representación compacta de patrones y a una estrategia simple de toma de decisiones. Por lo que, ninguna aproximación por sencilla que sea será la mejor ya que se han de utilizar diferentes técnicas y métodos. En consecuencia, la combinación de éstos es una práctica de uso común en el diseño de sistemas híbridos de reconocimiento de patrones (Fu 1983).

La literatura sobre el reconocimiento de patrones es extensa y dispersa encontrándose en numerosas revistas de diferentes disciplinas (ej. estadística aplicada, aprendizaje automático, redes neuronales y procesamiento de señales e imágenes). Un rápido vistazo de la tabla de contenidos de todos los temas de la *IEEE Transactions on Pattern Analysis and Machine Intelligence*, desde su primera publicación en enero de 1979, revela que aproximadamente 350 artículos tratan con el reconocimiento de patrones. Aproximadamente 300 de estos artículos cubren la aproximación estadística y pueden ser categorizados en los subtemas siguientes: Problema de dimensionalidad (15), Reducción de la dimensionalidad (50), Diseño de clasificadores (175), Combinación de clasificadores (10), Estimación del error (25), Clasificación no supervisada (59).

Además los excelentes libros de Duda y Hart (Duda and Hart 1973), Fukunaga (Fukunaga 1990), Devijver y Kittler (Devijver and Kittler 1982), Devroye, Gyorfi y Lugosi (Devroye and Lugosi 1996), Bishop (Bishop 1995), Ripley (Ripley 1996), Schurmann (Schuhfried 1992) y McLachlan (McLachlan 1992), Nagy (Nagy 1968) y Kanal (Kantrowitz 1994) en 1974 entre otros investigadores han contribuido notablemente al estado del arte de este tema.

Clasificación Conceptual

Alrededor de los años 80 Michalski (Michalski 1980) se interesa por formalizar la clasificación desde un punto de vista totalmente diferente. La idea de mimetizar el proceso de clasificación realizado por un humano, de naturaleza inherentemente cualitativa, y el interés por introducir consideraciones sobre la relevancia de los atributos, tal como hacen los humanos, llevan a presentar a la *clasiificación conceptual*, véase (Michalski and Stepp 1983).

La *clasiificación conceptual* es una línea de métodos de clasificación basados en a generalización de conceptos, entendida en un sentido más o menos amplio, que parten de una matriz de datos donde todas las variables (atributos en el contexto de la Inteligencia Artificial) son categóricas (cualitativas) y las descripciones de los objetos se asocian a los conceptos que se van a generalizar durante el proceso, construyendo las clases que irá teniendo cada vez mayor cobertura. La clave del método es minimizar la *esparsidad* (sparseness) que consiste en generalizar los conceptos al máximo pero evitando, en lo posible, que sean tan generales que cubran partes del dominio no observadas en la matriz de datos.

La estructura de los métodos de cluster conceptual descansa básicamente en la teoría de la lógica formal y los criterios para decidir las agregaciones sucesivas se basan en como de bien o de mal se generaliza el concepto (lógico) representado en cada clase. El problema es de complejidad combinatoria y la única forma de introducir variables numéricas en la clasificación es partiendo de una discretización previa que inherentemente sesgará de uno u otro modo los resultados.

En nuestra propuesta existe un especial interés por mantener los datos en su estructura original con el fin de evitar pérdida de información o sesgos más o menos arbitrarios. El

problema es altamente conocido y ha sido extensamente discutido en trabajos previos entre los cuales cabe citar (Gibert, Sonicki, and Martin 2002) donde se observa la conveniencia de mantener cada variable en su estado original.

Otra diferencia importante es que la propuesta presentada descansa sobre los métodos de clasificación basada en reglas donde, la agrupación de objetos se apoya en un aparato algebraico (clustering jerárquico) y se utiliza el formalismo lógico (en la base de conocimiento a priori) para introducir un sesgo de carácter semántico en el proceso de clasificación, el cual se realiza por criterios basados en distancia y nada tiene que ver con la generación de conceptos.

La clasificación basada en reglas representa una combinación algebraica lógica que permite mantener los datos con su estructura original, corregir la posición relativa de los objetos según su sesgo semántico y garantizar la coherencia semántica de los resultados.

Es precisamente el hecho de utilizar cluster jerárquico subyacentemente, cuyos resultados se limitan a proporcionar una partición de los datos, lo que hace necesario desarrollar un método posterior que identifique el concepto asociado a cada clase, que en cluster conceptual se obtiene de forma inherente.

No obstante los métodos de clustering conceptual y derivados no pueden clasificar matrices muy grandes por cuestiones de complejidad que los métodos algebraicos pueden resolver de forma más eficaz.

Las aplicaciones que se presentan en esta tesis contienen únicamente datos numéricos, pero no existen criterios claros para discretizar de antemano las variables. Además nuestra propuesta abre la puerta a realizar interpretaciones conceptuales de cualquier partición de unos datos realiza con cualquier método de clasificación jerárquico (basado en distancias, o conceptos lógicos, o conexiónistas, o cualquiera que sea su naturaleza).

Mapas SOM. Self Organization Mapping (Kohonen)

En 1982 Teuvo Kohonen (Kohonen 1995) presentó un modelo de red denominado mapas autoorganizados o Modelo de kohonen (Self-Organizing Maps), basado en ciertas evidencias descubiertas a nivel cerebral y con un gran potencial de aplicabilidad práctica. Este tipo de red se caracteriza por poseer un aprendizaje no supervisado competitivo. El objetivo de este aprendizaje es categorizar (clusterizar) los datos que se introducen en la red. De esta forma, las informaciones similares son clasificadas formando parte de la misma categoría y, por tanto, deben activar la misma neurona de salida. Las clases o categorías deben ser creadas por la propia red, puesto que se trata de un aprendizaje no supervisado, a través de las correlaciones entre los datos de entrada.

Las posibilidades que se contemplan en estos casos son las siguientes:

- Agrupamiento, Los datos de la entrada deben estar agrupados, y el sistema de procesamiento de datos, debe contar con los grupos como parámetro de entrada. La salida del sistema entrega el rotulado de los grupos en la entrada.
- Cuantificación de vectores, Este problema ocurre cuando un espacio continuo ha sido discretizado. La entrada del sistema es un vector n-dimensional y la salida es una representación discreta del espacio de la entrada.
- Reducción dimensional, La entrada es agrupada en sub-espacios de dimensión inferior a los datos de entrada. El sistema debe ser capaz de mantener la variación de los datos de entrada en los de la salida mediante el aprendizaje obtenido anteriormente.
- Extracción de rasgos. El sistema debe extraer rasgos a partir de las señales en la entrada. Lo anterior por lo general implica una reducción dimensional.

Dentro de las arquitecturas para la red de Kohonen, existen dos que son las más importantes y destacadas. La Primera de ellas se trata de la arquitectura L.V.Q. (Learning Vector Quantization). En el algoritmo asociado al modelo de Kohonen se puede considerar, por un lado, una etapa de funcionamiento donde se presenta, ante la red entrenada, un patrón de entrada y éste se asocia a la neurona o categoría cuyo vector de referencia es el más parecido y, por otro lado, una etapa de entrenamiento o aprendizaje donde se organizan las categorías que forman el mapa mediante un proceso no supervisado a partir de las relaciones descubiertas en el conjunto de los datos de entrenamiento.

Clasificadores Difusos

Otra de las nuevas vías de investigación es el *fuzzy mining*, esto es, la utilización de las técnicas de minería de datos con objetos simbólicos, que representen más fidedignamente la incertidumbre que se tiene de los objetos que se estudian (Aluja 2001).

El algoritmo Fuzzy C-means (Bezdek 1981) es la generalización de K-means (MacQueen 1967) al enfoque basado en la lógica difusa.

El algoritmo es el siguiente:

- Selección de la función de distancia
- Selección del número de clusters
- Interpretación del significado de los clusters
- Usar C.A. Cuando pensamos que hay grupos naturales en los datos
- Al agrupar en clusters, reducimos la complejidad de los datos de tal manera que en los clusters resultantes podamos emplear otras técnicas

3.5.3 Clustering en AI& Stats

La *clasificación basada en reglas* (Gibert and Cortés 1998a) es una metodología de clasificación consistente en la hibridación de un proceso inductivo (Inteligencia Artificial) con uno de clustering (Estadística). El desarrollo original ha sido publicado en (Gibert 1996b)(Gibert, Aluja, and Cortés 1998)(Gibert and Cortés 1998a) y se desarrolla formalmente en (Gibert and Cortés 1994) (detalles en (Gibert 1994)).

Los métodos clásicos de clasificación automática aplicada a *dominios poco estructurados* (Gibert 1994), muchas veces presentan resultados que no se pueden interpretar.

En muchas ocasiones el experto tiene suficiente conocimiento para organizar parte del dominio en entidades que tengan sentido. Sin embargo, los métodos estadísticos clásicos prácticamente ignoran esta información. Así una forma de resolver un problema de clasificación es el diseñar una metodología híbrida que permita cooperar a un proceso de clustering estadístico clásico con la construcción de una Base de Conocimiento a priori parcial.

Clustering based on rules

La idea fundamental del *Clustering based on rules*(Gibert 1994) es recoger este conocimiento y utilizarlo de forma cooperativa en el proceso de clustering. Dicho conocimiento se recoge en forma de reglas que subdividen el espacio de clasificación en entornos coherentes y la idea es que la clasificación final propuesta respete esta primera estructuración sugerida directamente por el experto. Con esto se pretende cubrir tres objetivos: incorporación a la clasificación de información antes ignorada (como relaciones entre atributos o restricciones), recogida de los

objetivos de la clasificación que se pretende obtener y garantizar la interpretabilidad de la clasificación obtenida (Gibert, Aluja, and Cortés 1998).

Visto que se utiliza información parcial que proporciona el experto, la clasificación basada en reglas, se sitúa a medio camino entre los métodos supervisados y los no supervisados dependiendo del grado de completitud de la base de conocimiento utilizada.

En este contexto, el conocimiento *a priori* proporcionado por el experto se formaliza en un conjunto de restricciones declarativas que la estructura final propuesta para el dominio ha de satisfacer (\mathcal{R}). Esas restricciones se utilizarán para inducir una primera *super-estructura* del dominio, que aunque parcial, guiará todo el proceso. La aproximación de la *clasificación basada en reglas* se basa en realizar clasificaciones *internas* a esta superestructura, respetando las restricciones del usuario, que pueden estar basadas (y de hecho es recomendable que así sea) en argumentos de naturaleza *semántica*.

Finalmente, todos los elementos del dominio serán integrados en una única estructura global. Los métodos de clasificación jerárquica son especialmente apropiados para nuestros propósitos, principalmente considerando que el conocimiento proporcionado por el experto va a ser heterogéneo (más específico en ciertas partes del dominio y más general en otras) y únicamente la organización jerárquica permite la generación de una única clasificación global que contemple este factor, poniendo de manifiesto qué parte de su conocimiento es más o menos genérica; y que los métodos jerárquicos, actualmente los más utilizados, permiten decidir el número de clases *a posteriori*, una vez construido el dendrograma y ya viendo la forma que presenta, lo que es muy recomendable en aplicaciones reales donde se busca la estructura subyacente a un dominio (presumiblemente porque no se conoce, o no está clara) y donde es más bien raro que se tenga previa información segura sobre el número de clases.

Dependiendo del carácter de la base de conocimiento proporcionada por el experto (\mathcal{R}), este proceso será no supervisado (como el clustering, $\mathcal{R} = \emptyset$) o supervisado (como la clasificación, \mathcal{R} será entonces una teoría completa del dominio o contendrá una clasificación de referencia). El método permite trabajar en cualquier situación intermedia, que use una bases de conocimiento *parciales*, lo que nos sitúa en el contexto de los métodos semisupervisados.

La *clasificación basada en reglas* consiste en, dado un conjunto $\mathcal{I} = \{i_1 \dots i_n\}$:

1. Construir la base de conocimiento inicial: El objetivo principal es permitir al experto introducir el conocimiento *a priori* del dominio de que dispone en forma de *restricciones* a la formación de clases (básicamente se trataría de materializar en esa base de conocimiento las cosas que se sabe que *son* y las que se sabe que *no pueden ser*); el experto proporciona dicho conocimiento de forma *declarativa*, lo que da lugar a un conjunto inicial de reglas lógicas, sea \mathcal{R}^0 .
 - Iniciar el proceso iterativo ($\xi = 1$):
2. Fase de proceso del conocimiento a priori:
 - (a) Determinar la partición de \mathcal{I} inducida por las reglas: $\mathcal{P}_{\mathcal{R}}^\xi$ a partir de \mathcal{R}^ξ . Incluir una *clase residual* \mathcal{C}_0^ξ en $\mathcal{P}_{\mathcal{R}}^\xi$ con los objetos para los que se proporcionó conocimiento inconsistente o no se proporcionó conocimiento alguno.
 - (b) Fase de resolución de conflictos: Analizar los objetos de \mathcal{C}_0^ξ seleccionados por reglas contradictorias:
 - i. Si es satisfactorio, ir a la fase de clasificación.
 - ii. Sino, volver a la construcción de \mathcal{R}^ξ y reformularla.
3. Fase de clasificación:

- (a) Clasificación *intra* restricciones del experto: $\mathcal{P}_{\mathcal{R}}^{\xi}$ satisfará *a priori* los requerimientos del experto. Realizar la clasificación para cada $\mathcal{C} \in \mathcal{P}_{\mathcal{R}}^{\xi}$. Notar que las clases $\mathcal{C} \subset \mathcal{I}$ lo que abarata la construcción de las clases. Determinar:
- i. Los correspondientes árboles jerárquicos (*dendrogramas*) $\tau_{\mathcal{C}}^{\xi}$,
 - ii. Sus prototipos $\bar{i}_{\mathcal{C}}^{\xi}$, vía sumarización de la clase,
 - iii. Sus masas $m_{\mathcal{C}}^{\xi} = \text{card } \mathcal{C}$ y
 - iv. Sus índices de nivel $h_{\mathcal{C}}^{\xi}$.

4. Fase de integración:

- (a) Extender la clase residual: Añadir los prototipos $\bar{i}_{\mathcal{C}}^{\xi}$ a la clase residual \mathcal{C}_0^{ξ} , como si fueran objetos ordinarios, pero teniendo en cuenta sus respectivas masas. El nuevo conjunto de datos es la llamada *clase residual extendida* $\tilde{\mathcal{I}}^{\xi}$:

$$\tilde{\mathcal{I}}^{\xi} = \left\{ (\bar{i}_{\mathcal{C}}^{\xi}, m_{\mathcal{C}}^{\xi}) : \mathcal{C} \in \mathcal{P}_{\mathcal{R}}^{\xi} \right\} \cup \left\{ (i, 1) : i \in \mathcal{C}_0^{\xi} \right\}$$

- (b) Realizar la integración: Clasificar $\tilde{\mathcal{I}}^{\xi}$ para integrar todos los objetos en una sola jerarquía, recuperando la estructura jerárquica de los prototipos $\bar{i}_{\mathcal{C}}^{\xi}$ previamente calculadas ($\tau_{\mathcal{C}}^{\xi}$, $(\mathcal{C} \in \mathcal{P}_{\mathcal{R}}^{\xi})$) y descolgándolos de su raíz a nivel $h_{\mathcal{C}}^{\xi}$ en la jerarquía global. Ello da lugar a la jerarquía τ^{ξ} .
- (c) Determinar el número final de clases: Analizar el dendrograma τ^{ξ} para elegir el mejor corte horizontal, utilizando criterios heurísticos (ya sea manuales o automáticos) (Gibert, Aluja, and Cortés 1998). Realizar el corte de τ^{ξ} identifica una partición de los datos en un conjunto de clases, \mathcal{P}^{ξ} . Entre los k mejores cortes (siendo k pequeño), elegir aquél que permita una mejor *interpretación*.

5. Fase de Evaluación: El experto debe confirmar también que la partición \mathcal{P}^{ξ} obtenida con \mathcal{R}^{ξ} mejora la partición $\mathcal{P}^{\xi-1}$ que se obtuvo con $\mathcal{R}^{\xi-1}$ en el modo deseado. Para ello se pueden analizar qué términos contribuyen más a las diferencias entre ellas o se pueden comparar distintas clasificaciones entre sí a través de tablas; es incluso posible probar la significación de dichas diferencias, utilizando un test no paramétrico (δ -test) diseñado para este propósito y presentado en (Gibert, Aluja, and Cortés 1998). Este paso puede producir el criterio de terminación del proceso:

- (a) Si la mejora no es significativa, parar la iteración y asumir los resultados de la iteración anterior como los mejores.
- (b) Sino, analizar los resultados para reformular la base de conocimiento. Construir $\mathcal{R}^{\xi+1}$, incrementar ($\xi = \xi + 1$) y repetir.

KLASS es el software que proporciona una implementación de esta metodología, actualmente disponible en versión LISP. A pesar de que la metodología es genérica y se podría utilizar cualquier motor de inferencia y cualquier formalismo lógico tanto para construir \mathcal{R}^{ξ} como para calcular las clases inducidas por las reglas, y asimismo cualquier método jerárquico de clustering para la clasificación, *KLASS* utiliza la lógica clásica con reglas de primer orden para el manejo del conocimiento, y una reescritura del método de los *vecinos recíprocos encadenados* (De Rham 1997) —convenientemente adaptado para tratar con datos heterogéneos vía las medidas de compatibilidad que se detallan en (Gibert and Nonell 2003), (Gibert, Nonell, Velarde, and Colillas 2004)— como método de clasificación subyacente. Una parte importante de nuestra investigación se ha centrado en las medidas de compatibilidad, lo que dio en su momento lugar a la definición de la *distancia mixta* (Gibert and Cortés 1997).

Las principales propiedades son:

- Permite tener en cuenta el conocimiento *a priori* existente sobre el dominio en estudio, incluso siendo este parcial.
- No requiere conocimiento completo sobre el dominio, admitiendo *BC* parciales.
- Las clases resultantes son consistentes con el conocimiento *a priori* proporcionado por el experto. Pueden generarla o especificarla, pero guardan consistencia con él.
- Mejora la *interpretabilidad* de las clases resultantes.

Con ello, su comportamiento frente a *dominios poco estructurados* supera las limitaciones de las técnicas clustering puras porque produce clases más *comprendibles* desde el punto de vista semántico. Lo mismo ocurre con las técnicas de aprendizaje inductivo por si mismas, puesto que reduce los efectos del conocimiento implícito, que necesariamente revierte en bases de conocimiento incompletas.

Sin embargo, en situaciones reales, es usual trabajar con dominios complejos, ver (Gibert and Cortés 1998a), tales como trastornos mentales (Gibert and Sonicki 1997), esponjas marinas (Gibert 1994), disfunciones tiroïdales (Gibert and Sonicki 1999), pruebas psicofisiológicas (Rodas, Gibert, and Rojo 2001), discapacidad de ancianos (Annichiarico and Gibert 2004) o rehabilitación (Gibert, García-Rudolph, García-Molina, Roig-Rovira, Bernabeu, and Tormos 2008) y muchas más, donde las bases de datos tienen tanto variables cualitativas como cuantitativas; y el experto posee algún conocimiento *a priori* (en general parcial) de la estructura del dominio que raramente un método de clustering sabría tener en cuenta—el cual es difícil de incluir en una *Base de Conocimiento*.

3.5.4 Validación de un clustering

Como el clustering produce clases donde se desconoce la estructura real subyacente, la validación de los resultados obtenidos se hace muy difícil. De hecho la validación de un cluster sigue siendo un *problema abierto* por no haberse encontrado aún un criterio objetivo para determinar la *calidad* de un conjunto de clases en el contexto del clustering (Hand 1996), que se aplica en situaciones en las que *no hay* un buen conocimiento de la estructura del dominio que pueda servir de referencia (como se hace en el caso supervisado), y si lo hay, es sólo parcial. Uno de los criterios utilizados habitualmente hasta ahora para la validación de clases es su *utilidad* que pasa por que tengan *significado* claro para el experto desde el punto de vista conceptual. Volle en 1985 (Volle 1985) hace toda una disertación ilustrando que el concepto de validez no es absoluto si no que va ligado a su utilidad, lo que es relativo a las condiciones del contexto. A pesar de que éste es un extremo poco objetivable, la *interpretación* se convierte así en una fase fundamental del proceso y sigue siendo, aún hoy, uno de los criterios más utilizados en la práctica para validar el cluster. Por esta razón la validación queda directamente ligada a la existencia de una interpretación clara para el clustering o partición. Actualmente es necesario introducir herramientas para asistir al usuario en las tareas de interpretación de una partición sobre un conjunto de objetos, para establecer el significado de las clases resultantes. Si las clases obtenidas no tienen sentido para el/los expertos, los resultados de la clasificación no son considerados válidos.

La aplicación de un algoritmo de clasificación a un conjunto de datos lleva a una partición o una jerarquía de particiones sobre los elementos en la clasificación. Después de una *interpretación*, este resultado nos dará información sobre las relaciones entre los elementos en la clasificación.

Como ya se ha visto el espacio de búsqueda asociado a un problema de clustering es inexplorable en un tiempo razonable y todos los métodos de *clustering* se basan en algún

heurístico para acotar la búsqueda y resultar viables. Sea cual sea la naturaleza del heurístico (lógico o algebraico) el resultado será un subóptimo respecto al criterio que se estuviera optimizando, pero a priori no hay garantía de que exista un buen ajuste con las clases reales existentes en el dominio. De hecho incluso si no hay clases de ningún tipo, la mayoría de métodos produce una u otra partición de todos modos, que, en el caso, sería sólo una división artificial y superficial de los datos. Así, ante el peligro de interpretar la existencia de diferentes clusters cuando estos no existen realmente es fundamental validar los resultados obtenidos.

En (Aluja 1996) se plantea hasta qué punto las clases obtenidas en un proceso de clasificación reflejan clases reales presentes en los datos, o si por el contrario, las clases obtenidas son el simple resultado de aplicar un algoritmo a los datos, es decir, una partición artificial de una realidad continua.

También se afirma que la experiencia prueba que, aunque nos encontramos en este último caso, la tipología obtenida puede ser igualmente útil, ya que aunque no se pueda hablar de clases realmente diferenciadas entre ellas, la partición obtenida suele facilitar la comprensión de los datos y por tanto su operatividad. En este caso hablamos de *clases instrumentales* en oposición a *clases reales* y volvemos al punto en que lo importante es la utilidad de las clases.

Varios autores han estudiado diferentes aproximaciones para la *validación de una clasificación*. Podemos mencionar (entre otros) las investigaciones de Bock (Bock 1985), (Bock 1996), que insertan los modelos de clasificación en un contexto probabilístico, asumiendo que los datos observados son una muestra de una población de estructura multivariada y tratando de ver qué diferencia significativa hay en la distribución de cada clase.

En (Gordon 1994), (Gordon 1996) y (Milligan. 1996) se apela al uso de herramientas empíricas, descriptivas o exploratorias; analizando la calidad de los grupos (clases) obtenidos.

Es usual aplicar varios métodos de clasificación (agrupamiento) a un mismo conjunto de datos con el fin de hacer la elección de uno de ellos o determinar uno nuevo a partir del conjunto de datos estudiado. Cuando se trata de elegir la mejor entre varias propuestas de partición, si se trata de particiones provenientes de clasificaciones jerárquicas (árboles de clasificación o dendogramas como se estudia en (Lapointe and Legendre 1990), (Lapointe and Legendre 1995)) y la teoría general en (Barthélemy and Monjardet 1986)) proponen la comparación directa de árboles o dendogramas. Otros autores proponen la combinación de todos los resultados en uno solo que sea más estable (clustering ensamble) (Gibert, Oliva, Sánchez-Marré, and Pinyol 2007).

Cuando se dispone de un único resultado es fundamental entender la calidad y estabilidad del mismo. Así la *validación de una clasificación* también puede ser considerada bajo diferentes perspectivas (Hubert 1987):

- Validar una sola clase o grupo obtenido, ha sido estudiada, entre otros, por (Gordon 1994).
- Validar una conjunto de clases o grupos obtenidos (partición) (Hubert 1987), o
- Validar una jerarquía, debido a su complejidad tiene menos referencias y se nota en la investigaciones que la validación de jerarquías aparece a menudo en el contexto de la validación de particiones, dado que las jerarquías son sucesiones de particiones.

En este trabajo nos centramos en la validación de una partición. Un índice de *validación de una partición* debe indicar la calidad de la clasificación obtenida (Halkidi, Batistakis, and Vazirgiannis 2001).

Existen en la literatura índices de validación de clasificaciones, algunos de ellos son:

1. *Dunn index* (Dunn 1974), intenta identificar qué tan compactos y separados son los grupos obtenidos, es un índice definido para un número específico de grupos. Grandes

valores del índice corresponden a un buen *clusters*, es decir, el número de clusters que maximiza el índice D es el número óptimo de *clusters*.

$$D_\xi = \min_{C \in \mathcal{P}_\xi} \left\{ \min_{C' \in \mathcal{P}_\xi, C > C'} \left\{ \frac{d(C, C')}{\max_{C'' \in \mathcal{P}_\xi} diam(C'')} \right\} \right\} \quad (3.1)$$

donde $d(C, C')$ es la distancia *intercluster*,

$$d(C, C') = \min_{i \in C, i' \in C'} d(i, i') \quad (3.2)$$

$diam(C)$ es el diámetro de la clase C , o la distancia *intraclass* y se define,

$$diam(C) = \max_{i, i' \in C} d(i, i') \quad (3.3)$$

y ξ es el número de clases de la partición de referencia \mathcal{P}_ξ .

2. *Davies-Bouldin index* (Bouldin and Davies 1979), es una medida de similitud entre los grupos y se define sobre la base de una medida de dispersión de un grupo y una medida de disimilitud entre dos grupos. Pequeños valores del índice identifican un buen *cluster*, es decir, el número de clusters que minimiza el índice DB es el número óptimo de *clusters*.

$$DB_\xi = \frac{1}{\xi} \sum_{C \in \mathcal{P}_\xi} \max_{C \neq C'} \left\{ \frac{diam(C) + diam(C')}{d(C, C')} \right\} \quad (3.4)$$

El diámetro $diam(C)$ se calcula utilizando la ecuación 3.3

3. *C-index* (Schultz and Hubert 1976), se basa en el cálculo de 3 medidas:

S es la suma de las distancias de todos los pares de elementos del mismo cluster. Se calcula con $l = \sum_{C \in \mathcal{P}_\xi} \frac{n_c(n_c-1)}{2}$ sumandos.

S_{min} suma los l elementos menores de la matriz de distancia de \mathcal{I} . Del mismo modo S_{max} suma los elementos mayores de la matriz de distancia de \mathcal{I} . Y el *C-index* se construye como:

$$C\text{-index} = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (3.5)$$

Pequeños valores del *C-index* se asocian con un buena partición porque $S - S_{min}$ tiende a cero si los elementos más parecidos de la base de datos van a parar a la misma clase y a $S_{max} - S_{min}$ si se agruparan en la misma clase los elementos menos parecidos.

4. *Goodman-Kruskal index* (Kruskal and Goodman 1954), comúnmente usado en estadística, grandes valores del índice Kruskal-Goodman se asocian con un buena partición, porque indica que hay mas cuádruples concordantes (los elementos de clases diferentes tienen distribuciones mayores que los de la misma clase).

$$GK = \frac{N_c - N_d}{N_c + N_d} \quad (3.6)$$

donde (i_1, i_2, i_3, i_4) una cuádrupla de elementos de la base de datos.

N_c es el número de *cuádruplas concordantes* que cumple una de las siguientes condiciones:

- $d(i_1, i_2) < d(i_3, i_4)$, i_1 y i_2 pertenecen al mismo *cluster*, i_3 y i_4 pertenecen a diferentes *clusters*
- $d(i_1, i_2) > d(i_3, i_4)$, i_1 y i_2 pertenecen a diferentes *clusters*, i_3 y i_4 pertenecen al mismo *cluster*.

y N_d es el número de *cuádruplas discordantes* que cumple una de las siguientes condiciones:

- $d(i_1, i_2) < d(i_3, i_4)$, i_1 y i_2 pertenecen a diferentes *clusters*, i_3 y i_4 pertenecen al mismo *cluster*
- $d(i_1, i_2) > d(i_3, i_4)$, i_1 y i_2 pertenecen al mismo *cluster*, i_3 y i_4 pertenecen a diferentes *clusters*.

5. *Silhouette index* (Rousseeuw 1987), calcula el *silhouette-width* de cada una de las muestras, la promedio de cada cluster y, en general, la *silhouette-width* total de un conjunto de datos. Este enfoque que se basa en la comparación de la *compresión* y la *separación* de la *silhouette* de cada cluster.

Sea $C = \{i_1, \dots, i_{n_c}\}$ un cluster de \mathcal{P}_ξ , donde $n_c = |C|$. La distancia media entre el i -ésimo vector del cluster C y otro vector del mismo cluster se calcula mediante la siguiente expresión (disimilitud promedio del i con todos los demás objetos en el mismo cluster):

$$a_i^C = \frac{1}{n_c - 1} \sum_{i' \in C, i \neq i'} d(i, i'), \quad i = 1, \dots, n_c \quad (3.7)$$

La distancia mínima entre el i -ésimo elemento del cluster C y todos los elementos pertenecientes al cluster C' , $C \neq C'$ se calcula de la siguiente forma:

$$b_i^C = \min_{C' \in \mathcal{P}_\xi; C \neq C'} \left\{ \frac{1}{n_{c'}} \sum_{i' \in C'; C' \neq C} d(i, i') \right\} \quad i \in C \quad (3.8)$$

Entonces la *silhouettes-width* del i -ésimo elemento del cluster C se define como:

$$s_i^C = \frac{b_i^C - a_i^C}{\max\{a_i^C, b_i^C\}} \quad (3.9)$$

De la expresión anterior se puede definir el *Silhouette index* del cluster C como:

$$S_C = \frac{1}{n_c} \sum_{i \in C} s_i^C \quad (3.10)$$

y finalmente

$$S = \frac{1}{\xi} \sum_{\forall C \in \mathcal{P}_\xi} S_C \quad (3.11)$$

Si el índice S toma valores cercanos a 1, significa que la muestra está *bien agrupada* y que cada elemento fue asignado a un grupo muy adecuado. Si índice S toma valores cercanos a cero, significa que la muestra se podría asignar a otro grupo más cercano y, además, significa que la muestra se encuentra igualmente alejada de ambos grupos. Si el índice S toma valores cercanos -1, significa que la muestra está *incorrectamente clasificada*.

6. En (Gibert, Oliva, Sàncchez-Marré, and Pinyol 2007) se propone medir la bondad de una partición usando:

(a) Una extensión al caso multivariado del cociente entre la variabilidad entre clases y la intra clases. El coeficiente I es, de hecho, una medida de la separabilidad de las clases y la homogeneidad dentro de las clases. Esta medida se basa en la F Estadística clásica, y calcula la relación entre la inercia intra clases (lo que aumenta con la heterogeneidad de una clase) y la inercia entre las clases (que aumenta con la distinguibilidad entre las clases). Aquí, se presenta una formulación multivariante. Es importante notar aquí que I se puede calcular sólo con variables numéricas.

Sea n_c el número de elementos C . ξ el número de clases de la partición \mathcal{P}_ξ y K el número de variables. $\bar{c} = (\bar{x}_{c1}, \dots, \bar{x}_{cK})$ el centroide de la clase c , donde $\bar{x}_{ck} = \frac{\sum_{i \in c} x_{ik}}{n_c}$ y sea $\bar{x} = (\bar{x}_1, \dots, \bar{x}_K)$ el centroide global de todo el conjunto de datos, donde $\bar{x}_k = \frac{\sum_{i=1}^n x_{ik}}{n}$. Si $d(i, j)^2 = \sum_{k=1}^K (x_{ik} - x_{jk})^2$ es la distancia euclídea al cuadrado entre los objetos i y i' . Entonces, la inercia intra clases (S_w^2) se define como:

$$S_w^2 = \frac{\sum_{\forall c \in \mathcal{P}_\xi} (n_c - 1) S_c^2}{n - \xi} \quad (3.12)$$

donde,

$$S_c^2 = \frac{\sum_{\forall i \in c} d(i, \bar{c})^2}{n_c - 1} \quad (3.13)$$

Por otro lado, la inercia entre clases (S_ξ^2) es

$$S_\xi^2 = \frac{\sum_{\forall c} d(\bar{c}, \bar{x})^2}{n - \xi} \quad (3.14)$$

Finalmente, la inercia total (I) se escribe como

$$I = \frac{S_\xi^2}{S_p^2} \quad (3.15)$$

El coeficiente I crece, cuando la variabilidad entre clases aumenta (S_ξ^2) y cuando la variabilidad intra clase disminuye (S_w^2). Es decir, cuanto más distinguibles y compactas sean las clases, mayor es el valor de I .

- (b) El coeficiente de Información Mutua (Shannon 1948) entre la partición y el conjunto de variables usados en la clasificación (ya usado por (Lee and Huh 2003) y (Huh and Song 2002) con los mismos fines). En (Gibert, Oliva, Sàncchez-Marré, and Pinyol 2007) se concluye que $MI(X_1, \dots, X_K, \mathcal{P}_\xi)$ presenta mejor comportamiento que I por valorar como mejores precisamente aquellas particiones que luego serán más fáciles de interpretar, lo cual no es sorprendente, porque precisamente el coeficiente de información mutua esta relacionado con la cantidad de información que aporta un grupo de variables a la clase (Benzécri 1973). En general, siendo \mathcal{P}_ξ la partición de referencia, se tiene que el coeficiente de información mutua entre esta y el conjunto de variables se calcula como:

$$\begin{aligned}
MI(X_1, \dots, X_K, \mathcal{P}_\xi) = & \\
& \int_{x_1} \dots \int_{x_K} \int_C f_{X_1 \dots X_K, \mathcal{P}_\xi}(x_1 \dots x_K, C) \log \left(\frac{f_{X_1 \dots X_K, \mathcal{P}_\xi}(x_1 \dots x_K, C)}{f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_K}(x_K) f_{\mathcal{P}_\xi}(C)} \right) \\
& dx_1 \dots dx_K dC
\end{aligned} \tag{3.16}$$

siendo $f_{X_k}(x)$ la función de densidad de X_k , $\forall k$ y $f_{X_1 \dots X_K, \mathcal{P}_\xi}(x_1 \dots x_K, C)$ la densidad conjunta entre $X_1 \dots X_K$ y $X_{\mathcal{P}_\xi}$.

En la práctica, esto es fácil de calcular cuando $X_1 \dots X_K$ son variables cualitativas, ya que en este caso es posible utilizar las frecuencias de las modalidades de las variables, como estimaciones de probabilidad. Se define $D_1 \dots D_K$ como el conjunto de valores tomado por las variables cualitativas $X_1 \dots X_K$.

$$M(X_1, \dots, X_K, \mathcal{P}_\xi) = \sum_{x_1 \in D_1} \dots \sum_{x_K \in D_K} \sum_{C \in \mathcal{P}_\xi} F(x_1, \dots, x_K, C) \tag{3.17}$$

donde

$$F(x_1, \dots, x_K, C) = P(x_1, \dots, x_K, C) \log \left(\frac{P(x_1 \dots x_K, C)}{P(x_1 \dots x_K)P(C)} \right) \tag{3.18}$$

En (Chieppa, Gibert, M., and Gómez-Sebastià 2008) se presentan propuestas para aproximar el cálculo de $F_{X_1, \dots, X_K, \mathcal{P}_\xi}$, F_{X_1, \dots, X_K} y F_{X_K} , $k = 1, \dots, K$ con variables numéricas basadas en distintas formas de discretizarlas.

Sin embargo todos ellos van orientados a validar la vertiente estructural de la partición, a ver que geométricamente las clases están bien formadas, son compactas y distinguibles entre sí.

Pero disponer de clases bien formadas estructuralmente no ofrece la menor garantía de que un experto vaya a ser capaz de asociar cada uno de esos grupos a una entidad semántica de su dominio, un concepto de su base de conocimiento personal sobre la cual él pueda tomar decisiones y razonar. Por ello es fundamental que, aparte de validar una clasificación estructuralmente se asista al usuario a comprender cuál fue el criterio de clasificación subyacente y a entender el significado de las clases.

Quizás por su naturaleza más semántica la generación automática de interpretaciones de una clasificación no se ha tratado formalmente desde el ámbito estadístico, aunque resolverlo es fundamental. Éste, de hecho, es uno de los problemas objeto del aprendizaje automático, del cual ID3 (Quinlan 1990) y C4.5 (Quinlan 1993) y sus sucesores son exponentes característicos. Tanto los árboles de decisión, como los métodos de inducción de reglas generan modelos completos y consistentes con los datos. Se produce aquí con frecuencia el fenómeno del *over-fitting*, lo que significa que se ajusta tanto a la muestra observada que se aprende incluso el ruido y no se generaliza a otras muestras.

Por otro lado, la construcción de árboles de decisión suele tener coste exponencial, aunque existen técnicas para reducir el espacio de búsqueda. Así también puede generar conceptos de alto valor predictivo pero poco compactos y por lo tanto de comprensión compleja si tratamos bases de datos con muchas clases y/o variables. Hemos explorado una posibilidad más barata, también útil con variables numéricas y que genere conceptos de comprensión más inmediata por parte del experto sea cual sea el tratamiento de la Base de Datos. Nuevamente nos hallamos ante un problema común a la Estadística y a la Inteligencia Artificial, y ligado a los sistemas de *KDD* (Fayyad 1996), muy en la línea de nuestros trabajos.

Capítulo 4

Antecedentes

4.1 Antecedentes

Esta tesis es la continuación de una línea de investigación consolidada en el seno del grupo donde se inserta, especialmente orientada a la generación de herramientas de soporte a la interpretación en *clustering*. La propuesta se fundamenta en el estudio de la distribución conjunta entre pares de clases. Lo anterior da lugar a descripciones conceptuales que contienen la conjunción de dos atributos en cada clase. A partir de ello se observó enseguida que limitarse a pares de atributos impide conseguir buenas caracterizaciones de las clases, porque muchas veces, lo típico de ciertas clases es la interacción entre más de dos variables. Además, a mayor complejidad de la estructura del dominio, mayor suele ser el orden de estas interacciones. Es más, incluso nos atreveríamos a decir que, es precisamente en esa propiedad donde radica la característica de que un dominio sea complejo y se sitúe en lo que (Gibert 1994) denomina *dominios poco estructurados*.

Así, en (Gibert, Aluja, and Cortés 1998) se describe la forma de caracterizar una clasificación utilizando los representantes de clase a partir de variables cualitativas y se presenta la primera versión sobre el uso de condicionamientos sucesivos, en ese caso utilizando una hipótesis de mundo cerrado y conjuntandos negativos en los conceptos generados, como primera aportación al manejo de interacciones de orden superior a dos con variables cualitativas.

Paralelamente, (Rodríguez, D. 1999) aborda una primera aproximación para visualizar la distribución de una variable numérica común respecto algunos grupos (clases), haciendo uso del *box-plot* múltiple¹. En este trabajo se constató que esta herramienta proporciona toda la información necesaria para identificar las variables caracterizadoras de una clase. Las primeras aplicaciones del método se realizaron sobre bases de datos médicas previamente analizadas manualmente (Gibert and Sonicki 1997) (Gibert and Sonicki 1999) y en (Gibert and Roda 2000) se consolida el uso de ésta (como alternativa al método de variables cualitativas) para el caso de variables numéricas, aplicándolo por primera vez en el análisis de plantas depuradoras.

Detectado en las variables numéricas el problema que ya se había presentado en las cualitativas, que se requiere estudiar interacciones de orden superior a dos, con el agravante de que el trabajo con rangos continuos no aconseja el análisis por casos, en (Gibert and Salvador 2000) se plantea la posibilidad de estudiar las intersecciones entre las cajas de los *box-plot*, lo

¹El *box-plot* múltiple es una herramienta gráfica introducida en (Tukey 1977) y funciona del siguiente modo: para cada clase el intervalo de valores que toma la variable se visualiza y las observaciones atípicas (outliers) se marca con “*”. Se despliega una caja desde Q_1 (primer cuartil) hasta Q_3 (tercer cuartil) y la mediana se marca con un signo horizontal al centro de la caja. Las cajas incluyen, entonces, el 50% de los elementos de la clase y los bigotes se extienden hasta el mínimo y máximo para cada clase.

que da lugar a una primera versión de un mecanismo que genera reglas probabilizadas y se sitúa en el paradigma difuso.

En realidad a lo largo de esta investigación se ha tomado conciencia de cómo un experto procede manualmente a interpretar a partir del Boxplot múltiple y se ha implementado en *KLASS* un *Class Panel Graph (CPG)* (Gibert, Nonell, Velarde, and Colillas 2005) como primera herramienta que permite ver en perspectiva. Se observó que los expertos daban prioridad a las variables que tomaban valores específicos en las clases y ello condujo a la definición de los conceptos:

1. Valores propios, (Gibert 1994).
2. Valor totalmente caracterizador, (Gibert 1996b).
3. Valor parcialmente caracterizador, (Gibert 1996b).
4. Variable *totalmente caracterizadora*, (Gibert and Cortés 1998a).
5. Variable *parcialmente caracterizadora*, (Gibert and Cortés 1998a).

Todo este trabajo cristaliza en la formulación de dos elementos básicos para el desarrollo de esta tesis:

1. La discretización basada en boxplots cuyas primeras formulaciones arrancan en (Gibert and Cortés 1998b) y que se puede hallar formalizado en (Gibert 2004) que supone una forma eficiente de categorizar una variable numérica a intervalos de longitud variable teniendo en cuenta los puntos donde cambian las intersecciones entre clases (Gibert 2004).

La propiedad principal de este método es que sin entrar en el análisis de intersecciones de orden superior entre clases, calcula con complejidad lineal los puntos exactos donde esa intersección cambia y ello será enormemente útil en la generación de interpretaciones porque la variable categórica generada es la que mejor se asocia a la variable de clase que se quiere explicar.

2. La inducción de reglas basada en el boxplot que constituye un método de generación de conceptos probabilizados, con un número mínimo de atributos en el antecedente. En (Vázquez and Gibert 2001) se incide en la implementación de estas ideas. Aunque se obtienen soluciones satisfactorias desde un punto de vista aplicado, no está clara su optimalidad en términos de cobertura.

En (Comas, Dzeroski, Gibert, Roda, and Sàncchez-Marrè 2001) se comparan los resultados de ésta primera versión con otros métodos de aprendizaje en el ámbito de plantas depuradoras y se observó que los resultados son fácilmente comprensibles por el experto, aunque el método presenta algunas deficiencias todavía.

En (Vázquez and Gibert 2002) se estudia la sensibilidad de las interpretaciones generadas a partir de clasificaciones obtenidas por diferentes procedimientos automáticos. En este trabajo se observó que en dominios de estructura compleja existen núcleos de estructura fuerte (los días de tormenta, etc.) que son reconocidos desde cualquier método de clustering, generando conceptos muy estables, mientras las situaciones de estructura más débil si son sensibles al método.

Esta tesis se apoya en las investigaciones previas en las que se ha analizado el *proceso de caracterización automática de clases* y que en el párrafo anterior se han comentado. Inspirado en la forma cómo los expertos realizan (manualmente) el proceso de interpretación

Capítulo 5

Conceptos previos

Antes de hacer la definición formal de la metodología introducimos algunos conceptos que se van a utilizar como base para el desarrollo de esta tesis. Algunos proceden de la literatura. Otros del trabajo anteriormente desarrollado en el grupo de investigación. Si bien parte de la notación ya se ha introducido en la sección 1.1, reproducimos aquí parte de la formulación para facilitar la lectura.

5.1 Notación general

Sea $\mathcal{I} = \{i_1, \dots, i_n\}$ un conjunto de individuos u objetos, que está descrito por una serie de atributos cualitativos y/o cuantitativos $X_1 \dots X_K$, cuyos valores para cada uno de los individuos $i \in \mathcal{I}$ se representan por una matriz rectangular \mathcal{X} de dimensión (n, K) , como se muestra en la Tabla 5.1:

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1K-1} & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K-1} & x_{2K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n-11} & x_{n-12} & \dots & x_{n-1K-1} & x_{n-1K} \\ x_{n1} & x_{n2} & \dots & x_{nK-1} & x_{nK} \end{pmatrix}$$

Tabla 5.1: Matriz de datos \mathcal{X} .

donde x_{ik} con $1 \leq i \leq n$ y $1 \leq k \leq K$, es el valor que, el individuo i -ésimo toma para el k -ésimo atributo; es decir, que las filas de la matriz de datos \mathcal{X} contienen información relativa a las características de los individuos, la cual se puede representar como un vector de atributos de la forma:

$$x_i = (x_{i1} \ x_{i2} \ \dots \ x_{ik} \ \dots \ x_{nK})$$

y las columnas hacen referencia a los K atributos X_K .

Sea $\mathcal{P}_\xi = \{C_1, \dots, C_\xi\}$, una partición en ξ clases de \mathcal{I} y $\mathcal{P}_2 = \{C_1, C_2\}$, una partición binaria de \mathcal{I} .

Sea $\tau = \{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4, \dots, \mathcal{P}_n\}$, una jerarquía indexada sobre \mathcal{I} . Es importante señalar que dados $\mathcal{P}_\xi \in \tau$ y $\mathcal{P}_{\xi+1} \in \tau$, entre ambas siempre hay una clase y sola una, que se divide exactamente en 2 clases.

5.2 Algoritmo genérico de clasificación ascendente jerárquica

Una clasificación jerárquica es una secuencia de clasificaciones en la que los *clusters* más grandes se forman a través de la fusión consecutiva de clusters más pequeños.

Sea $\mathcal{I} = \{i_1, \dots, i_n\}$ un conjunto de individuos u objetos de un universo de discurso. Las técnicas de clasificación ascendente jerárquica construyen una sucesión de particiones encajadas sobre \mathcal{I} con cada vez menos clases.

Comienzan con la partición más fina, que está formada por tantas clases como individuos a clasificar (n) y terminan con la partición más grosera, formada por una sola clase, con todos los individuos. Por el camino, se pasa de la partición \mathcal{P}_ξ a la $\mathcal{P}_{\xi-1}$ agregando dos clases de \mathcal{P}_ξ en una sola de $\mathcal{P}_{\xi-1}$. Las dos clases que se fusionan en cada paso son siempre las *más próximas*, en algún sentido que habrá que precisar previamente. Esta relación entre \mathcal{P}_ξ y $\mathcal{P}_{\xi-1}$ permite hacer un representación gráfica de \mathcal{P} en forma de árbol jerárquico o dendrograma (ver §5.3). Algoritmos de clasificación ascendente jerárquica también hay muchísimos y cada uno con variedades que le son propias y que llevan a diferentes clasificaciones; pero si se quisiera presentar un algoritmo genérico para los métodos de clasificación ascendente jerárquica, éste podría ser el que plantea (Diday and Moreau 1984), con la ayuda de un índice de distancia δ . Dado un conjunto de individuos, $\mathcal{I} = \{\{i_1\}, \{i_2\}, \dots, \{i_j\}, \{i_k\}, \dots, \{i_n\}\}$ a clasificar:

1. Empezar con la partición $\mathcal{P}_n = \{\{i_1\}, \{i_2\}, \dots, \{i_n\}\}$, con cada cluster formado por un único elemento.
2. Construir una nueva partición, fusionando los dos clusters de la partición previa que más se parecen (de acuerdo con un cierto criterio que será función de δ y al que llamamos *criterio de agregación*).
Supongamos que los clusters a fusionar sean $\{i_j\}$ e $\{i_k\}$, y que forman la nueva clase $C' = \{i_j, i_k\}$, entonces,

$$\mathcal{P}_{n-1} = \{\{i_1\}, \{i_2\}, \dots, \{i_j, i_k\}, \dots, \{i_n\}\} = \{\{i_1\}, \dots, C', \{i_n\}\}$$

Si existe más de una pareja de clusters candidatos a ser agregados en una nueva clase, entonces se elegirá una al azar.
3. Repetir el segundo punto hasta que todos los clusters se hayan fusionado en uno solo que contenga a todos.

Para pasar de \mathcal{P}_ξ a la $\mathcal{P}_{\xi-1}$ se requiere que se especifique cómo determinar las entidades a fusionar en cada paso, así como la forma de efectuar esta fusión. Hay diversas posibilidades que dan lugar a diferentes formaciones de las clases. Es decir, según el criterio de agregación, la característica relevante de una clase variará. Entre otras, se pueden utilizar como selectores de los nodos a agregar: la distancia entre sus centros de gravedad (*Método del centroide* (Sokal and Michener 1958)) o entre sus medianas (*Método de la mediana*); la distancia entre los dos individuos más separados dentro de la clase (*Complete*); o la de los más próximos (*Single linkage*) o la distancia media entre los individuos de las dos clases (*Average linkage*) o Método de Ward, la mínima pérdida de inercia (Ward 1963); etc., siempre en términos de la distancia δ (Völle 1985).

Por otro lado, queda claro que uno de los puntos críticos de este algoritmo es la existencia de una medida δ que permita comparar objetos entre sí y decidir cuáles se parecen más y cuáles se parecen menos. De acuerdo con la naturaleza de los datos, δ tomará una forma u otra. En esta tesis se trabaja con *KLASS*(Gibert and Nonell 2005), ver Anexo A, que implementa varios algoritmos de clasificación ascendente jerárquica que se enmarcan en el esquema general que se acaba de presentar. En concreto se trata del algoritmo de los vecinos recíprocos encadenados (De Rham 1997) que es de complejidad algo menor que $O(n^2)$. Sobre

este esquema se implementa también el algoritmo de Clasificación basada en reglas (Gibert 1994) y el de clasificación condicionada (Raya 2007).

Klass además permite clasificar con el criterio del centroide (Sokal and Michener 1958) (el resultado de la fusión de 2 individuos es uno que tiene como coordenadas las del centro de gravedad de sus componentes) y el de Ward (Ward 1963)(decide qué dos elementos (ya sea objetos individuales o clases) deben fusionarse cada vez basándose, no en una noción pura de distancia, sino en el concepto de inercia) y métricas diferentes según la naturaleza de las variables de la matriz de datos (Gibert, Nonell, Velarde, and Colillas 2005).

En esta tesis se han usado 2 criterios para realizar las clasificaciones de las secciones §16 y §21; el criterio del centroide y el criterio de Ward, aunque en las clasificaciones elegidas para ser interpretadas se ha usado siempre el criterio de Ward por tener mejores propiedades y la métrica utilizada es Euclídea ascendente jerárquica clásica Normalizada (Diday and Moreau 1984), al trabajar con variables numéricas.

5.3 Dendrograma

El proceso de clasificación jerárquica se representa en forma de dendrograma que es un árbol (Dendro=árbol) binario que organiza los datos en subgrupos que se van uniendo de 2 en 2, hasta llegar al nivel de agrupamiento deseado (asemejándose a las ramas de un árbol que se van uniendo unas a otras sucesivamente hasta llegar al tronco). En cada paso se juntan las dos clases más próximas y se representa por un nuevo nodo. La altura del nuevo nodo interno del árbol viene dada por la distancia que separaba las dos clases que se agregan (esta distancia es cada vez mayor al progresar hacia la parte superior del árbol).

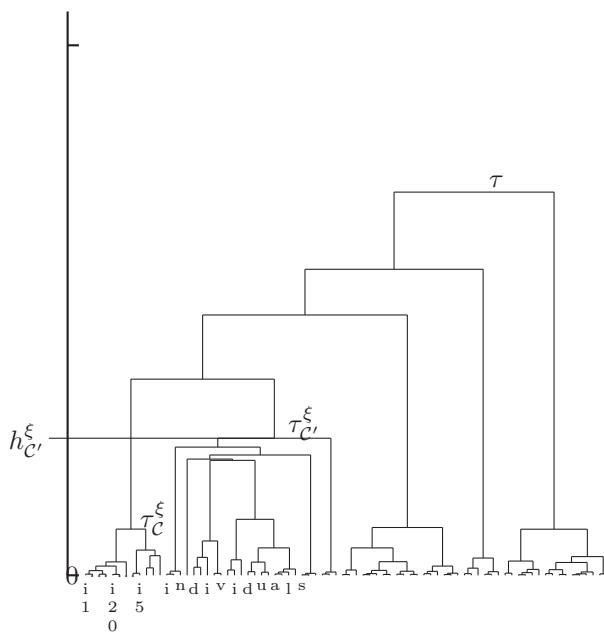


Figura 5.1: Estructura de τ .

Un dendrograma o árbol jerárquico se compone de:

- Una hoja para cada elemento de $\mathcal{I} = \{i_1, \dots, i_n\}$,
- Los nodos internos representan las diferentes subagrupaciones establecidas sobre los elementos,

- Las hojas del subárbol correspondiente a cada nodo interno son los elementos que forman parte de cada subclase,
- Las ramas de diferente longitudes sitúan los nodos internos a diferentes niveles respecto de la horizontal sobre la que se dispone los elementos. El nivel de los nodos, normalmente cifrado y conocido como índice de nivel (ver §5.4), indica el grado de semejanza (similitud) de sus hijos, y está relacionado directamente con la distancia entre ellos en el espacio de las variables. Cuanto más grande sea la longitud de la rama que une dos hijos a su padre, menos parecidas son las subclases que estos nodos representan.

Al cortar el árbol a determinado nivel horizontal se determina una partición \mathcal{P}_ξ de \mathcal{I} . Haciendo variar el nivel de corte, obtenemos diversas particiones. Con esto se tienen particiones de \mathcal{I} a diferentes grados de abstracción, y se podrá elegir la que más se ajuste a los propósitos del usuario como ya se ha comentado. Es a partir de un estudio a posteriori de este árbol que se encuentra la partición definitiva.

La Figura 5.1 representa un posible árbol jerárquico. La representación en dendrograma, como se puede ver, aporta mas información que la representación en forma de sucesión de particiones, dado que indica qué relación de semejanza hay entre los nodos que se agregan entre una partición \mathcal{P}_ξ y la siguiente $\mathcal{P}_\xi + 1$.

Criterios de corte: En cuanto a la determinación del nivel adecuado del corte, en (Milligan and Cooper 1985) se proponen pruebas para identificar el nivel de corte apropiado en una jerarquía. En cambio (Völle 1985) y (Lebart, Morineau, and Fenelon 1985) se inclinan por la construcción de una gráfica que visualice la evolución de los índices de nivel §5.4 de las sucesivas agregaciones. Las discontinuidades pronunciadas en esta gráfica se interpretan como agregaciones *forzadas* de clases diferentes (que generan grandes aumentos de la inercia intra-clase). Se recomienda cortar en uno de estos saltos siempre y cuando la partición resultante admita la interpretación. Dichos saltos coinciden con la discontinuidad en el cociente entre la varianza entre clases y la intra que optimiza a su vez un criterio de homogeneidad de las clases y separabilidad entre ellas. De hecho, la gráfica no hace más que representar un fenómeno que también es perceptible, de forma más velada, en el dendrograma. La Figura 5.2 muestra la gráfica correspondiente a la clasificación realizada con la base de datos de la aplicación que se presenta en el Capítulo §21.

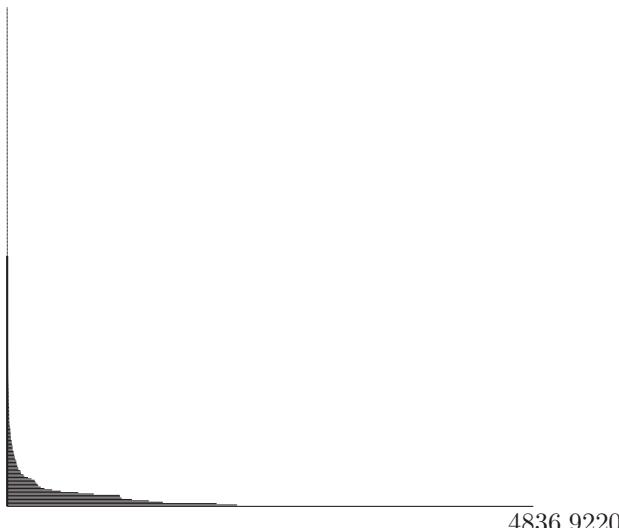


Figura 5.2: Gráfico de inercia interna de las clases $[\tau_{Lj3,R2}^{EnW,G}]$.

Por otro lado, *KLASS* implementa una heurística para determinar el nivel de corte de forma automática (Gibert 1994). La heurística es muy sencilla y consiste en partir de una lista con los índices de nivel de cada nodo, ordenados de mayor a menor, luego se calcula la diferencia entre el índice de nivel de un nodo y los siguientes y estas diferencias se ordenan de mayor a menor, asociando a cada una el número de clases del corte que determina y finalmente se realiza el corte para el mayor incremento de inercia, pero se indican los sucesores inmediatos para que el usuario pueda optar por otro si fuera necesario. Esta es una forma eficaz de identificar el punto donde el cociente de inercias es mayor.

5.4 Índice de nivel de la jerarquía

Puesto que se ha dado una idea de la distancia existente entre los hijos, el *índice de nivel* de un nodo de jerarquía $C = \{C_e, C_d\}$ se define habitualmente como $\nu(C) = d(C_e, C_d)$ y se representa como la ordenada de los nodos del dendrograma o árbol ascendente jerárquico. Evidentemente los nodos terminales tienen un índice cero.

5.5 Jerarquías indexadas (τ) y ultramétricas

Se denota $P_\xi(\mathcal{I})$ el conjunto de partes de \mathcal{I} .

Definición: $\tau \subset P_\xi(\mathcal{I})$ es una jerarquía de objetos sobre \mathcal{I} si (Cuadras 1991):

1. $\mathcal{I} \in \tau, \emptyset \notin \tau$
2. $\forall i \in \mathcal{I}, \{i\} \in \tau$
3. $\forall h_1, h_2 \in \tau$, se tiene $h_1 \cap h_2 = \emptyset \vee h_1 \subset h_2 \vee h_2 \subset h_1$
Si además,
4. $\forall h \in \tau$ no unitario, existen $h_1, h_2 \in \tau$ tales que $h_1 \cap h_2 = \emptyset$ y $h_1 \cup h_2 = h$,
se dice que τ es una jerarquía binaria.

La sucesión de particiones obtenida con un método ascendente jerárquico cuyo criterio de comparación entre objetos es una métrica, cumple que el índice ν de un nodo C tq $C \subset C' \Rightarrow \nu(C) < \nu(C')$ es una jerarquía indexada y los nodos del dendrograma tienen entonces propiedades de ultramétrica.

Si $\nu(C)$ satisface que $C \subset C' \Rightarrow \nu(C) < \nu(C')$, entonces se tiene una *jerarquía indexada* y construir una jerarquía indexada sobre el conjunto \mathcal{I} es equivalente a definir una ultramétrica sobre \mathcal{I} (Benzécri 1973). Las α -particiones de una jerarquía indexada tienen una interpretación clara como clasificaciones de \mathcal{I} . Es decir, que de alguna manera, encontrar una sobre \mathcal{I} garantiza la existencia de un dendrograma que permite hacer particiones sobre este conjunto.

Una ultramétrica se define como una función d sobre $\mathcal{I} \times \mathcal{I}$ con las siguientes propiedades:

1. $d(i, i') = 0 \iff i = i', \quad \forall i, i'$
2. $d(i, i') = d(i', i), \quad \forall i, i'$
3. $d(i, i') \leq \max\{d(i, i''), d(i', i'')\}, \quad \forall i, i', i''$

Así una ultramétrica siempre es una distancia, mientras que el recíproco no tiene porque serlo.

Se dice que un dendrograma presenta inversiones cuando no es una jerarquía indexada, en otras palabras, cuando hay nodos del dendrograma que tienen un índice de nivel menor que alguno de sus hijos:

$$\exists C \subset C' \text{ tq } \nu(C) > \nu(C')$$

Ello puede ocurrir si se clasifica con el criterio de centroide.

5.6 Boxplot simple

El *boxplot simple*, Figura 5.3, es una herramienta gráfica introducida por (Tukey 1977) y funciona del siguiente modo: el intervalo de valores que toma la variable se visualiza y las observaciones atípicas (outliers) se marcan con “*”. Se despliega una caja desde Q_1 (primer cuartil) hasta Q_3 (tercer cuartil) y la mediana se marca con un signo horizontal al centro de la caja. Las cajas incluyen, entonces, el 50% de los elementos y los bigotes se extienden hasta el valor mínimo y máximo que toma la variable.

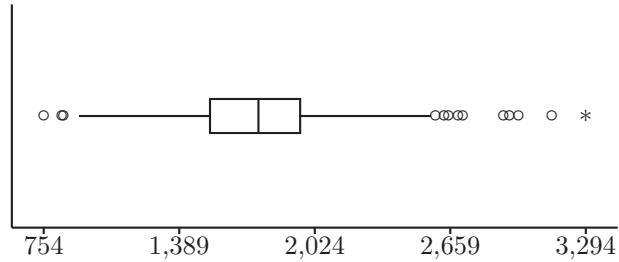


Figura 5.3: Boxplot de la variable MLSS-B.

Aparece como herramienta gráfica que resume la información suficiente sobre la distribución de la variable.

5.7 Boxplot múltiple

El *boxplot múltiple*, Figura 5.4, visualiza las distribuciones de una variable numérica condicionada a un conjunto de grupos (o clases) y, en consecuencia, permite analizar la relación entre ellos.

Para cada clase, se representa el boxplot de la variable numérica para el intervalo de valores tomados por la variable en esa clase de acuerdo a lo introducido en §5.6. Se yuxtaponen los boxplots de cada grupo con un eje común que mantiene la misma graduación y permite comparaciones. La yuxtaposición puede ser vertical u horizontal.

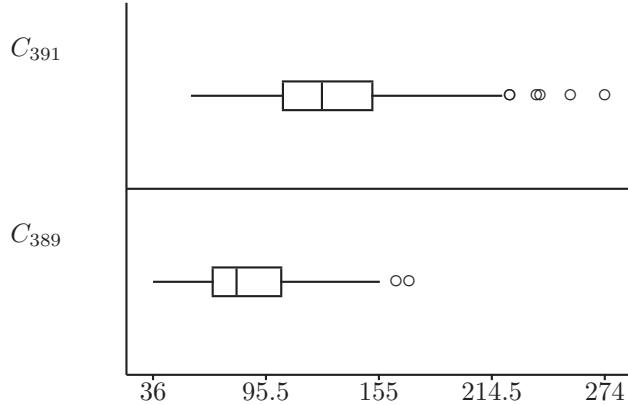


Figura 5.4: Boxplot múltiple de la variable DBO-D versus $\mathcal{P}_2 = \{C_{391}, C_{389}\}$.

En nuestro caso, estos gráficos se usan para visualizar la distribución de una variable numérica en todas las clases de \mathcal{P}_ξ (ver figura 5.4). De acuerdo con las ideas introducidas por Tukey en el campo del análisis descriptivo (Tukey 1977), se empezó por observar la representación gráfica de dichas distribuciones condicionales y se utilizó para extraer toda la información relevante sobre el problema. Así fue como el *box-plot múltiple* se hizo fundamental en el desarrollo de la metodología que se propone en esta tesis y fue la base para identificar, primero visualmente, variables *caracterizadoras* de una clase y para después definir cómo calcularlas (Gibert 2004).

5.8 Variables y valores caracterizadores de una clase

La idea es identificar lo que se denominan *variables caracterizadoras* de la clase C (Gibert and Cortés 1998b), concepto que descansa a su vez en el concepto de *valor propio* de una clase C . Las primeras formulaciones ya aparecen en (Gibert 1994). Existen diversos trabajos donde se elaboran éstos conceptos; (Gibert 1996b), (Gibert and Cortés 1998a) hasta que en (Gibert 2004) se presenta la formulación definitiva:

1. Valor propio: Un valor $c_s^k \in \mathcal{D}_k$ es *propio* de la clase C , si cumple:

$$(\exists i \in C : x_{ik} = c_s^k) \wedge (\forall i \notin C : x_{ik} \neq c_s^k)$$

Son valores que se dan de forma exclusiva en una clase. Estos valores, cuando ocurren, identifican una clase con toda seguridad, por lo que, actúan como *valores caracterizadores* de C y los denotamos por λ_{sc}^k .

Llamaremos Λ_C^k al conjunto de los *valores propios* de la variable X_K para la clase C .

2. Valor caracterizador: $\lambda \in \Lambda_C^k$, y λ se puede utilizar para identificar la clase entera o parte de ella, dependiendo de si existen otros valores de X_K en C . Ello depende de si existe no interacción entre clases. A lo largo de los diferentes estudios se ha visto que hay 4 tipos de valores, ver Tabla 5.2, los dos últimos explicitados en trabajos posteriores (Vázquez Torres 2002).

- (a) λ es un valor *parcialmente caracterizador* de C si $\{i \in C : x_{ik} = \lambda\} \subset C$. Sea V_C^k , el conjunto de valores *parcialmente caracterizadores* de C . Se da exclusivamente en una clase pero no la cubre.
- (b) λ es un valor *totalmente caracterizador* de C si $\{i \in C : x_{ik} = \lambda\} = C$. Se da exclusivamente en una clase y la cubre totalmente.
- (c) Un valor *caracterizador no propio* de C es aquel que se da en la clase y la cubre entera pero no es exclusivo de ella.
- (d) Un valor *genérico* de la clase C se da en la clase, no la cubre y no es exclusivo de la clase.

	Tipo de Valor	Cobertura de la Clase	
		Cubre todo C	Cubre sólo parte C
Interacción con otras clases	Exclusivo	<i>Caracterizador total</i>	<i>Caracterizador parcial</i>
	No exclusivo	<i>Caracterizador no propio</i>	<i>Genérico</i>

Tabla 5.2: Valores caracterizadores.

3. Variables caracterizadoras: Las *variables caracterizadoras*, según (Gibert 1994), son “*las variables más relevantes en cada una de las clases formadas; dicho de otra forma, las que han resultado más decisivas en la construcción de éstas y, eventualmente, permiten detectar la pertenencia de un objeto a una clase determinada, excluyéndolo de las restantes*”, ver Tabla 5.3.
- (a) Variable parcialmente caracterizadora: X_K es *parcialmente caracterizadora* de la clase $C \in \mathcal{P}$ si tiene al menos un valor propio de la clase C , ($\Lambda_C^k \neq \emptyset$) y ($V_C^k \neq \Lambda_C^k$), aunque puede compartir alguno con otra(s) clase(s).
- (b) Variable totalmente caracterizadora: X_K es *totalmente caracterizadora* de la clase $C \in \mathcal{P}$, si todos los valores que toma X_K en C son *propios* de C , es decir, no existen objetos de otras clases que tomen esos valores. Sea $Ext(\Lambda_C^k) = \{i \in \mathcal{I} \text{ tq } x_{ik} \in \Lambda_C^k\}$, si $Ext(\Lambda_C^k) = C$, X_K es totalmente caracterizadora de C .

	Tipo de Valor	Cobertura de la Clase	
		Cubre todo C	Cubre sólo parte C
Interacción con otras clases	Exclusivo	$Ext(\Lambda_C^k) = C$	$\Lambda_C^k \neq \emptyset \wedge V_C^k \neq \Lambda_C^k$
	No exclusivo	$Ext(\Lambda_C^k) \supseteq C \wedge card(\Lambda_C^k) = 1$	$Ext(\Lambda_C^k) \supsetneq C \wedge card(\Lambda_C^k) > 1$

Tabla 5.3: Variables caracterizadoras.

4. Grado de caracterización: Sea $(1 - \varepsilon)$, $\varepsilon \in [0, 1]$ el grado de caracterización de una clase C , para un valor. Dada una variable X_K ,

Un valor $(1 - \varepsilon)$ —*caracterizador* de C es aquel *valor propio* de C que sólo identifica $(1 - \varepsilon)\%$ de C .

$$Card(Ext(\Lambda_C^k)) = (1 - \varepsilon)Card(C)$$

Ya en (Gibert 1994) aparece la idea de $(1 - \varepsilon)$ —*caracterización*. Ello conduce a situaciones en apariencia complejas como el hecho de que X_K sea $(1 - \varepsilon_1)$ —*caracterizadora* de C y también $(1 - \varepsilon_2)$ —*caracterizadora* de C' con $\varepsilon_1 \neq \varepsilon_2$.

En realidad esto sucede porque lo que determina el poder de caracterización no es la variable en sí, sino los valores que toma y su distribución a lo largo de las clases.

Con toda esta información se establece la siguiente relación entre la pertenencia a una clase C de un objeto i y sus valores en X_K ó c_s^k , ver Tabla 5.4.

	Tipo de Valor	Cobertura de la Clase	
		Cubre todo C	Cubre sólo parte C
Interacción con otras clases	Exclusivo	$i = c_s^k \Leftrightarrow i \in C$	$i = c_s^k \Rightarrow i \in C$ $i = c_s^k \Leftarrow i \in C$
	No exclusivo	$i = c_s^k \Leftarrow i \in C$ $i = c_s^k \Rightarrow i \in C$	$i = c_s^k \nLeftrightarrow i \in C$

Tabla 5.4: relación de valores caracterizadores y una clase C .

5.8.1 Alcance

Las variables caracterizadoras se utilizaron para definir un primer procedimiento de caracterización para detectar conjuntos mínimos de variables que distingan una clase de otra

utilizando únicamente variables cualitativas (Gibert and Cortés 1998b). Ello pasa por estudiar cómo interaccionan las clases. Para el estudio de *interacciones entre clases* se consideran las variables, en su estado natural, evitando cualquier transformación arbitraria sobre su naturaleza, que pudieran alterar el sentido de la iteracción. Esto consiste en identificar todas las intersecciones que se dan entre los valores de las variables y las distintas clases, determinando en qué puntos del rango de las variables están cambiando estas intersecciones; así se puede identificar las distintas combinaciones de clases donde se puede dar un mismo valor de cierta variable y como consecuencia hacer emergir los valores caracterizadores de una clase; éstos identificarán variables total o parcialmente caracterizadoras.

Los conceptos definidos en (Gibert 1996b) son formales y generales y se propone un método de caracterización basado en hallar los valores propios de las clases.

El cálculo de valores propios de una clase descansa necesariamente en las distribuciones condicionadas de cada variable en dicha clase.

En un principio de analizaba el boxplot múltiple de forma visual, ya que es muy fácil observar si el boxplot de cierta clase no interseca con el de las demás, ver Figura 5.5, y gráficamente es fácil ver los valores propios. Sin embargo, en la práctica no se puede basar un proceso automático en la interpretación de una representación gráfica, por lo que en (Vázquez and Gibert 2001) se propone una primera alternativa equivalente, pero automatizable, un sistema de intervalos o ventanas de longitud variable sobre el que hacer contajes relacionados con las distribuciones condicionadas y su interacción.

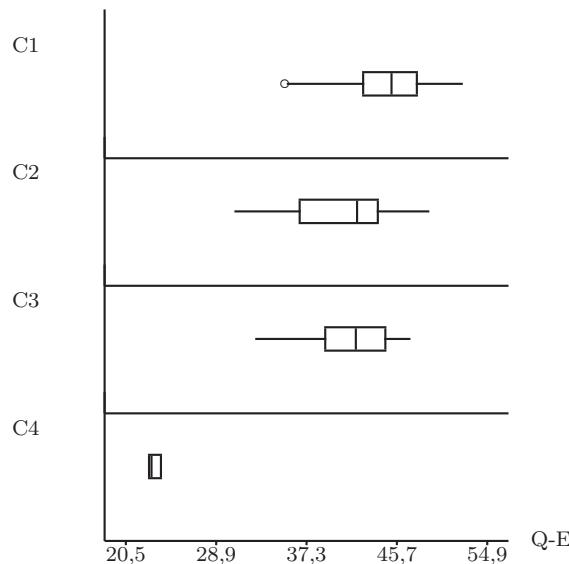


Figura 5.5: Boxplot múltiple de la variable Q_E vs partición en 4 clases.

Como en lo habitual se encuentran pocos valores *totalmente caracterizadores*, en sentido estricto, lo común, son los valores *parcialmente caracterizadores* (Vázquez Torres 2002). Es decir, valores que determinan parte de una clase, la cual tiene que cuantificarse para poder determinar el poder de caracterización de dichos valores. No existe garantía alguna de que existan valores propios en una clase cualquiera, lo que en seguida requiere plantear propuestas de solución cuando nos encontramos con este caso.

5.9 Boxplot based Discretization (BbD)

Como ya hemos dicho, para identificar las variables total o parcialmente caracterizadoras se estudian los valores propios que toma una variable X_K en una clase C , en relación a las otras

y se ve si son de la clase o no; para ello se analiza cómo son las interacciones entre clases. En (Gibert and Cortés 1998b) se propone un método para asociar a una variable numérica una categorización que facilite el análisis de valores propios, que en ese caso se deberían llamar propiamente *intervalos propios*.

El *Boxplot based Discretization (BbD)* presentado formalmente por primera vez en (Gibert 2004) es un método para asociar a una variable numérica una categorización que permita identificar dónde cambian las intersecciones entre clases con respecto a X_K y facilite el análisis de valores propios. Los puntos donde varían las intersecciones entre clases se pueden encontrar de forma exacta con un coste computacional mínimo, solamente calculando los valores mínimos y máximos por variable y clase y ordenándolos en forma conveniente. A partir de dicha ordenación, se define una discretización de la variable en un conjunto de intervalos, sobre los que se podrá identificar los valores propios de una variable en todas las clases.

Dada una variable numérica X_K se transformará en una variable categórica cuyos valores \mathcal{D}^k correspondan a un sistema de intervalos inducidos por \mathcal{P}_ξ sobre X_K utilizando el método del *Boxplot based Discretization* que se describe a continuación.

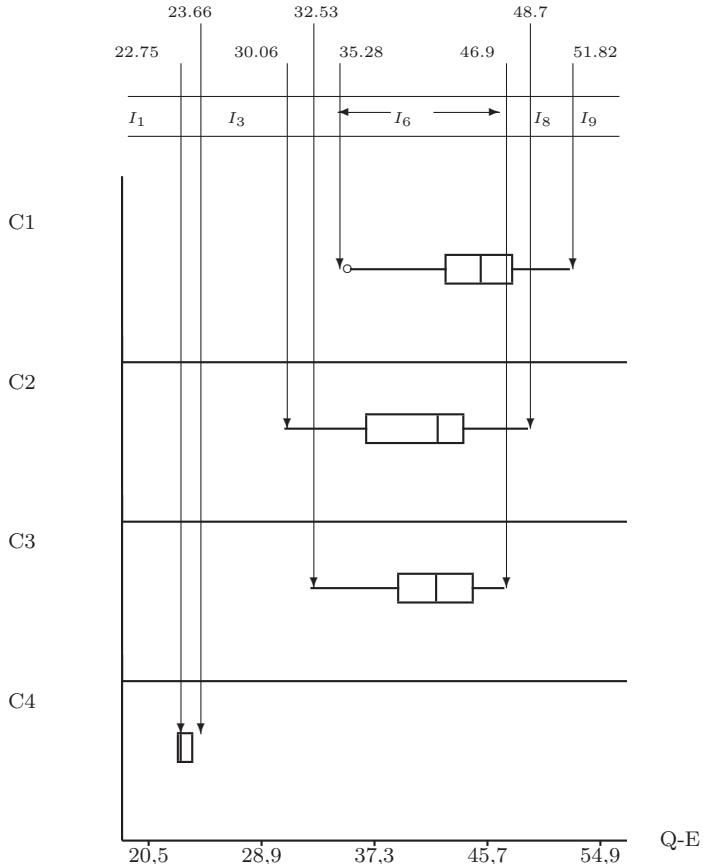


Figura 5.6: Boxplot múltiple de la variable Q_E vs partición en 4 clases, con ventanas de longitud variable para inducir la codificación I^k de Q_E .

Sean,

$$m_C^k = \min X_K | C = \min_{i \in C} \{x_{ik}\} \text{ y } M_C^k = \max X_K | C = \max_{i \in C} \{x_{ik}\}.$$

los mínimos y los máximos observados para la variable X_K en la clase $C \in \mathcal{P}_\xi$. Los pasos a seguir según se propone en (Gibert and Roda 2000) son los que se describen a continuación (para detalles ver (Vázquez and Gibert 2001)):

1. Construir $\mathcal{M}^k = \{m_{c_1}^k, \dots, m_{c_\xi}^k, M_{c_1}^k, \dots, M_{c_\xi}^k\}$, siendo la $card(\mathcal{M}^k) = 2\xi$
2. Construir el *conjunto de puntos de corte*
 \mathcal{Z}^k ordenando \mathcal{M}^k de menor a mayor q $\mathcal{Z}^k = \{z_i^k ; i = 1 : 2\xi\}$, tal que:
 i) $z_1^k = \min \mathcal{M}^k$
 ii) $z_i^k = \min(\mathcal{M}^k \setminus \{z_j^k ; j < i\})$, $i = \{2, \dots, 2\xi\}$
 $\mathcal{Z}^k = \{z_i^k\}$ es un conjunto tal que: $\mathcal{Z}^k = \{z_j^k | z_{j-1}^k < z_j^k; 1 < j \leq 2\xi\}$
3. Construir el *sistema de intervalos I^k inducido por \mathcal{P}_ξ sobre X_K* , en la siguiente forma:

$$\mathcal{D}^k = \{I_s^k : 1 \leq s \leq 2\xi - 1\}$$

donde

- i) $I_1^k = [z_1^k, z_2^k]$.
- ii) $I_s^k = (z_s^k, z_{s+1}^k)$, ($s = 2 : 2\xi - 1$), I_s^k siendo intervalos de longitud variable.
4. Definir la nueva variable categórica I^k cuyo conjunto de valores es $\mathcal{D}^k = \{I_1^k, \dots, I_{2\xi-1}^k\}$.
 Siendo $D^k = \cup_{s=1}^{2\xi-1} I_s^k = [z_1^k, z_{2\xi-1}^k] = [\min_{\forall i \in \mathcal{I}} x_{ik}, \max_{\forall i \in \mathcal{I}} x_{ik}] = r_k$ el dominio de X_K , \mathcal{D}^k representa una categorización del mismo, con $card(\mathcal{D}^k) = 2\xi - 1$ valores, pero no es arbitraria en absoluto sino que identifica *todas las intersecciones entre clases* según X_K , y además se calcula de forma inmediata.

Así, se tiene 2ξ puntos de corte diferentes se generán como máximo $2\xi - 1$ intervalos y la $card(\mathcal{D}^k) = 2\xi - 1$, recordando que ξ es el número de clases de la partición de referencia que se quiere caracterizar.

Por último, hay que observar que para construir I^k ya no hace falta realizar el *box-plot múltiple*, aunque éste sigue siendo una excelente representación de lo que se está haciendo.

La idea central se presenta en la Figura 5.6. Identificando los valores extremos de las distribuciones condicionales de X_K en cada clase (lo que es directo sobre el boxplot múltiple) y ordenándolos en una lista global, se puede determinar un conjunto de intervalos en el rango de la variable numérica que se puede utilizar como codificación de la misma. Son intervalos de longitud variable tales que:

- Identifican *todos* los valores de las variables numéricas donde la superposición entre clases varía; donde cambia, en definitiva, la *aridad* de la intersección entre clases.
- Se pueden computar independientemente de la representación gráfica fácilmente y con poco coste computacional, tras una simple ordenación, sin necesidad de entrar en el estudio de intersecciones $n \times n$, de coste combinatorio.

A partir de lo anterior, la idea fundamental, es usar \mathcal{D}^k para caracterizar las clases de \mathcal{P}_ξ . Para ello se busca si \mathcal{D}^k tiene algún valor *propio* o *parcialmente caracterizador* en alguna clase.

A partir del número de observaciones n_{sc} que simultáneamente están en C y cumple que $X_K \in I_s^k$. Si $n_{sc} = card\{i \in C \wedge x_{ik} \in I_s^k\}$ se puede observar que la característica de un valor *propio* o *parcialmente caracterizador* de \mathcal{D}^k en la clase C , es tal que cumple :

1. I_s^k es valor *parcialmente caracterizador* de la clase C si:
 (a) $n_{sc} \neq 0$ y

(b) $\forall C' \neq C, n_{sc'} = 0$

2. I_s^k es un valor *totalmente caracterizador* de la clase C si:

(a) $\forall s' \neq s$ y

(b) $n_{s'c} = 0$

3. I_s^k es un *valor caracterizador no propio* de la clase C si:

(a) $n_{sc} \neq 0$ y

(b) $\forall s' \neq s \ n_{s'c} = 0$

A partir de estos intervalos, se pueden calcular las *variables caracterizadoras* de una clase, ya sea totales o parciales, y en ese último caso, las probabilidades condicionales se pueden utilizar para expresar la incertidumbre ligada a la interpretación generada (a la inducción).

5.10 Boxplot based Induction Rules (BbIR)

En (Gibert, Aluja, and Cortés 1998) se describe la forma de caracterizar una clasificación utilizando los representantes de clase a partir de variables cualitativas y se presenta la primera versión sobre el uso de condicionamientos sucesivos, en ese caso utilizando una hipótesis de mundo cerrado y conjuntandos negativos en los conceptos generados, como primera aportación al manejo de interacciones de orden superior a dos con variables cualitativas.

Ya se usan versiones previas del BbIR y se combina la inducción de todas las variables numéricas. Algunos trabajos previos como (Gibert 1996a) y (Gibert and Salvador 2000) se presenta el método general utilizando variables numéricas versus una partición, donde se puede asociar a un objeto cualquiera su grado de pertenencia a cada clase mediante la probabilización de las reglas generadas; y en (Gibert and Roda 2000) se utiliza el método también con variables numéricas logrando identificar variables totalmente caracterizadoras para obtener una base de conocimientos que permite generar una primera interpretación de una partición en 4 clases validada por el experto.

La *inducción de reglas basada en box-plots (Boxplot-based Induction Rules)* presentada formalmente en (Gibert 2004), se basa en una idea muy simple, que mimetiza bastante bien lo que los expertos realmente hacen cuando interpretan un boxplot múltiple.

Una descripción breve del método es la siguiente:

1. Para todas las variables numéricas de $C \in \mathcal{P}_\xi$, obtener con el *BbD* (ver §5.9) el sistema de intervalos $\mathcal{D}^k = \{I_1^k, \dots, I_{2\xi-1}^k\}$:

(a) Construir $\mathcal{M}^k = \{m_{c_1}^k, \dots, m_{c_\xi}^k, M_{c_1}^k, \dots, M_{c_\xi}^k\}$, siendo la $card(\mathcal{M}^k) = 2\xi$

(b) Construir el *conjunto de puntos de corte*,

\mathcal{Z}^k ordenando \mathcal{M}^k de menor a mayor $\mathcal{Z}^k = \{z_i^k ; i = 1 : 2\xi\}$, tal que:

i) $z_1^k = \min \mathcal{M}^k$.

ii) $z_i^k = \min(\mathcal{M}^k \setminus \{z_j^k ; j < i\})$, $i = \{2, \dots, 2\xi\}$.

$\mathcal{Z}^k = \{z_i^k\}$ es un conjunto tal que: $\mathcal{Z}^k = \{z_j^k | z_{j-1}^k < z_j^k; 1 < j \leq 2\xi\}$.

(c) Construir el *sistema de intervalos I^k inducido por \mathcal{P}_ξ sobre X_K* , en la siguiente forma:

$$\mathcal{D}^k = \{I_s^k : 1 \leq s \leq 2\xi - 1\}$$

donde

- i) $I_1^k = [z_1^k, z_2^k]$.
- ii) $I_s^k = (z_s^k, z_{s+1}^k], \quad (s = 2 : 2\xi - 1), I_s^k$, siendo intervalos de longitud variable.
- (d) Definir la nueva variable categórica I^k cuyo conjunto de valores es $\mathcal{D}^k = \{I_1^k, \dots, I_{2\xi-1}^k\}$, con $card(\mathcal{D}^k) = 2\xi - 1$ y hacer $x_{iI^k} = I_s^k \quad tq \quad x_{ik} \in I_s^k$.

2. Para todas las variables:

- (a) Si X_K es numérica, $\mathcal{D}^k = \{I_1^k, \dots, I_{2\xi-1}^k\}$

Construir la tabla de frecuencias de las clases condicionadas a los intervalos:

$\mathcal{I}^k \mathcal{P}_\xi$	C_1	C_2	\dots	C_ξ
I_1^k	p_{11}	p_{12}		
I_2^k				
\vdots				
$I_{2\xi-1}^k$		p_{sc}		$p_{(2\xi-1)\xi}$
	1	1		1

donde $p_{sc} = \frac{card\{i : i \in C \wedge x_{ik} \in I_s^k\}}{card\{i : x_{ik} \in I_s^k\}}$
y el caracterizador total es tal que $p_{sc} = 1$

- (b) Si X_K es categórica, $\mathcal{D}^k = \{I_1^k, \dots, I_{n_k}^k\}$, donde n_k es el número de modalidades de la variable categórica X_K y I_s^k será una modalidad de X_K .

Construir la tabla de frecuencias de las clases condicionadas a las categorías de la variable:

$\mathcal{I}^k \mathcal{P}_\xi$	C_1	C_2	\dots	C_ξ
I_1^k	p_{11}	p_{12}		
I_2^k				
\vdots				
$I_{n_k}^k$		p_{sc}		$p_{n_k\xi}$
	1	1		1

donde $p_{sc} = \frac{card\{i : i \in C \wedge x_{ik} = I_s^k\}}{card\{i : x_{ik} = I_s^k\}}$
y el caracterizador total es tal que $p_{sc} = 1$

3. Identificar las frecuencias condicionadas empíricas con grados de certeza. Esto se puede representar gráficamente y puede ser utilizado como una herramienta de soporte a la interpretación.

La figura 5.7 muestra los grados de pertenencia de una variable X_K en una partición de 4 clases.

- 4. Para cada casilla no vacía de la tabla $\mathcal{P}_\xi | \mathcal{I}^k$ construir una regla probabilística tal como:

$$\mathcal{R} = \{r : x_{ik} \in I_s^k \xrightarrow{p_{sc}} C : \forall p_{sc} > 0\}$$
- 5. Finalmente, si se requieren decisiones crisp, decidir el nivel de incertidumbre α y cortar de \mathcal{R} todas las reglas con un grado de incertidumbre menor que α para tener una interpretación automática de \mathcal{P}_ξ , a nivel α .

A partir de los conceptos anteriores, se facilita aún más la identificación de los valores *caracterizadores* (Vázquez and Gibert 2001). Así, tenemos que:

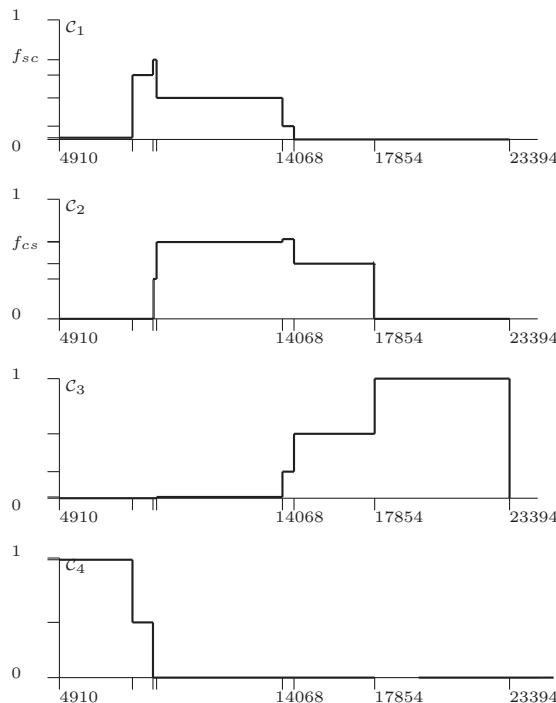
- Un valor *totalmente caracterizador* de C es: $I_s^k \quad tq \quad p_{sc} = 1, \quad p_{s'c} = 0, \quad \forall s' \neq s$.

- Un valor *parcialmente caracterizador* de C es: $I_s^k \text{ tq } p_{sc} = 1, p_{s'c} > 0, \forall s' \neq s$.
- Un valor *caracterizador no propio* de C es: $I_s^k \text{ tq } p_{sc} \in (0, 1), p_{s'c} = 0, \forall s' \neq s$.
- Un *valor genérico* de la clase C es: $I_s^k \text{ tq } p_{sc} \in (0, 1)$ y $\exists s'$ tal que $p_{s'c} > 0, s' \neq s$ y $\exists c'$ tal que $p_{sc'} > 0, c' \neq c$. Estos valores se pueden interpretar como el subconjunto de individuos i de la clase C que comparten su valor I_s^k tanto con las demás clases, existiendo a su vez en la misma clase C algunos otros elementos que pertenecen a otros intervalos.

	Tipo de Valor	Cobertura de la Clase	
		Cubre todo C	Cubre sólo parte C
Interacción con otras clases	Exclusivo	$I_s^k \text{ tq } p_{sc} = 1 \wedge p_{s'c} = 0, \forall s' \neq s$	$I_s^k \text{ tq } p_{sc} = 1 \wedge p_{s'c} > 0, \forall s' \neq s$
	No exclusivo	$I_s^k \text{ tq } p_{sc} \in (0, 1) \wedge p_{s'c} = 0, \forall s' \neq s$	$I_s^k \text{ tq } p_{sc} \in (0, 1) \wedge p_{s'c} > 0, \forall s' \neq s$

Tabla 5.5: Relación entre valores caracterizadores y p_{sc} .

Diagrama de grados de pertenencia. La idea de asociar a un objeto cualquiera su grado de pertenencia a cada clase a partir de la tabla de frecuencias de las clases condicionadas a los intervalos da lugar a un gráfico de grados de pertenencia que se podía adaptar fácilmente al paradigma difuso (Gibert and Salvador 2000) para cada clase y para cada variable como se puede ver en la Figura 5.7 y que lleva a elementos de razonamiento aproximado. Cabe mencionar, que el área bajo estas funciones ya no es 1, puesto que se componen de probabilidades que provienen de distintas distribuciones condicionadas (las de $C|\mathcal{I} = I_s^k, \forall s$).

Figura 5.7: Grados de pertenencia de X_K a una partición \mathcal{P}_4 en 4 clases.

5.11 Uso de los condicionamientos sucesivos en la interpretación de clases

Como ya se ha dicho, en (Gibert, Aluja, and Cortés 1998) se describe la forma de caracterizar una clasificación utilizando los representantes de clase a partir de variables cualitativas ($i \in C \iff (x_{ik}, x_{ik'}) \in (\Lambda_C^k \times \Lambda_{C'}^k)$).

Se presenta la primera versión sobre el uso de condicionamientos sucesivos, en ese caso utilizando una hipótesis de mundo cerrado y conjuntandos negativos en los conceptos generados, como primera aportación al manejo de interacciones de orden superior a dos con variables cualitativas. En este trabajo aparece por primera vez la idea de buscar jerárquicamente los caracterizadores pero no utilizando el dendrogram subyacente al proceso de clustering realizado, sino descartando iterativamente las clases que ya se pudieron caracterizar a base de valores totalmente caracterizadores.

La idea es identificar, si es posible, la variable X_K y los valores de \mathcal{D}_C^k que permitan la identificación de cada clase ($i \in C \iff x_{ik} \in \Lambda_C^k$). En ocasiones hace falta un par de variables para distinguir una cierta clase de las demás.

La propuesta consiste en introducir la *negación* de la información y utilizar un método basado en el uso de condicionamientos sucesivos de las descripciones. Las caracterizaciones de las clases que se dan en términos de lógica y la semántica es evidente para el experto. La caracterización automática de la partición estudiada en (Gibert, Aluja, and Cortés 1998) se basa en hallar $C \in \mathcal{P}_\xi$ con valores propios $tq i \in C \iff \Lambda_C^k$ y definir C a través del concepto de $X_K \in \Lambda_C^k$ y $\mathcal{P}'_\xi = \{C' \in \mathcal{P}_\xi \text{ tq } C \neq C'\}$ a través del concepto $X_K \notin \Lambda_C^k$. A través de la repetición consecutiva del proceso para \mathcal{P}'_ξ y la concatenación de caracterizadores se genera la descripción final. Pero los trabajos sucesivos en el área (Gibert and Roda 2000), (Gibert and Salvador 2000), (Vázquez and Gibert 2001) pone en evidencia que si lo que caracteriza una clase es una interacción de variables de orden superior el método presenta dificultades para alcanzar caracterizaciones consistentes. Es así como en esta tesis (Gibert and Pérez-Bonilla 2005c), (Gibert and Pérez-Bonilla 2006b), (Pérez-Bonilla and Gibert 2006), (Gibert, Pérez-Bonilla, and Rodríguez 2006), (Pérez-Bonilla and Gibert 2007b), (Pérez-Bonilla and Gibert 2007c), (Pérez-Bonilla, Gibert, and Vrecko 2008) se empieza a explorar una posibilidad nueva que es afrontar el problema de forma paulatina aprovechando que se obtuvo la partición con un método de clasificación jerárquico.

5.12 Lógica proposicional

En el Capítulo 3 ya se ha presentado el concepto de la Lógica proposicional, en esta sección se describe las propiedades necesarias para manejar en el desarrollo y evaluación de la propuesta metodológica.

Proposiciones Una proposición en lógica clásica, para los fines de esta exposición, es una declaración la cual puede evaluar a verdadero o falso ante un cierto estado del universo que modela, por ejemplo: $5 > 4$, $2 + 2 = 5$, “Pedro comió a las 3”, “Me gusta la sopa”. Algunas veces es más difícil que otras determinar si la declaración (o proposición) es verdadera o falsa, en otras palabras, si toma el valor de verdad o falsedad. La lógica proposicional es un lenguaje formal que se puede describir con una gramática BNF. El Backus-Naur form (BNF) (también conocido como Backus-Naur formalism, Backus normal form o Panini-Backus Form) es una metasintaxis usada para expresar gramáticas libres de contexto: es decir, una manera formal de describir lenguajes formales.

Sintaxis y notación

1. Sintaxis: El primer paso en el estudio de un lenguaje es definir los símbolos básicos que lo constituyen (alfabeto) y cómo se combinan para formar sentencias. La teoría sobre un dominio en lógica proposicional está constituido por:
 - Símbolos de veracidad: \top para verdadero y \perp para falso. Alternativamente se pueden usar V para verdadero y F para falso.
 - Símbolos de variables: p, q, \dots, z
 - Símbolos de conectivas: Negación (\neg), Conjunción (\wedge), Disyunción (\vee), Implicación (\rightarrow), Coimplicación (\leftrightarrow).
 - Símbolos: paréntesis (), corchetes [] y llaves {} para evitar ambigüedades.
2. Reglas de formación: Las clases de sentencias bien formadas se definen por reglas puramente sintácticas, llamadas reglas de formación, y que son:
 - Una variable proposicional es una sentencia (también llamada fórmula) bien formada. Son sentencias bien formadas: $\neg p$, $p \vee q$, $p \wedge q$, $p \rightarrow q$, $p \leftrightarrow q$. Si p y q son a su vez sentencias bien formadas.
 - A las conjunciones y disyunciones se les puede permitir tener más de dos argumentos.

Tablas de verdad Las tablas de verdad permiten evaluar sentencias compuestas y bien formadas a partir de los valores de las variables que las componen.

p	$\neg p$	p	q	$p \vee q$	$p \wedge q$	$p \rightarrow q$	$p \leftrightarrow q$
V	F	V	V	V	V	V	V
F	V	F	V	V	F	V	F
		V	F	V	F	F	F
		F	F	F	F	V	V

Tabla 5.6: Negación (\neg), Disyunción (\vee), Conjunción (\wedge), Condicional (\rightarrow) y Bicondicional (\leftrightarrow).

Leyes de Morgan (1806-1871)

1. Ley de morgan 1: $\neg(A \wedge B) \equiv \neg A \vee \neg B$.
2. Ley de morgan 2: $\neg(A \vee B) \equiv \neg A \wedge \neg B$.

Axiomas y reglas Los axiomas para el cálculo proposicional (Farré, Nieuwenhuis, Nivela, Oliveras, and Rodríguez 2008) son:

Sean P, Q y R sentencias, entonces;

1. Axioma de idempotencia: $(P \vee P) \rightarrow P$.
2. Axioma de adjunción: $P \rightarrow (P \vee Q)$.
3. Axioma de conmutatividad: $(P \vee Q) \rightarrow (Q \vee P)$.
4. Axioma de adición: $(P \rightarrow Q) \rightarrow [(R \vee P) \rightarrow (R \vee Q)]$.

A partir de estos axiomas y aplicando las dos reglas de transformación siguientes se puede demostrar cualquier teorema:

1. Regla de sustitución: el resultado de reemplazar cualquier variable en un teorema por una sentencia bien formada es un teorema.
2. Regla de separación: si S y $(S \rightarrow R)$ son teoremas, entonces R es un teorema.

Relativo a un criterio de validación, un sistema axiomático debería cumplir las siguientes propiedades para ser un sistema perfecto:

- Debe ser lógico o razonable: en el sentido de que todo teorema o es un axioma o la última secuencia de una deducción que se sigue de operaciones lógicas deductivas según las reglas especificadas.
- Completo: toda sentencia bien formada válida es un teorema y se debe poder demostrar a partir de los axiomas.
- Consistente: no se pueden demostrar como teoremas, sentencias bien formadas que no sean tautologías.
- Deben ser independientes: ningún axioma debe ser derivable a partir de los otros. Sin embargo, el teorema de Gödel demuestra que tal sistema perfecto no es posible.

Parte II

Marco Teórico y Desarrollo de la Metodología

Capítulo 6

Introducción

Como ya se ha dicho, en un proceso de clasificación automática en que se descubren las clases que componen un determinado dominio, quizás uno de los problemas más importantes y menos sistematizados que hay que enfrentar es el proceso de *interpretación* de las clases, íntimamente ligado a la *validación* de las mismas (Gordon 1994) y decisivo en la posterior *utilidad* del conocimiento adquirido. La interpretación de las clases, tan fundamental para entender el significado de la clasificación obtenida y, en consecuencia, la estructura del dominio, se realiza habitualmente de forma muy artesanal. Pero este proceso se complica enormemente conforme el número de clases crece. Con la Metodología de Caracterización Conceptual por Condicionamientos Sucesivos (*CCCS*) (Pérez-Bonilla, Gibert, and Vrecko 2008) se pretende abordar el problema de la generación automática de interpretaciones de una clasificación, con el objetivo de consolidar, a largo plazo, una herramienta que dé apoyo a esta tarea y contribuya a la sistematización de la misma.

Esta metodología aprovecha la estructura jerárquica de la clasificación objetivo para inducir conceptos iterando sobre las divisiones binarias de la secuencia de clases que indica el dendrograma. Un paso fundamental es definir cómo se integra el conocimiento inducido a cada iteración final de una única conceptualización de la partición a interpretar.

En el capítulo §11 se encuentra la metodología de caracterización conceptual por condicionamientos sucesivos (*CCCS*), resultado final de ésta investigación. La propuesta metodológica de Caracterización Conceptual por Condicionamientos Sucesivos (*CCCS*) permite construir una base de conocimiento con tantos conceptos como clases se quieren interpretar, siempre y cuando procedan de un *clustering* jerárquico y con la propiedad de que la aridad de cada concepto será la profundidad de la clase en el dendrograma.

Para elaborar la propuesta final, se ha trabajado en distintos aspectos que se presentan en la primera parte de este apartado en el Capítulo §7 en donde en primer lugar se hace una tipificación de los sistemas de reglas inducidos por el BbIR, presentado en el Capítulo §5, a continuación se presentan los cardinales y propiedades de los sistemas de reglas y luego una revisión del Boxplot based induction rules (BbIR) para el caso particular de particiones binarias.

Una idea central en la que se basa el desarrollo de la metodología es que la existencia de una jerarquía indexada de clases permite abordar el problema de la interpretación de forma recursiva descendiendo en el dendrograma, y por ello en el capítulo §8 se presentan las ventajas de trabajar con una jerarquía indexada de clases lo que permitirá abordar el problema de la interpretación de forma recursiva. Lo anterior reduce cada iteración a la interpretación de un partición binaria y por ello buena parte del trabajo que aquí se presenta hace referencia al caso particular de particiones binarias simplificando el problema de hallar distintivos en las clases.

Seguidamente se definen, en el capítulo §9, los criterios con los que se selecciona el o los

mejores conceptos (reglas probablizadas) asociado a las clases que se quieren interpretar.

Por último, como la propuesta se fundamenta en el estudio de la distribución conjunta entre pares de clases. Lo anterior da lugar a descripciones conceptuales que contienen la conjunción de dos atributos en cada clase, es por ello que en el capítulo §10, se proponen 5 formas de combinar el conocimiento, extraído en forma de conceptos (reglas probablizadas), en cada iteración.

Capítulo 7

Marco teórico de la metodología

Una idea central de la tesis es que la existencia de una jerarquía indexada de clases permite abordar el problema de la interpretación de forma recursiva descendiendo en el dendograma, ver capítulo §8. Ello reduce cada iteración a la interpretación de un partición binaria y por ello buena parte del marco teórico de la metodología que se presenta a continuación hace referencia al caso particular de particiones binarias simplificando el problema de hallar distintivos en las clases.

La propuesta metodológica de Caracterización Conceptual por Condicionamientos Sucesivos (CCCS), ver Capítulo §11, permite construir una base de conocimiento con tantos conceptos como clases se quieran interpretar, siempre y cuando procedan de un *clustering* jerárquico, ver capítulo §5, y con la propiedad de que la aridad de cada concepto será la profundidad de la clase en el dendrograma.

Para elaborar la propuesta metodológica final, se ha trabajado en distintos aspectos que se presentan en este capítulo, en primer lugar se hace una tipificación de los sistemas de reglas inducidos por el BbIR, presentado previamente en el Capítulo §5, un análisis las propiedades que de estos sistemas se derivan tanto para el caso general de ξ clases como para el caso particular de particiones binarias, finalmente la revisión del Boxplot based induction rules (BbIR).

7.1 Tipificación de los sistemas de reglas inducidos por el BbIR

Partiendo del Boxplot-based induction rules (BbIR)(Vázquez and Gibert 2001) presentado en la sección §5.10 de esta tesis, se ha realizado un análisis sobre los resultados en distintos casos reales y se ha visto que los sistemas de reglas se pueden estructurar en formas diferentes según los intereses del análisis y del tipo de reglas que contienen.

Así se propone considerar los siguiente conjuntos de reglas:

Dada \mathcal{P}_ξ una partición en ξ clases de \mathcal{I} y X_k una variable numérica o categórica y dado $\mathcal{D}^k = \{I_1^k, \dots, I_{n_k}^k\}$, el conjunto de valores que toma X_k (I_s^k será una modalidad si X_k es categórica o un intervalo si X_k es numérica (en cuyo caso $D^k = \bigcup_{s=1}^{\xi} I_s^k$), eventualmente obtenido por una discretización basada en boxplots) y siendo $p_{sc} = P(C | I^k = I_s^k) = \frac{\text{card}\{i : i \in C \wedge x_{ik} \in I_s^k\}}{\text{card}\{i : x_{ik} \in I_s^k\}}$ cuando I_s^k es un intervalo y $p_{sc} = P(C | I^k = I_s^k) = \frac{\text{card}\{i : i \in C \wedge x_{ik} = I_s^k\}}{\text{card}\{i : x_{ik} = I_s^k\}}$ cuando I_s^k es una modalidad.

7.1.1 Sistema de reglas completo $\mathcal{R}(X_k, \mathcal{P}_\xi)$

Basándonos en la idea original de (Vázquez Torres 2002), conservamos la definición e incorporamos una notación más conveniente para esta tesis:

$$\begin{aligned} \mathcal{R}(X_k, \mathcal{P}_\xi) = \{ & r_{s,c}^k : x_{ik} \in I_s^k \xrightarrow{p_{sc}} i \in C \text{ con } p_{sc} > 0, \quad p_{sc} = P(C | I^k = I_s^k) \\ & s = \{1, \dots, (2\xi - 1)\}, \quad c \in \mathcal{P}_\xi, \quad k = \{1, \dots, K\} \} \end{aligned}$$

Finalmente, se obtiene un sistema global, a partir del cual, para cierto valor del atributo X_k se da un mayor o menor grado de pertenencia a cada clase de cierta partición de referencia \mathcal{P}_ξ .

Los sistemas de reglas, pueden contener hasta 3 tipos de reglas que serán relevantes en la determinación de la calidad del Sistema, estos tipo son:

1. Reglas no efectivas:

Diremos que una regla $r_{s,c}^k : x_{ik} \in I_s^k \xrightarrow{p_{sc}} i \in C$ es no efectiva si $p_{sc} = 0$.

Las reglas con $p_{sc} = 0$ son totalmente prescindibles y solo aumentan innecesariamente la cardinalidad del sistema de reglas completo.

2. Reglas efectivas:

Diremos que una regla $r_{s,c}^k : x_{ik} \in I_s^k \xrightarrow{p_{sc}} i \in C$ es efectiva si $p_{sc} > 0$.

3. Reglas seguras:

Diremos que una regla $x_{ik} \in I_s^k \xrightarrow{p_{sc}} i \in C$ es segura si $p_{sc} = 1$.

Son un caso particular de las efectivas.

7.1.2 Sistema de reglas completo global, $\mathcal{R}(\mathcal{P}_\xi)$

Es es sistema de reglas que une a todos y cada uno de los sistemas de reglas completo generados para cada variable X_K . Definimos el Sistema de reglas completo global como:

$$\mathcal{R}(\mathcal{P}_\xi) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_\xi)$$

7.1.3 Sistema de reglas reducido, $\mathcal{R}^*(X_k, \mathcal{P}_\xi)$

Es el sistema que contiene una y sólo una regla de antecedente I_s^k y es la de probabilidad máxima de entre las de $\mathcal{R}(X_k, \mathcal{P}_\xi)$. En los trabajos anteriores no se llega a expresar formalmente y ahora lo definimos como:

$$\begin{aligned} \mathcal{R}^*(X_k, \mathcal{P}_\xi) = \{ & r_{s,c}^k : x_{ik} \in I_s^k \xrightarrow{p_{sc}} i \in C \text{ con } p_{sc} = \max_{\forall C' \in \mathcal{P}_\xi} \{p_{sc'}\} \\ & s = \{1, \dots, (2\xi - 1)\}, \quad k = \{1, \dots, K\} \} \end{aligned}$$

Propiedades:

- $\mathcal{R}^*(X_k, \mathcal{P}_\xi) \subseteq \mathcal{R}(X_k, \mathcal{P}_\xi)$.
- Puede contener reglas de cualquiera de los tipos definidos anteriormente.

7.1.4 Sistema de reglas reducido global, $\mathcal{R}^*(\mathcal{P}_\xi)$

Es es sistema de reglas que une a todos y cada uno de los sistemas de reglas reducidos generados para cada variable X_K . Definimos el Sistema de reglas completo global como:

$$\mathcal{R}^*(\mathcal{P}_\xi) = \bigcup_{k=1}^K \mathcal{R}^*(\mathcal{P}_\xi)$$

7.1.5 Sistemas de reglas de nivel η

Se podrán construir subsistemas de reglas a partir del sistema de reglas completo global condicionando qué el valor puede tener p_{sc} para que la regla $r_{s,c}^k$ pertenezca al sistema.

Para todas las clases $C \in \mathcal{P}_\xi$:

$$\mathcal{R}^\xi(X_k, \eta) = \{r_{s,c}^k \in \mathcal{R}(\mathcal{P}_\xi) \text{ tq } r_{s,c}^k : x_{ik} \in I_s^{k,\xi} \xrightarrow{p_{sc}} i \in C \text{ y } p_{sc} \geq \eta, \forall C \in \mathcal{P}_\xi\}$$

7.1.6 Sistema de reglas efectivas $\mathcal{R}e(X_k, \mathcal{P}_\xi)$

Es el caso particular de sistema de reglas de nivel $\mathcal{R}^\xi(\eta)$ para $\eta > 0$.

El sistema reglas efectivas es el que contiene solamente reglas efectivas de entre las del sistema de reglas completo $\mathcal{R}(X_k, \mathcal{P}_\xi)$, si existen, siendo $p_{sc} > 0$ la probabilidad asociada a una regla $r_{s,c}^k$,

$$\mathcal{R}e(X_k, \mathcal{P}_\xi) = \{r_{s,c}^k \in \mathcal{R}(X_k, \mathcal{P}_\xi) \text{ tq } p_{sc} > 0, s = \{1, \dots, (2\xi - 1)\}, c \in \mathcal{P}_\xi, k = \{1, \dots, K\}\}$$

Propiedades:

- $\mathcal{R}e(X_k, \mathcal{P}_\xi) \subseteq \mathcal{R}(X_k, \mathcal{P}_\xi)$.
- $\mathcal{R}e(X_k, \mathcal{P}_\xi) \subseteq \mathcal{R}^*(X_k, \mathcal{P}_\xi)$.
- No contiene reglas no efectivas.

7.1.7 Sistema de reglas efectivas global, $\mathcal{R}e(\mathcal{P}_\xi)$

Es el sistema de reglas que une a todos y cada uno de los sistemas de reglas efectivas generados para cada variable X_K . Definimos el Sistema de reglas efectivas global como:

$$\mathcal{R}e(\mathcal{P}_\xi) = \bigcup_{k=1}^K \mathcal{R}e(X_k, \mathcal{P}_\xi)$$

7.1.8 Sistema de reglas efectivas reducido $\mathcal{R}e^*(X_k, \mathcal{P}_\xi)$

Es un subconjunto de $\mathcal{R}e(X_k, \mathcal{P}_\xi)$.

El sistema reglas efectivas reducido equivale a un sistema de reglas reducido en el que se han eliminado todas las reglas no efectivas que pudiera contener. Siendo $p_{sc} > 0$ la probabilidad asociada a una regla $r_{s,c}^k$.

$$\mathcal{R}e^*(X_k, \mathcal{P}_\xi) = \{r_{s,c}^k \in \mathcal{R}e(X_k, \mathcal{P}_\xi) \text{ } p_{sc} = \max_{\forall C' \in \mathcal{P}_\xi} \{p_{sc'}\} \text{ y } p_{sc} > 0, s = \{1, \dots, (2\xi - 1)\}, k = \{1, \dots, K\}\}$$

Propiedades:

- $\mathcal{R}e^*(X_k, \mathcal{P}_\xi) \subseteq \mathcal{R}^*(X_k, \mathcal{P}_\xi) \subseteq \mathcal{R}(X_k, \mathcal{P}_\xi)$.
- $\mathcal{R}e^*(X_k, \mathcal{P}_\xi) = \mathcal{R}^*(X_k, \mathcal{P}_\xi) \setminus \{r_{s,c}^k \in \mathcal{R}^*(X_k, \mathcal{P}_\xi) \text{ tq } p_{sc} = 0\}$.
- $\mathcal{R}e^*(X_k, \mathcal{P}_\xi) \subseteq \mathcal{R}e(X_k, \mathcal{P}_\xi)$.
- No contiene reglas no efectivas.
- Contiene una o ninguna regla de antecedente I_s^k pero nunca más de 1.

7.1.9 Sistema de reglas efectivas reducido global, $\mathcal{R}e^*(\mathcal{P}_\xi)$

Es el sistema de reglas que une a todos y cada uno de los sistemas de reglas efectivas reducido generados para cada variable X_K . Definimos el Sistema de reglas efectivas reducido global como:

$$\mathcal{R}e^*(\mathcal{P}_\xi) = \bigcup_{k=1}^K \mathcal{R}e^*(X_k, \mathcal{P}_\xi)$$

7.1.10 Sistema de reglas Seguras $\mathcal{S}(X_k, \mathcal{P}_\xi)$

Es el caso particular de sistema de reglas de nivel $\mathcal{R}^\xi(\eta)$ para $\eta = 1$ y está incluido en el sistema de reglas efectivas, en el sistema de reglas completo y en el sistema de reglas reducido.

El sistema reglas seguras es el que contiene solamente las reglas seguras ($p_{sc} = 1$) de entre las del sistema de reglas reducido $\mathcal{R}^*(X_k, \mathcal{P}_\xi)$, si existen, siendo p_{sc} la probabilidad asociada a una regla $r_{s,c}^k$,

$$\mathcal{S}(X_k, \mathcal{P}_\xi) = \{r_{s,c}^k \in \mathcal{R}^*(X_k, \mathcal{P}_\xi) \text{ tq } p_{sc} = 1, s = \{1, \dots, (2\xi - 1)\}, c \in \mathcal{P}_\xi, k = \{1, \dots, K\}\}$$

Propiedades:

- Es un caso particular de $\mathcal{R}^\xi(\eta)$ para $\eta = 1$.
- $\mathcal{S}(X_k, \mathcal{P}_\xi) \subseteq \mathcal{R}(X_k, \mathcal{P}_\xi)$.
- $\mathcal{S}(X_k, \mathcal{P}_\xi) \subseteq \mathcal{R}^*(X_k, \mathcal{P}_\xi)$.
- $\mathcal{S}(X_k, \mathcal{P}_\xi) \subseteq \mathcal{R}e(X_k, \mathcal{P}_\xi)$.
- Sólo contiene reglas seguras.
- No contiene reglas no efectivas.

7.1.11 Sistema de reglas seguras global, $\mathcal{S}(\mathcal{P}_\xi)$

Es es sistema de reglas que une a todos y cada uno de los sistemas de reglas seguras generados para cada variable X_K . Definimos el Sistema de reglas completo global como:

$$\mathcal{S}(\mathcal{P}_\xi) = \bigcup_{k=1}^K \mathcal{S}(X_k, \mathcal{P}_\xi)$$

Propiedades:

- $\mathcal{S}(\mathcal{P}_\xi) = \mathcal{R}^\xi(1)$.

7.2 Tipos de valores

El hecho de que una variable X_k genere sistemas de reglas con más o menos reglas seguras está relacionado con el comportamiento de los valores que toma dicha variable.

Sea X_k una variable categórica o la categorización en intervalos de una variable numérica, hemos llamado \mathcal{D}^k al conjunto de modalidades de X_k . Eventualmente \mathcal{D}^k puede ser el conjunto de intervalos en que se ha categorizado una variable originalmente numérica.

Los valores de \mathcal{D}^k , ya sean modalidades o intervalos pueden ser de 4 tipos básicos, ya definidos en el capítulo §5 y que se muestran en la tabla 7.1.

	Tipo de Valor	Cobertura de la Clase	
		Cubre todo C	Cubre sólo parte C
Interacción con otras clases	Exclusivo	<i>Caracterizador total</i>	<i>Caracterizador parcial</i>
	No exclusivo	<i>Caracterizador no propio</i>	<i>Genérico</i>

Tabla 7.1: Valores caracterizadores.

En esta tesis se aporta una caracterización alternativa en función de p_{sc} , lo que permite identificar los tipos de valores a partir de una sistema de reglas. Así

1. Un valor *totalmente caracterizador* de C es: I_s^k tq $p_{sc} = 1$, $p_{s'c} = 0$, $\forall s' \neq s$.
Si I_s^k es el valor totalmente caracterizador para C , entonces I_s^k producirá una regla segura para C .
2. Un valor *parcialmente caracterizador* de C es: I_s^k tq $p_{sc} = 1$, $p_{s'c} > 0$, $\forall s' \neq s$.
Los valores total o parcialmente caracterizadores siempre generan reglas seguras.
Si I_s^k es el valor parcialmente caracterizador para C , entonces I_s^k producirá una regla segura para C .
3. Un valor *caracterizador no propio* de C es: I_s^k tq $p_{sc} \in (0, 1)$, $p_{s'c} = 0$, $\forall s' \neq s$.
4. Un *valor genérico* de la clase C es: I_s^k tq $p_{sc} \in (0, 1)$ y $\exists s'$ tal que $p_{s'c} > 0$, $s' \neq s$ y $\exists c'$ tal que $p_{sc'} > 0$, $c' \neq c$. Estos valores se pueden interpretar como el subconjunto de individuos i de la clase C que comparten su valor I_s^k tanto con las demás clases, existiendo a su vez en la misma clase C algunos otros elementos que pertenecen a otros intervalos.

En (Pérez-Bonilla and Gibert 2006) se llaman intervalos comunes para el caso de variables numéricas discretizadas por el BbD o por algún otro método.

Los valores no vacíos *generan siempre reglas efectivas*, es decir, $n_{I_s^k} \neq 0 \iff p_{sc} > 0$. Las reglas efectivas son de gran importancia para la posterior generación de las interpretaciones de las clases.

Nunca generan reglas no efectivas.

5. Además se ha observado que a efectos de los sistemas de reglas es relevante considerar un nuevo tipo de valor que es el *valor vacío*.

Sea $n_{I_s^k}$ el $card\{I_s^k\}$ o el número de elementos de \mathcal{I} que para X_k toman valores en I_s^k :

- (a) $n_{I_s^k} = card\{I_s^k\} = card\{i \in \mathcal{I} : x_{ik} = I_s^k\}$, si X_k es categórica.
- (b) $n_{I_s^k} = card\{I_s^k\} = card\{i \in \mathcal{I} : x_{ik} \in I_s^k\}$, si X_k es numérica.

Diremos que un intervalo o modalidad es vacía si:

$$n_{I_s^k} = 0$$

Según X_k sea numérica o no:

- (a) $n_{I_s^k} = 0 \iff I_s^k = \emptyset$, si X_k es numérica.
- (b) $n_{I_s^k} = 0 \iff I_s^k \neq x_{ik}$, si X_k es categórica.

Los valores vacíos *generan siempre reglas no efectivas*, es decir $n_{I_s^k} = 0 \iff p_{sc} = 0$, es decir producen siempre reglas de probabilidad nula de las que se puede prescindir. El número de valores vacíos de X_k determinará la calidad de un sistema de reglas.

	Tipo de Valor	Cobertura de la Clase	
		Cubre todo C	Cubre sólo parte C
Interacción con otras clases	Exclusivo	$I_s^k \text{ tq } p_{sc} = 1 \wedge p_{s'c} = 0, \forall s' \neq s$	$I_s^k \text{ tq } p_{sc} = 1 \wedge p_{s'c} > 0, \forall s' \neq s$
	No exclusivo	$I_s^k \text{ tq } p_{sc} \in (0, 1) \wedge p_{s'c} = 0, \forall s' \neq s$	$I_s^k \text{ tq } p_{sc} \in (0, 1) \wedge p_{s'c} > 0, \forall s' \neq s$

Tabla 7.2: Relación entre valores caracterizadores y p_{sc} .

La configuración de un sistema de reglas con más o menos valores de un cierto tipo determinará criterios de calidad para evaluar los sistemas de reglas.

Por otro lado, sobre los tipos de intervalos construidos a partir la discretización de X_K utilizando el *BbD*, hemos visto en función de cómo es su intersección con las distintas clases, que existen tres tipos de intervalos diferentes (Gibert and Pérez-Bonilla 2006a), (Pérez-Bonilla and Gibert 2006) que resulta relevante considerar para la generación de interpretaciones:

1. El *intervalo vacío* (V): Los intervalos vacíos son siempre de la forma $(z_s^k, z_s^k]$ o (z_s^k, z_s^k) o $[z_s^k, z_s^k)$. Dan lugar a valores vacíos y por tanto producen ξ reglas no efectivas.
2. El *intervalo propio* (P): Constituye un *valor propio* de la clase, es decir se da de forma exclusiva en una clase y puede ser caracterizador total o parcial de la misma según si lo cubre total o parcialmente. Estos valores, cuando ocurren, identifican una clase con toda seguridad, por lo que, actúan como *valores caracterizadores* de la clase C .

Todo intervalo propio cumple que $p_{sc} = 1$. En efecto:

- si se trata de totalmente caracterizador $I_s^k \text{ tq } p_{sc} = 1 \wedge p_{s'c} = 0, \forall s' \neq s$.
- si se trata de parcialmente caracterizador, $I_s^k \text{ tq } p_{sc} = 1 \wedge p_{s'c} > 0, \forall s' \neq s$.

Entonces un intervalo propio cumple que:

$$(p_{sc} = 1 \wedge p_{s'c} = 0, \forall s' \neq s) \vee (p_{sc} = 1 \wedge p_{s'c} > 0, \forall s' \neq s)$$

Aplicando las leyes del cálculo proposicional;

$$\begin{aligned} (p_{sc} = 1 \wedge p_{s'c} = 0, \forall s' \neq s) \vee (p_{sc} = 1 \wedge p_{s'c} > 0, \forall s' \neq s) &= \\ (p_{sc} = 1) \wedge (p_{s'c} = 0 \vee p_{s'c} > 0, \forall s' \neq s) &= \\ (p_{sc} = 1) \wedge (p_{s'c} \geq 0 \forall s' \neq s) &= \\ \text{por ser } p_{sc} \in [0, 1] \text{ una proporción} \\ (p_{sc} = 1 \wedge \text{True}) &= \\ (p_{sc} = 1) \end{aligned}$$

Éstos dan siempre lugar a una única regla segura ($p_{sc} = 1$) y a $\xi - 1$ reglas no efectivas.

3. El *intervalo común* (G): Constituye un valor que cumple siempre que $p_{sc} \in (0, 1)$. Entonces un intervalo común puedo comportarse como:

- *valor caracterizador no propio*, es decir se da en la clase y la cubre entera pero no es exclusivo de ella y cumple que $I_s^k \text{ tq } p_{sc} \in (0, 1) \wedge p_{s'c} = 0, \forall s' \neq s$ o
- *valor genérico* de la clase que cumple que $I_s^k \text{ tq } p_{sc} \in (0, 1) \wedge p_{s'c} > 0, \forall s' \neq s$.

Entonces un intervalo propio cumple que:

$$\begin{aligned}
& (p_{sc} \in (0, 1) \wedge p_{s'c} = 0, \forall s' \neq s) \vee (p_{sc} \in (0, 1) \wedge p_{s'c} > 0, \forall s' \neq s) = \\
& (p_{sc} \in (0, 1) \wedge (p_{s'c} = 0 \vee p_{s'c} > 0) \forall s' \neq s) = \\
& (p_{sc} \in (0, 1) \wedge (p_{s'c} \geq 0) \forall s' \neq s) = \\
& \text{por ser } p_{sc} \in [0, 1] \text{ una proporción} \\
& (p_{sc} \in (0, 1) \wedge \text{True}) = \\
& (p_{sc} \in (0, 1))
\end{aligned}$$

En el caso general da lugar a ξ reglas de probabilidades complementarias que suman 1.

El intervalo común es un caso general, pero cumple siempre que no genera ninguna regla segura, por lo menos 2 reglas son efectivas y en consecuencia genera entre 0 y $\xi - 2$ reglas no efectivas según el número de clases que solapan en ese intervalo.

Para particiones binarias da lugar siempre a dos reglas de probabilidades complementarias ($p_{sc} = 1 - p_{sc'}$).

Si \mathcal{D}^k es el conjunto de valores que toma X_k

$$\text{En general se cumple que } \text{card}\mathcal{D}^k = n_p^k + n_g^k + n_v^k$$

Donde,

n_p^k , es el número de intervalos propios.

n_g^k , es el número de intervalos comunes y

n_v^k , es el número de intervalos vacíos.

$$\text{En el caso particular del BbD } \mathcal{D}^k = 2\xi - 1, \text{ y en consecuencia } n_p^k + n_g^k + n_v^k = 2\xi - 1.$$

7.3 Cardinales y propiedades de los Sistemas de reglas

La valoración de la calidad de los sistemas de reglas inducidos a partir de variables categóricas o numéricas discretizadas (ya sea por *BbIR*(ver §5.10) u otros métodos) se puede estructurar alrededor de ciertos parámetros de interés:

- Cardinalidad de un sistema de reglas ($n_{\mathcal{R}}$):

La cardinalidad de un sistema de reglas cualquiera es el número de reglas que contiene.

Sea,

n_0 , el número de reglas no efectivas en un Sistema de reglas.

n_e , el número de reglas efectivas en un Sistema de reglas (cardinalidad de $\mathcal{R}e$).

$n_{\mathcal{S}}$, el número de reglas seguras ($p_{sc} = 1$) en un Sistema de reglas(cardinalidad de \mathcal{S}).

Podemos enunciar las siguientes propiedades que sirven tanto para sistemas globales o locales a una variable.

- Si n^k es la cardinalidad de un sistema de reglas de la variable X_k entonces, $n = \sum_{\forall k} n^k$ es la cardinalidad del correspondiente sistema global, ya que los conjuntos de reglas que genera cada variable son disjuntos.
- $n_{\mathcal{R}}^k = \xi \text{card } \mathcal{D}^k$, si ξ es el número de clases a interpretar.
- $n_{\mathcal{R}*}^k = \text{card } \mathcal{D}^k = n_p^k + n_g^k + n_v^k$.
- $n_{\mathcal{R}} = n_e + n_0$ y es constante dado ξ .

- $n_e = n_{\mathcal{R}} - n_0$, que el número de reglas *no efectivas* aumente en un sistema de reglas es consecuencia directa de que aumente el número de valores vacíos en X_k e implica, si X_k es la discretización de una variable originalmente continua, que se tiene una discretización más compacta.
- El cardinal del sistema de reglas efectivas reducido es $n_{\mathcal{R}e^*} = n_p + n_g$.
- Toda regla segura es también efectiva. Así $n_{\mathcal{S}} \leq n_e$, o también $n_e - n_{\mathcal{S}} \geq 0$.
- Si n_e disminuye es consecuencia directa de que aumente el número de valores vacíos de X_k e implica mayor compactación.
- Si $\frac{n_e}{n_{\mathcal{R}}} \rightarrow 1$, el sistema de reglas completo es mejor porque todas las reglas que se generan son efectivas ($n_0 \rightarrow 0$).
- Si $\frac{n_0}{n_{\mathcal{R}}} \rightarrow 0$, el sistema de reglas completo es mejor porque no hay ninguna que regla que se pueda despreciar, con lo cual la cardinalidad del sistema de reglas completo y la del sistema de reglas efectivas se aproximan ($n_e \rightarrow n_{\mathcal{R}}$).
- Si $\frac{n_{\mathcal{S}}}{n_{\mathcal{R}}} \rightarrow 1$, esta es la mejor situación que se puede dar, ya que las reglas seguras identificarían variables caracterizadoras y son las de menor incertezza.
- Si $n_e \rightarrow n_{\mathcal{S}}$, como en el caso anterior, idealmente el número de reglas efectivas de un sistema de reglas completo debiera tender al número de reglas seguras lo que facilitaría encontrar variables caracterizadoras.
- Si $n_{\mathcal{S}} \rightarrow \xi$, cuando esto ocurre podríamos decir que hay al menos una regla segura asociada a cada clase de la partición \mathcal{P}_ξ y podemos reconocer todas las clases sin incertezza.
- $n_e \leq \xi n_g$. El número de reglas efectivas es menor o igual que el número de intervalos comunes por el número de clases de la partición \mathcal{P}_ξ .
- $n_0 \geq \xi n_v$. Un intervalo vacío genera siempre ξ reglas no efectivas, por lo tanto el número total de reglas no efectivas de un sistema es como mínimo ξn_v .
- La mejor situación de todas es $n_e \rightarrow n_{\mathcal{S}}$ y $n_{\mathcal{S}} \rightarrow n_{\mathcal{R}}$.
- Así, entre dos sistemas de reglas será mejor el de mayor número de reglas seguras. Si coinciden en $n_{\mathcal{S}}$ será mejor el de más reglas no efectivas.

7.3.1 Relación entre los tipos de valores de una variable y los tipos de reglas que genera

El número de reglas efectivas o seguras de $\mathcal{R}(X_k, \mathcal{P}_\xi)$ depende de cómo son los valores de X_k respecto de \mathcal{P}_ξ .

Ya se ha visto que cada variable X_k genera

- n_p^k valores propios.
- n_g^k valores comunes.
- n_v^k valores vacíos.

También se vio que:

- Si s es un *valor propio* de X_k genera:
 - 1 regla segura.
 - 0 reglas efectivas y no seguras.
 - $(\xi - 1)$ reglas no efectivas.
- Si s es un *valor común* de X_k genera:
 - 0 reglas seguras.
 - E_{ks} reglas efectivas y no seguras, $2 \leq E_{ks} \leq \xi$.
 - $\xi - E_{ks}$ reglas no efectivas.
- Si s es un *valor vacío* de X_k genera:
 - 0 reglas seguras.
 - 0 reglas efectivas.
 - ξ reglas no efectivas.

Así, considerando todos los valores de X_k conjuntamente, la variable X_k genera:

- Para todos sus *valores propios*:
 - n_p reglas seguras y efectivas.
 - 0 reglas efectivas y no seguras.
 - $n_p(\xi - 1)$ no efectivas.
- Para todos sus *valores comunes*:
 - 0 reglas seguras.
 - $\sum_{\forall s \text{ valor comun}} (E_{ks})$ reglas efectivas y no seguras, $2 \leq E_{ks} \leq \xi$.
 - $\sum_{\forall s \text{ valor comun}} (\xi - E_{ks})$ reglas no efectivas.
- Para todos sus *valores vacíos*:
 - 0 reglas seguras.
 - 0 efectivas.
 - $n_v\xi$ reglas no efectivas.

En total la variable X_k da lugar al siguiente número de reglas:

- Número de reglas seguras que genera X_k en $\mathcal{R}(X_k, \mathcal{P}_\xi)$:

$$n_s^k = n_p^k + 0 + 0 = n_p^k$$

- Número de reglas efectivas que genera X_k en $\mathcal{R}(X_k, \mathcal{P}_\xi)$:

$$n_e^k = n_p^k + \sum_{\forall s \text{ valor comun}} (E_{ks}) + 0 = n_p^k + \sum_{\forall s \text{ valor comun}} (E_{ks})$$

- Número de reglas no efectivas que genera X_k en $\mathcal{R}(X_k, \mathcal{P}_\xi)$:

$$\begin{aligned}
n_0^k &= n_p^k(\xi - 1) + \sum_{\forall s \text{ valor comun}} (\xi - E_{ks}) + n_v^k \xi \\
&= n_p^k(\xi - 1) + \xi n_g^k - \sum_{\forall s \text{ valor comun}} (E_{ks}) + n_v^k \xi \\
&= \xi(n_p^k + n_g^k + n_v^k) - n_p^k - \sum_{\forall s \text{ valor comun}} (E_{ks}) \\
&= \xi \text{ card} \mathcal{D}^k - n_p^k - \sum_{\forall s \text{ valor comun}} (E_{ks}) \quad 2 \leq E_{ks} \leq \xi
\end{aligned}$$

Teniendo en cuenta que los sistemas de reglas de cada variable son disjuntos:

- El número total de *reglas seguras* de $\mathcal{R}(\mathcal{P}_\xi)$:

$$n_S = n_p = \sum_{k=1}^K n_p^k$$

El número de reglas seguras es siempre igual al número de intervalos propios del sistema porque cada valor propio genera una única regla segura y no hay ningún otro tipo de valor que pueda generar reglas seguras.

- El número total de *reglas efectivas* de $\mathcal{R}(\mathcal{P}_\xi)$:

$$n_e = \sum_{\forall k} n_p^k + \sum_{\forall k} \sum_{\forall s \text{ valor comun}} (E_{ks}) \quad 2 \leq E_{ks} \leq \xi$$

- El número total de *reglas no efectivas* de $\mathcal{R}(\mathcal{P}_\xi)$:

$$n_0 = \sum_{\forall k} \xi \text{ card} \mathcal{D}^k - \sum_{\forall k} n_p^k - \sum_{\forall k} \sum_{\forall s \text{ valor comun}} (E_{sk}) \quad 2 \leq E_{ks} \leq \xi$$

De ahí se puede establecer un criterio final de calidad ligado a la estructura de los valores de X_k . La mejor situación es que $n_p \rightarrow \xi$ ya que $n_p = n_S$

7.3.2 Variables procedentes de la discretización por el BbD de una variable continua

Cuando la variable X_k es producto de la discretización realizada por el Boxplot based Discretization (BbD) de una variable continua, la cardinalidad siempre es:

$$\text{card } \mathcal{D}^k = 2\xi - 1$$

De acuerdo a lo anterior se cumplen las siguientes propiedades:

- Puesto que el sistema de reglas completo genera ξ reglas por modalidad la cardinalidad del sistema de reglas completo es:

$$n_{\mathcal{R}}^k = \xi \text{card} \mathcal{D}^k = \xi(2\xi - 1) = 2\xi^2 - \xi$$

y entonces,

$$n_{\mathcal{R}} = \sum_{\forall k} \xi \text{card} \mathcal{D}^k = K\xi(2\xi - 1) = K(2\xi^2 - \xi)$$

- Puesto que el sistema de reglas reducido genera una regla por modalidad, la cardinalidad es:

$$n_{\mathcal{R}^*}^k = 2\xi - 1$$

y entonces,

$$n_{\mathcal{R}^*} = \sum_{\forall k} 2\xi - 1 = K(2\xi - 1)$$

- Número de reglas no efectivas que genera X_k en $\mathcal{R}(X_k, \mathcal{P}_\xi)$

$$n_0^k = 2\xi^2 - \xi - n_p^k - \sum_{\forall s \text{ valor comun}} (E_{sk}) \quad 2 \leq E_{ks} \leq \xi$$

7.4 Particiones binarias

Según se vio en (Gibert and Pérez-Bonilla 2006a) y (Pérez-Bonilla and Gibert 2006), para el caso que nos ocupa en el que se tratarán sólo particiones binarias $\xi = 2$, $\mathcal{P}_2 = \{C_i, C_j\}$.

Partiendo de \mathcal{P}_2 y utilizando la propuesta original del *Boxplot based Discretization*, presentada en la sección §5, el sistema de intervalos inducidos por \mathcal{P}_2 sobre la variable X_k es: $\mathcal{D}^k = \{I_1^k, I_2^k, I_3^k\}$ y tendrá siempre 3 elementos tales que $D^k = I_1^k \cup I_2^k \cup I_3^k$, sin perjuicio de que algún intervalo pueda ser vacío.

En un sistema de reglas completo a lo sumo tendremos 2 reglas de igual antecedente una de consecuente C_i y otra de consecuente C_j , que además presentarán siempre probabilidades complementarias siendo;

$$p_{si} = \frac{\text{card}\{i : i \in C_i \wedge x_{ik} \in I_s^k\}}{\text{card}\{i : x_{ik} \in I_s^k\}} \quad y \quad p_{sj} = \frac{\text{card}\{i : i \in C_j \wedge x_{ik} \in I_s^k\}}{\text{card}\{i : x_{ik} \in I_s^k\}}$$

y se cumplirá siempre que $p_{si} + p_{sj} = 1$ por lo que $p_{sj} = 1 - p_{si}$. Por ser \mathcal{P}_2 una partición $C_i \cup C_j = \mathcal{I}$ y $C_i \cap C_j = \emptyset$.

Así $\text{Card}\{C_i\} + \text{Card}\{C_j\} = n$ y $p_{si} > p_{sj} \Leftrightarrow \text{Card}\{C_i\} > \text{Card}\{C_j\}$.

Por ello;

$$\max\{p_{si}, p_{sj}\} = \max\{p_{si}, 1 - p_{si}\} = \begin{cases} p_{si} & \text{si } p_{si} > 0.5 \\ p_{sj} & \text{si } \text{no} \end{cases}$$

Lo que significa que $p_{si} > 0.5 \Leftrightarrow \text{Card}C_i > \text{Card}C_j$.

Para el caso particular de $\xi = 2$, $\mathcal{P}_2 = \{C_i, C_j\}$ el sistema de reglas reducido $\mathcal{R}^*(X_k, \mathcal{P}_2)$ será:

$$\mathcal{R}^*(X_k, \mathcal{P}_2) = \{ r_{s,c}^k : x_{ik} \in I_s^k \xrightarrow{psc} i \in C \text{ con } r_{s,c}^k \text{ tq } p_{sc} = \max\{p_{si}, p_{sj}\} \\ s = \{1, 2, 3\}, c \in \mathcal{P}_2, k = \{1, \dots, K\} \}$$

Además, puesto que un sistema de reglas reducido estaría formado por las reglas que presenten una mayor probabilidad, cuando los tamaños de las clases son muy desequilibrados, la probabilidades que se generan también lo son. Por eso el sistema de reglas reducido asignará los individuos a la clase más grande.

De todo lo anterior se deduce que en particiones binarias siempre se cumple que:

1. La cardinalidad del sistema de reglas completo es constante y $n_{\mathcal{R}}^k = 2(n_p^k + n_g^k + n_v^k) = 6$.
2. La cardinalidad del sistema de reglas reducido es constante y $n_{\mathcal{R}^*}^k = n_p^k + n_g^k + n_v^k = 3$.
3. Como la cardinalidad del sistema de reglas completo es constante, la cardinalidad del sistema de reglas completo global dependerá del número de variables, así $n_{\mathcal{R}} = 6K$.
4. En cuanto a la cardinalidad del sistema de reglas reducidos tenemos que $n_{\mathcal{R}^*}^k = 3$ y para el reducido global $n_{\mathcal{R}^*} = 3K$.
5. En el caso de la cardinalidad de los sistemas de reglas efectivas se cumple que $n_e^k = 6 - n_0^k$ y $n_e^k = n_p^k + 2n_g^k \leq 4$. Para el sistema de reglas efectivas global se cumplirá que $n_e = 6K - n_0 = n_p + 2n_g \leq 4$.
6. Finalmente la cardinalidad del sistema de reglas seguras cumplirá que $n_{\mathcal{S}}^k = n_p^k \leq 3$ y por tanto $n_{\mathcal{S}} = n_p \leq 3K$.
7. El número de *reglas no efectivas* (de probabilidad nula) que contiene $\mathcal{R}(X_k, \mathcal{P}_2)$, se cumple siempre que $n_0^k = 6 - n_p^k - \sum_{s=1}^2 (E_{sk})$.
8. Considerando todos los valores de X_k en una partición binaria:
 - (a) $0 \leq n_p^k \leq 3$
 - (b) $0 \leq n_g^k \leq 3$
 - (c) $0 \leq n_v^k \leq 2$
 - (d) $n_p^k + n_g^k + n_v^k = 3$
 - (e) Si $n_v^k = 0$, entonces;
 - $1 \leq n_p^k \leq 3$
 - $1 \leq n_g^k \leq 2$
 - (f) Si $n_p^k = 0$, entonces;
 - $0 \leq n_v^k \leq 1$
 - $0 \leq n_g^k \leq 2$
 - (g) Si $n_g^k = 0$, entonces;
 - $2 \leq n_p^k \leq 3$
 - $1 \leq n_v^k \leq 1$

Como ya hemos definido en el Capítulo §5 a partir de la discretización realizada con el *Bbd* se puede realizar una *inducción de reglas basada en box-plots* mediante el método *Boxplot-based Induction Rules*. En (Gibert and Pérez-Bonilla 2006a) se vio que según como se definen los límites de los intervalos I_s^k , lo que es de vital importancia para la generación de interpretaciones automáticas, el sistema de reglas inducido a partir de él producirá más o menos reglas seguras.

7.5 Revisión del Boxplot based Discretization (BbD)

En esta tesis se presenta una propuesta metodológica que se basa en el análisis iterativo de particiones binarias. Así, se parte siempre de una partición de $\mathcal{I} = \{i_1, \dots, i_n\}$, o de uno de sus subconjuntos, en 2 clases: $\mathcal{P}_2 = \{C_i, C_j\}$. El análisis de particiones binarias generará sistemas que presentan algunas particularidades. En (Gibert and Pérez-Bonilla 2006a) y (Pérez-Bonilla and Gibert 2006) se observó que la propuesta original del *Boxplot based Discretization* al aplicarla a particiones binarias (Vázquez and Gibert 2001) presentaba las siguientes anomalías:

1. Existe un único patrón establecido de manera uniforme para la generación de los 3 intervalos en particiones binarias el que se materializa de siguiente forma:

$$I_1^{k,2} = [z_1^k, z_2^k], \quad I_2^{k,2} = (z_2^k, z_3^k], \quad I_3^{k,2} = (z_3^k, z_4^k]$$

2. Hay situaciones en que el intervalo del centro contiene un sólo punto aislado lo que genera una regla segura para un intervalo que en realidad es vacío excepto por un único punto extremo. Siendo ésta una división artificial del recorrido de esa clase (ver Figura 7.1).

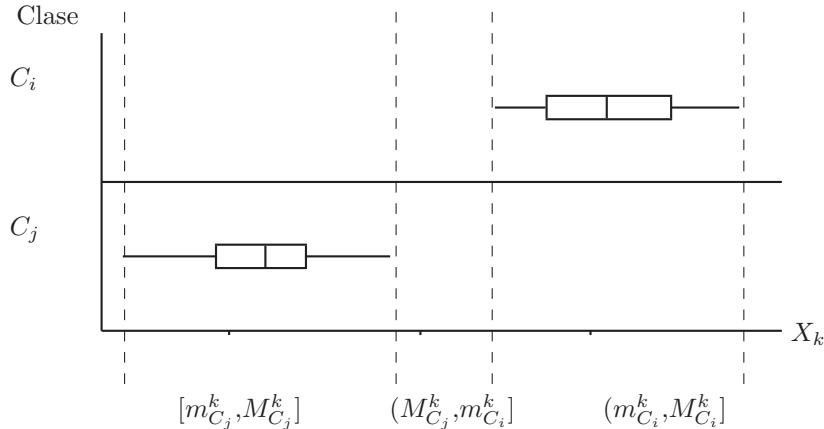


Figura 7.1: Caso 1: Boxplot based Discretization.

3. En realidad el problema que presenta la propuesta original es que los puntos de corte de los intervalos (es decir, los elementos de Z) vienen incluídos en el intervalo de su izquierda sean extremo inferior o extremo superior de su clase. Cuando se desplaza al intervalo de su izquierda el extremo inferior de una clase se produce un punto aislado de esa clase que, si se solapa con la otra clase genera reglas no seguras sólo porque excluye al límite del intervalo y genera reglas seguras donde deberían ser vacías, ver Figura 7.1.

$$I_1^{k,2} = [z_1^k, z_2^k] = [m_{Cj}^k, M_{Cj}^k]$$

$$I_2^{k,2} = (z_2^k, z_3^k] = (M_{Cj}^k, m_{Ci}^k]$$

$$I_3^{k,2} = (z_3^k, z_4^k] = (m_{Ci}^k, M_{Ci}^k]$$

Así para el caso particular de 2 clases se realizó un análisis por casos a fin de sistematizar dichas anomalías y diseñar una propuesta de corrección.

7.5.1 Análisis de los sistemas de intervalos inducidos a partir de una partición en 2 clases. Análisis por caso.

Sea $\mathcal{P}_\xi = \{C_i, \dots, C_\xi\}$ una partición en ξ clases de \mathcal{I} y uno de sus subconjuntos en 2 clases: $\mathcal{P}_2 = \{C_i, C_j\}$. Sobre la base del boxplot múltiple de X_k versus \mathcal{P}_2 se realiza un análisis por casos que permite proponer correcciones para mejorar la capacidad predictiva de los sistemas de reglas inducidos a partir de X_k de la discretización de las variables numéricas mediante el Boxplot based discretization (BbIR), ver §5.9 y §5.10.

Existen 13 casos posibles, en este apartado sólo se presentará uno de ellos y en el Apéndice B se puede ver en detalle el análisis de los otros 12 casos.

Caso 1, $M_{C_j}^k < m_{C_i}^k$

El caso 1, se presenta como uno de los casos extremos y además muy claro. La Figura 7.2, presenta un boxplot múltiple que corresponde a esta situación.

Como m_C^k y M_C^k son los mínimos y los máximos de la variable X_k en la clase $C \in \mathcal{P}$, tenemos que:

$$\begin{aligned} r_k^{C_i} &= [m_{C_i}^k, M_{C_i}^k], \text{ es el rango de la variable } X_k \text{ en la clase } C_i. \\ r_k^{C_j} &= [m_{C_j}^k, M_{C_j}^k], \text{ es el rango de la variable } X_k \text{ en la clase } C_j. \end{aligned}$$

Según la propuesta original presentada en (Vázquez and Gibert 2001), el sistema de intervalos inducido por \mathcal{P} sobre X_k sería:

$$I_1^{k,2} = [m_{C_j}^k, M_{C_j}^k],$$

$$I_2^{k,2} = (M_{C_j}^k, m_{C_i}^k],$$

$$I_3^{k,2} = (m_{C_i}^k, M_{C_i}^k]$$

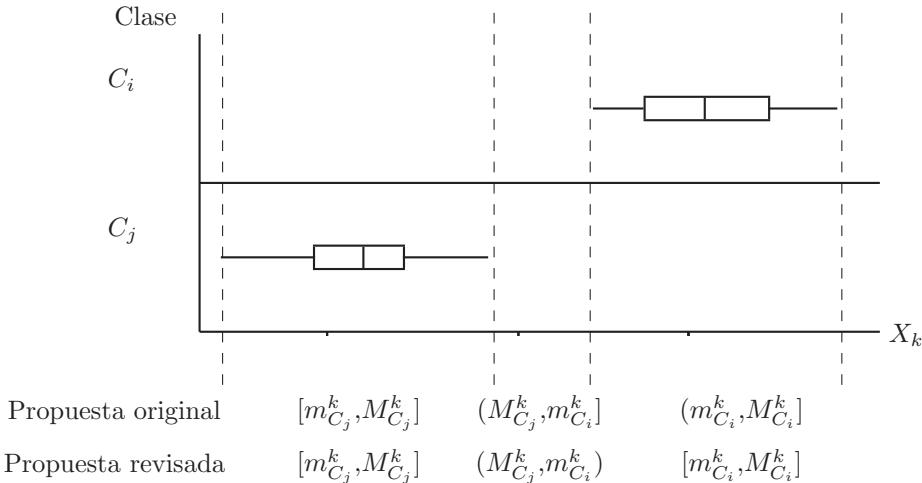


Figura 7.2: Caso 1: Boxplot based Discretization.

Lo que genera el siguiente sistema de reglas $\mathfrak{R}(X_k, \mathcal{P})$:

$$\Re(X_k, \mathcal{P}_2) = \{ \begin{array}{l} r_{1,j}^k : x_{ik} \in I_1^k \xrightarrow{p_{1j}=1} i \in C_j, \\ r_{1,i}^k : x_{ik} \in I_1^k \xrightarrow{p_{1i}=0} i \in C_i, \\ r_{2,j}^k : x_{ik} \in I_2^k \xrightarrow{p_{2j}=0} i \in C_j, \\ r_{2,i}^k : x_{ik} \in I_2^k \xrightarrow{p_{2i}=1} i \in C_i, \\ r_{3,j}^k : x_{ik} \in I_3^k \xrightarrow{p_{3i}=0} i \in C_j, \\ r_{3,i}^k : x_{ik} \in I_3^k \xrightarrow{p_{3i}=1} i \in C_i \end{array} \}$$

El sistema de reglas reducido es:

$$\Re^*(X_k, \mathcal{P}_2) = \{ \begin{array}{l} r_{1,j}^k : x_{ik} \in I_1^k \xrightarrow{p_{1j}=1} i \in C_j, \\ r_{2,i}^k : x_{ik} \in I_2^k \xrightarrow{p_{2i}=1} i \in C_i, \\ r_{3,i}^k : x_{ik} \in I_3^k \xrightarrow{p_{3i}=1} i \in C_i \end{array} \}$$

En este caso el sistema de reglas reducido es igual al sistema de reglas seguras e igual al sistema de reglas efectivas:

$$\Re^*(X_k, \mathcal{P}_2) = \mathcal{S}(X_k, \mathcal{P}_2) = \Re e(X_k, \mathcal{P}_2) = \Re e^*(X_k, \mathcal{P}_2)$$

Sin embargo I_2^k es un intervalo que contiene un único punto pegado a I_3^k , $m_{C_i}^k$, lo cual tiene muchos sentido si lo que se pretende es interpretar las clases, que en este caso toma valores altos en C_i y bajos en C_j sin más discusión, mientras $\Re(X_k, \mathcal{P}_2)$ sugiere que los valores medios también van a C_i .

Nuestra propuesta consiste en redefinir el sistema de intervalos de la siguiente forma:

$$\begin{aligned} I_1^{k,2} &= [m_{C_j}^k, M_{C_j}^k] \\ I_2^{k,2} &= (M_{C_j}^k, m_{C_i}^k) \\ I_3^{k,2} &= [m_{C_i}^k, M_{C_i}^k] \end{aligned}$$

Si cerramos los intervalos I_1^k y I_3^k por ambos lados y dejamos abierto el I_2^k en sus 2 extremos, el intervalo del centro es vacío, puesto que al tener $M_{C_j}^k < m_{C_i}^k$, $I_2^k = \emptyset$. En la Figura 7.2 se ha indicado cómo se formularían los intervalos bajo esta propuesta.

A partir del nuevo sistema de intervalos podríamos conseguir un sistema de reglas más compacto que la propuesta de (Vázquez and Gibert 2001) con sólo 2 reglas seguras que separan claramente las 2 clases.

Entonces el nuevo sistema de reglas completo $\mathcal{R}(X_k, \mathcal{P}_2)$ sería:

$$\mathcal{R}(X_k, \mathcal{P}_2) = \{ \begin{array}{l} r_{1,j}^k : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}=1} i \in C_j, \\ r_{1,i}^k : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=0} i \in C_i, \\ r_{2,j}^k : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}=0} i \in C_i, \\ r_{2,j}^k : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}=0} i \in C_j, \\ r_{3,i}^k : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3i}=1} i \in C_i, \\ r_{3,j}^k : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3j}=0} i \in C_j \end{array} \}$$

El sistema de reglas reducido es:

$$\begin{aligned}\mathcal{R}^*(X_k, \mathcal{P}_2) = \{ & r_{1,j}^k : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}=1} i \in C_j, \\ & r_{2,i}^k : x_{ik} \in I_3^{k,2} \xrightarrow{p_{2i}=0} i \in C_i, \\ & r_{3,i}^k : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3i}=1} i \in C_i \quad \}\end{aligned}$$

Teniendo en cuenta que las reglas no efectivas son prescindibles, el sistema de reglas efectivas:

$$\begin{aligned}\mathcal{R}e(X_k, \mathcal{P}_2) = \{ & r_{1,j}^k : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}=1} i \in C_j, \\ & r_{3,i}^k : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3i}=1} i \in C_i \quad \}\end{aligned}$$

Pero en el caso de la propuesta revisada sólo se cumple que el sistema de reglas efectivas es igual al sistema de reglas seguras:

$$\mathcal{R}e(X_k, \mathcal{P}_2) = \mathcal{R}e^*(X_k, \mathcal{P}_2) = \mathcal{S}(X_k, \mathcal{P}_2)$$

	n_v^k	n_g^k	n_p^k	$n_{\mathcal{R}}^k$	$n_{\mathcal{R}^*}^k$	n_e^k	$n_{\mathcal{S}}^k$	$n_{\mathcal{R}e^*}^k$
Original	0	0	3	6	3	3	3	3
Revisado	1	0	2	6	3	2	2	2

Tabla 7.3: Relación de valores y cardinalidad de los Sistema de reglas para el Caso 1.

Con lo cual obtenemos un sistema de reglas efectivas y un sistema de reglas seguras más compacto que en la propuesta original, ver tabla 7.3. Hemos logrado identificar una variable totalmente caracterizadora y representa un modelo más fiel a la situación vista en los boxplots.

Un razonamiento parecido ha permitido corregir los otros 12 casos estudiados, el detalle del análisis de éstos casos se presenta en el Anexo B.

7.5.2 Propuesta de corrección

Presentamos en la Tabla 7.4 el resumen de los 13 casos identificados. Por su parte la Tabla 7.5 resume como son los límites de los intervalos que se generan para cada caso. La columna de la derecha, llamada *columna patrón*, es una representación esquemática de la estructura del \mathcal{D}^k propuesto a partir del análisis por casos del punto anterior.

Este análisis evidencia que los sistemas de reglas que se desprenden de estos \mathcal{D}^k son sensibles a la forma como se definan los límites de los $I_s^{k,\xi}$.

Por su parte la Tabla 7.5 resume también los *tipos de intervalos* que se generan para cada caso y según sea el *BdD* revisado o no. Esto se puede observar en la columna llamada Tipo de $I_s^{k,\xi}$, para cada caso se indica el tipo de los 3 intervalos generados según la propuesta original (centro) y revisada (derecha). A partir de ésta información se ha podido establecer la propuesta de corrección que presentamos a continuación.

Tipos de patrones inducidos

A partir las Tablas 7.4, 7.5 y de la *columna patrón*, para la revisión, podemos identificar en qué casos la construcción del \mathcal{D}^k se puede hacer de una u otra forma partiendo de Z^k , lo que resulta de sumo interés para la posterior caracterización de las clases.

Observando la *columna patrón*, de la Tabla 7.5, se desprende que hay únicamente 2 formas de construir \mathcal{D}^k a partir de Z^k . Ambas generarán \mathcal{D}^k con 3 intervalos igual que la propuesta original, pero identificamos 2 patrones según como sean los límites de dichos intervalos (Gibert and Pérez-Bonilla 2006a):

1. Patrón *Centro Abierto*

En los casos 1 ($M_{C_j}^k < m_{C_i}^k$) y 2 ($M_{C_i}^k < m_{C_j}^k$), definimos 3 intervalos de tal forma que el del centro (I_2^k) es un intervalo *abierto* por ambos lados y un punto de corte para X_k $z_1^k \leq z_2^k \leq z_3^k \leq z_4^k$.

Dada una partición binaria y una variable X_k , el patrón centro abierto consiste en un sistema $I^{k,\xi}$ de intervalos que responde a la siguiente estructura;

$$I^{k,2} = \{I_1^{k,2}, I_2^{k,2}, I_3^{k,2}\}$$

donde:

$$I_1^{k,2} = [z_1^k, z_2^k]$$

$$I_2^{k,2} = (z_2^k, z_3^k)$$

$$I_3^{k,2} = [z_3^k, z_4^k]$$

2. Patrón *Centro Cerrado*

En los demás casos (4 a 13), con la revisión del *Boxplot based Discretization* original, definimos 3 intervalos de tal forma que el del centro (I_2^k) es un intervalo *cerrado* por ambos lados y un punto de corte para X_k $z_1^k \leq z_2^k \leq z_3^k \leq z_4^k$.

Dada una partición binaria y una variable X_k , el patrón centro cerrado consiste en un sistema $I^{k,\xi}$ de intervalos que responde a la siguiente estructura;

$$I^{k,2} = \{I_1^{k,2}, I_2^{k,2}, I_3^{k,2}\}$$

donde:

$$I_1^{k,2} = [z_1^k, z_2^k)$$

$$I_2^{k,2} = [z_2^k, z_3^k]$$

$$I_3^{k,2} = (z_3^k, z_4^k]$$

Caso	Característica	Boxplot Múltiple
1	$M_{C2}^k < m_{C1}^k$	
2	$m_{C1}^k < M_{C2}^k$	
3	$m_{C1}^k = m_{C2}^k \wedge M_{C1}^k = M_{C2}^k$	
4	$m_{C1}^k > M_{C2}^k \wedge M_{C1}^k > M_{C2}^k \wedge m_{C1}^k < M_{C2}^k$	
5	$m_{C1}^k < M_{C2}^k \wedge M_{C1}^k < M_{C2}^k \wedge m_{C2}^k > M_{C1}^k$	
6	$m_{C1}^k < M_{C2}^k \wedge M_{C1}^k > M_{C2}^k \wedge m_{C1}^k < M_{C2}^k$	
7	$m_{C1}^k > M_{C2}^k \wedge M_{C1}^k < M_{C2}^k \wedge m_{C1}^k < M_{C2}^k$	
8	$m_{C1}^k = M_{C2}^k \wedge M_{C1}^k < M_{C2}^k$	
9	$m_{C1}^k = M_{C2}^k \wedge M_{C1}^k > M_{C2}^k$	
10	$m_{C1}^k > M_{C2}^k \wedge M_{C1}^k = M_{C2}^k$	
11	$m_{C1}^k < M_{C2}^k \wedge M_{C1}^k = M_{C2}^k$	
12	$M_{C1}^k = m_{C2}^k$	
13	$M_{C2}^k = m_{C1}^k$	

Tabla 7.4: Descripción de los casos estudiados, ver detalles en Anexo B.

Caso Nº	Característica	Boxplot Múltiple	Patrón \mathcal{D}^k	Tipo de I_s^k (1)Original; (2)Revisada
1	$M_{C2}^k < m_{C1}^k$		[], (), []	$I_1^k P P$ $I_2^k P V$ $I_3^k P P$
2	$M_{C1}^k < m_{C2}^k$		[], (), []	$I_1^k P P$ $I_2^k P V$ $I_3^k P P$
3	$m_{C1}^k = m_{C2}^k \wedge M_{C1}^k = M_{C2}^k$		[], [], ()	$I_1^k O V$ $I_2^k O O$ $I_3^k V V$
4	$m_{C1}^k > m_{C2}^k \wedge M_{C1}^k > M_{C2}^k \wedge m_{C1}^k < M_{C2}^k$		[], [], ()	$I_1^k O P$ $I_2^k O O$ $I_3^k P P$
5	$m_{C1}^k < m_{C2}^k \wedge M_{C1}^k < M_{C2}^k \wedge m_{C2}^k > M_{C1}^k$		[], [], ()	$I_1^k G P$ $I_2^k G O$ $I_3^k P P$
6	$m_{C1}^k < m_{C2}^k \wedge M_{C1}^k > M_{C2}^k \wedge m_{C1}^k < M_{C2}^k$		[], [], ()	$I_1^k G P$ $I_2^k G O$ $I_3^k P P$
7	$m_{C1}^k > m_{C2}^k \wedge M_{C1}^k < M_{C2}^k \wedge m_{C1}^k < M_{C2}^k$		[], [], ()	$I_1^k G P$ $I_2^k G O$ $I_3^k P P$
8	$m_{C1}^k = m_{C2}^k \wedge M_{C1}^k < M_{C2}^k$		[], [], ()	$I_1^k G V$ $I_2^k G O$ $I_3^k P P$
9	$m_{C1}^k = m_{C2}^k \wedge M_{C1}^k > M_{C2}^k$		[], [], ()	$I_1^k G V$ $I_2^k G O$ $I_3^k P P$
10	$m_{C1}^k > m_{C2}^k \wedge M_{C1}^k = M_{C2}^k$		[], [], ()	$I_1^k G P$ $I_2^k G O$ $I_3^k V V$
11	$m_{C1}^k < m_{C2}^k \wedge M_{C1}^k = M_{C2}^k$		[], [], ()	$I_1^k G P$ $I_2^k G O$ $I_3^k P V$
12	$M_{C1}^k = m_{C2}^k$		[], [], ()	$I_1^k G P$ $I_2^k G O$ $I_3^k G P$
13	$M_{C2}^k = m_{C1}^k$		[], [], ()	$I_1^k G P$ $I_2^k V O$ $I_3^k G P$

Tabla 7.5: Patrón y descripción de los casos estudiados, ver detalles en Anexo B.

casos	Original	Revisada	Patrón	
caso 1	$I_1^k = [m_{C2}^k, M_{C2}^k]$ $I_2^k = (M_{C2}^k, m_{C1}^k)$ $I_3^k = (m_{C1}^k, M_{C1}^k)$	$I_1^k = [m_{C2}^k, M_{C2}^k]$ $I_2^k = (M_{C2}^k, m_{C1}^k)$ $I_3^k = [m_{C1}^k, M_{C1}^k]$	$I_2^k = \emptyset$	[], (), []
caso 2	$I_1^k = [m_{C1}^k, M_{C1}^k]$ $I_2^k = (M_{C1}^k, m_{C2}^k)$ $I_3^k = (m_{C2}^k, M_{C2}^k)$	$I_1^k = [m_{C1}^k, M_{C1}^k]$ $I_2^k = (M_{C1}^k, m_{C2}^k)$ $I_3^k = [m_{C2}^k, M_{C2}^k]$	$I_2^k = \emptyset$	[], (), []
caso 3	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = (m_{C1}^k, M_{C1}^k)$ $I_3^k = (M_{C1}^k, M_{C2}^k)$	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = [m_{C1}^k, M_{C1}^k]$ $I_3^k = (M_{C1}^k, M_{C2}^k)$	$I_1^k = \emptyset$ $I_3^k = \emptyset$	[], [], ()
caso 4	$I_1^k = [m_{C2}^k, m_{C1}^k]$ $I_2^k = (m_{C1}^k, M_{C2}^k)$ $I_3^k = (M_{C2}^k, M_{C1}^k)$	$I_1^k = [m_{C2}^k, m_{C1}^k]$ $I_2^k = [m_{C1}^k, M_{C2}^k]$ $I_3^k = (M_{C2}^k, M_{C1}^k)$		[], [], ()
caso 5	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = (m_{C2}^k, M_{C1}^k)$ $I_3^k = (M_{C1}^k, M_{C2}^k)$	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = [m_{C2}^k, M_{C1}^k]$ $I_3^k = (M_{C1}^k, M_{C2}^k)$		[], [], ()
caso 6	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = (m_{C2}^k, M_{C2}^k)$ $I_3^k = (M_{C2}^k, M_{C1}^k)$	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = [m_{C2}^k, M_{C2}^k]$ $I_3^k = (M_{C2}^k, M_{C1}^k)$		[], [], ()
caso 7	$I_1^k = [m_{C2}^k, m_{C1}^k]$ $I_2^k = (m_{C1}^k, M_{C1}^k)$ $I_3^k = (M_{C1}^k, M_{C2}^k)$	$I_1^k = [m_{C2}^k, m_{C1}^k]$ $I_2^k = [m_{C1}^k, M_{C1}^k]$ $I_3^k = (M_{C1}^k, M_{C2}^k)$		[], [], ()
caso 8	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = (m_{C2}^k, M_{C1}^k)$ $I_3^k = (M_{C1}^k, M_{C2}^k)$	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = [m_{C2}^k, M_{C1}^k]$ $I_3^k = (M_{C1}^k, M_{C2}^k)$	$I_1^k = \emptyset$	[], [], ()
caso 9	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = (m_{C2}^k, M_{C2}^k)$ $I_3^k = (M_{C2}^k, M_{C1}^k)$	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = [m_{C2}^k, M_{C2}^k]$ $I_3^k = (M_{C2}^k, M_{C1}^k)$	$I_1^k = \emptyset$	[], [], ()
caso 10	$I_1^k = [m_{C2}^k, m_{C1}^k]$ $I_2^k = (m_{C1}^k, M_{C2}^k)$ $I_3^k = (M_{C1}^k, M_{C2}^k)$	$I_1^k = [m_{C2}^k, m_{C1}^k]$ $I_2^k = [m_{C1}^k, M_{C2}^k]$ $I_3^k = (M_{C1}^k, M_{C2}^k)$	$I_3^k = \emptyset$	[], [], ()
caso 11	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = (m_{C2}^k, M_{C2}^k)$ $I_3^k = (M_{C1}^k, M_{C2}^k)$	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = [m_{C2}^k, M_{C2}^k]$ $I_3^k = (M_{C1}^k, M_{C2}^k)$	$I_3^k = \emptyset$	[], [], ()
caso 12	$I_1^k = [m_{C1}^k, M_{C1}^k]$ $I_2^k = (M_{C1}^k, m_{C2}^k)$ $I_3^k = (m_{C2}^k, M_{C2}^k)$	$I_1^k = [m_{C1}^k, m_{C2}^k]$ $I_2^k = [M_{C1}^k, m_{C2}^k]$ $I_3^k = (m_{C2}^k, M_{C2}^k)$		[], [], ()
caso 13	$I_1^k = [m_{C2}^k, M_{C2}^k]$ $I_2^k = (M_{C1}^k, M_{C1}^k)$ $I_3^k = (m_{C1}^k, M_{C1}^k)$	$I_1^k = [m_{C2}^k, m_{C1}^k]$ $I_2^k = [M_{C2}^k, m_{C1}^k]$ $I_3^k = (m_{C1}^k, M_{C1}^k)$		[], [], ()

Tabla 7.6: Comparación entre la propuesta original y la revisada para la construcción de los intervalos, ver detalles en Anexo B.

Propuesta BdD revisado

Desde el punto de vista algorítmico queda muy delimitado el planteamiento para construir los \mathcal{D}^k . Entonces el ***Boxplot based discretization (BdD) revisado para particiones binarias*** (Gibert and Pérez-Bonilla 2006a), (Pérez-Bonilla and Gibert 2006), tiene los siguientes pasos (ver ejemplo en la Figura 7.3):

1. Calcular *mínimo* ($m_{C_i}^k$) y *máximo* ($M_{C_i}^k$) de X_k en cada clase. Construir $\mathcal{M}^k = \{m_{C_i}^k, m_{C_j}^k, M_{C_i}^k, M_{C_j}^k\}$, donde $\text{card}(\mathcal{M}^k) = 4$
2. Construir el *conjunto de puntos de corte* \mathcal{Z}^k ordenando \mathcal{M}^k de manera ascendente; $\mathcal{Z}^k = \{z_i^k ; i = 1, \dots, 4\}$. Cada z_i^k es el punto donde la intersección entre clases cambia.
3. Construir el conjunto de intervalos $I^{k,\xi}$ inducido por \mathcal{P}_ξ sobre X_k , definiendo el intervalo $I_s^{k,\xi}$ entre cada par de valores consecutivos de \mathcal{Z}^k de la siguiente forma:

Si ($M_{C_j}^k < m_{C_i}^k$) o ($M_{C_i}^k < m_{C_j}^k$) *entonces generar un* \mathcal{D}^k ***centro abierto***:

$$\begin{aligned} I_1^{k,\xi} &= [z_1^k, z_2^k] \\ I_2^{k,\xi} &= (z_2^k, z_3^k) \\ I_3^{k,\xi} &= [z_3^k, z_4^k] \end{aligned}$$

sino generar un \mathcal{D}^k ***centro cerrado***:

$$\begin{aligned} I_1^{k,\xi} &= [z_1^k, z_2^k] \\ I_2^{k,\xi} &= [z_2^k, z_3^k] \\ I_3^{k,\xi} &= (z_3^k, z_4^k) \end{aligned}$$

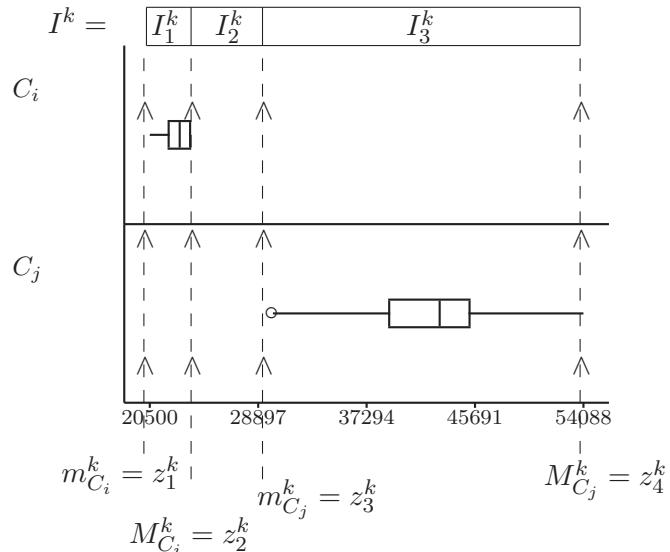


Figura 7.3: Ejemplo BdD revisado con patrón *centro abierto* para X_k vs \mathcal{P}_2 .

Éstos representarían un modelo más fiel a la situación vista en los Boxplot múltiples para cada caso descritos en detalle en el reporte de investigación (Pérez-Bonilla and Gibert 2005b).

Ninguna de estas formas coincide con la propuesta de (Vázquez and Gibert 2001).

7.5.3 Propiedades

De todo lo expuesto en las secciones anteriores, se desprenden algunas propiedades interesantes que podemos generalizar para la revisión del *BbD*, considerando los 2 tipos de patrones que tendrán los 3 intervalos inducidos.

	BbD revisado	
casos	$n_{Re^*}^k$	n_S^k
caso 1	2	2
caso 2	2	2
caso 3	1	0
caso 4	3	2
caso 5	3	2
caso 6	3	2
caso 7	3	2
caso 8	2	1
caso 9	2	1
caso 10	2	1
caso 11	2	1
caso 12	3	2
caso 13	3	2

Tabla 7.7: Nueva propuesta para los sistemas de reglas efectivas.

En términos específicos del análisis por casos y utilizando la Tabla 7.7 observamos que:

1. $\mathcal{S}(X_k, \mathcal{P}_2) = \emptyset$ ($n_S^k = 0$), cuando:

- $minC_i = minC_j \wedge MaxC_i = MaxC_j$

Esto ocurre en el caso 3

2. $n_S^k = 1$, cuando:

- $minC_i = minC_j \wedge MaxC_i \neq MaxC_j$

- $MaxC_i = MaxC_j \wedge minC_i \neq minC_j$

Esto ocurre en los casos 8, 9, 10, 11

3. $n_S^k = 2$, cuando:

- $minC_i \neq minC_j \wedge MaxC_i \neq MaxC_j$

Esto ocurre en los casos 1, 2, 4, 5, 6, 7, 12, 13

4. El *patrón centro abierto* identifica una variable totalmente caracterizadora y cumple que:

- $n_p^k = \xi$

- $n_g^k = 0$

- $n_v^k = 1$

- $n_S^k = n_p^k$

5. El número de reglas seguras siempre será menor o igual que 2, con lo cual:

$$n_{\mathcal{S}}^k = \text{card } \{\mathcal{S}(X_k, \mathcal{P}_2)\} \leq 2$$

6. Considerando todos los valores de X_k en una partición binaria, con la propuesta revisada tenemos que (ver Tabla 7.8):

- (a) $0 \leq n_p^k \leq 2$
- (b) $0 \leq n_g^k \leq 1$
- (c) $0 \leq n_v^k \leq 2$

Entonces con la nueva propuesta:

- Si $n_p^k = 0 \Rightarrow n_{\mathcal{R}e}^k = 2$
- Si $n_p^k = 1 \Rightarrow n_{\mathcal{R}e}^k = 3$
- Si $n_p^k = 2 \Rightarrow n_{\mathcal{R}e}^k = \{2, 4\}$
 - Si $n_g^k = 0 \Rightarrow n_{\mathcal{R}e}^k = 2$
 - Si $n_g^k = 1 \Rightarrow n_{\mathcal{R}e}^k = 4$

Caso	BbD Original			BbD Revisada		
	n_p^k	n_g^k	n_v^k	n_p^k	n_g^k	n_v^k
caso 1	3	0	0	2	0	1
caso 2	3	0	0	2	0	1
caso 3	0	2	1	0	1	2
caso 4	1	2	0	2	1	0
caso 5	1	2	0	2	1	0
caso 6	1	2	0	2	1	0
caso 7	1	2	0	2	1	0
caso 8	1	2	0	1	1	1
caso 9	1	2	0	1	1	1
caso 10	0	2	1	1	1	1
caso 11	0	2	1	1	1	1
caso 12	1	1	1	2	1	0
caso 13	1	1	1	2	1	0

Tabla 7.8: Tipos de valores.

7.5.4 Comparación de las propuestas

- En la estructura de los intervalos que genera

Observando la Tabla 7.8 tenemos que se da una de las siguientes situaciones:

- Se pierde un intervalo propio para ganar uno vacío. Esta situación se da en los casos 1 y 2, correspondiente a variables totalmente caracterizadoras.

Esto no satisface el criterio de comparación definido para decir que la nueva propuesta es mejor y puede parecer como no apropiado, pero no es así, pues la

efectividad sigue siendo la misma. Lo que ocurría con la propuesta anterior es que el intervalo del centro contenía un punto aislado que coincidía con el extremo inferior de la clase distribuida más a la derecha y generaba una regla segura para un único punto, siendo ésta una división artificial del recorrido de esa clase. La modificación que proponemos fusiona los 2 intervalos contiguos y propios de la misma clase en uno solo equivalente.

Hemos mejorado porque detectamos la variable caracterizadora que teníamos donde no se detectaba.

- Se pierde un intervalo genérico para ganar un vacío, en los casos 3, 8 y 9.

Todos coinciden en el número de valores propios, pero ganan en el número de intervalos vacíos.

Son todos casos donde el mínimo de las 2 clases coincide y da lugar a un intervalo genérico de un único punto que representa una división artificial del recorrido de X_k . Abrir el intervalo por la derecha lo deja vacío y el intervalo central, que se cierra por la izquierda, abraza este punto compactando el sistema de reglas resultante.

Aquí se mejora porque se reduce el número de reglas efectivas del sistema al fundirse en uno varios antecedentes.

- Se pierde un intervalo genérico para ganar uno propio, en los casos 4, 5, 6, 7, 10 y 11.

Aumenta el número de intervalos propios. Son todos los casos donde una clase se alarga más por la izquierda que la otra. En la versión original el primer intervalo era cerrado a ambos lados y abrazaba los valores mínimos de ambas clases. Al abrirlo por la derecha el punto mínimo de una de las 2 clases sale del intervalo y éste pasa de común a propio generando únicamente una regla segura, que antes no existía.

El nuevo sistema es mejor porque reconoce un intervalo propio donde no se detectaba.

- Se pierde un vacío para ganar un propio, en los casos 12 y 13.

Aumenta el número de intervalos propios. Los casos 12 y 13 se refieren a variables que serían *totalmente caracterizadoras* de no ser porque el extremo superior de una clase coincide con el inferior de la otra, lo que genera una coincidencia en un único punto que, en la propuesta original se considera en el intervalo izquierdo, y convierte en genérico un intervalo que conceptualmente sería propio dejando vacío el intervalo central. El patrón centro cerrado aquí deja este punto común en un intervalo central aislado y genérico que no enmascara el potencial caracterizador de la variable.

Aquí hemos mejorado también por aislar la intersección de las dos clases en un intervalo separado que da valor de propios a los otros intervalos.

• En los sistemas de reglas que induce

Teniendo en cuenta las características de los sistemas de intervalos generados en una y otra propuesta, cómo son estos intervalos, el número de intervalos vacíos o el cardinal de los sistemas de reglas derivados, o el número de reglas seguras y observando la Tabla 7.9 podemos señalar algunas características de la propuesta que presentamos aquí, en relación con la propuesta anterior.

Caso	BbD Original					BbD Revisada				
	n_0^k	n_e^k	$n_{Re^*}^k$	n_S^k	$n_e^k - n_S^k$	n_0^k	n_e^k	$n_{Re^*}^k$	n_S^k	$n_e^k - n_S^k$
caso 1	3	3	3	3	0	4	2	2	2	0
caso 2	3	3	3	3	0	4	2	2	2	0
caso 3	2	4	2	0	4	4	2	1	0	2
caso 4	1	5	3	1	4	2	4	3	2	2
caso 5	1	5	3	1	4	2	4	3	2	2
caso 6	1	5	3	1	4	2	4	3	2	2
caso 7	1	5	3	1	4	2	4	3	2	2
caso 8	1	5	3	1	4	3	3	2	1	2
caso 9	1	5	3	1	4	3	3	2	1	2
caso 10	2	4	2	0	4	3	3	2	1	2
caso 11	2	4	2	0	4	3	3	2	1	2
caso 12	3	3	2	1	2	2	4	3	2	2
caso 13	3	3	2	1	2	2	4	3	2	2

Tabla 7.9: Comparación para los sistemas de reglas completos.

La Tabla 7.9 confronta el número de reglas no efectivas, número de reglas efectivas, número de reglas seguras y el número de reglas efectivas y no seguras que se generan a partir de estos intervalos según las 2 propuestas. El valor que toma $n_{Re^*}^k = n_p^k + n_g^k$ mide el número de intervalos útiles (intervalos no vacíos), es decir, los que generan reglas efectivas, en el caso de particiones binarias y representa el cardinal del sistema de reglas efectivas reducido.

Sabemos que $n_e^k = n_p^k + 2n_g^k$ y que $n_e^k = n_R^k - n_0^k$ y el valor que toma $n_e^k - n_S^k$ corresponde al número de reglas que son efectivas, pero no seguras. De aquí podemos observar que:

1. El número de reglas seguras n_S es mayor en la propuesta revisada en 8 de los 13 casos, lo que representa un 62% de los casos.
2. En 3 de los 13 casos n_S se mantiene pero aumenta n_0 , lo que presenta ventajas, pues reduce la cardinalidad de los sistemas de reglas efectivas que de esos sistemas de intervalos se deriva, haciendo el sistema mas compacto.
3. En 2 de los 13 casos n_S es menor, pero corresponde al caso particular de variables totalmente caracterizadoras.
4. Por lo que se refiere al número de reglas no efectivas n_0 , vemos que, *en todos los casos, a excepción de los casos 12 y 13, el número de reglas no efectivas inducidas es mayor en la propuesta revisada*, esto producirá sistemas de reglas más compactos.

En realidad ocurre lo siguiente:

1. En los casos 4 a 7 y 10 a 13 (ver detalles en Anexo B), el número de reglas seguras aumenta.

Podemos distinguir 2 situaciones:

- (a) En los casos 4 a 7, 10 y 11, son todos los casos donde la revisión genera un valor propio donde había uno genérico. Al cambiar el patrón para la construcción de

los intervalos, el intervalo común que antes daba lugar a dos reglas de probabilidades complementarias, en la propuesta actual se convierte en propio, dando lugar únicamente a una regla segura, que antes no existía. Es decir se eliminarán pares de reglas de probabilidad complementarias que antes existían y que ahora se sustituyen por una única regla segura y una no efectiva. Así se reduce el número de reglas efectivas.

- (b) En los casos 12 y 13, son los casos donde se pierde un intervalo vacío para ganar uno propio. La consecuencia directa de la modificación es que se generará una regla segura en lugar de dos de probabilidades complementarias y por otro lado se generarán dos reglas de probabilidades complementarias en lugar de ninguna. Así, en la propuesta actual tendremos un sistema de 4 reglas, de las cuales 2 serán seguras, en lugar de las 3 que se tenían en la propuesta anterior (de las que únicamente una era segura). Aparece un intervalo común que antes no teníamos. De todos modos, éste contiene un único punto, y su cobertura será menor que la de los dos intervalos propios que se identifican, con lo que la modificación resultará beneficiosa en general.

Cabe observar que los casos 12 y 13 se refieren a variables que serían *totalmente caracterizadoras* de no ser porque el extremo superior de una clase coincide con el inferior de la otra, lo que genera una coincidencia en un único punto que conviene tratar de forma aislada.

Son los casos que cumplen que $m_{C_i}^k \in (m_{C_j}^k, M_{C_j}^k) \vee m_{C_j}^k \in (m_{C_i}^k, M_{C_i}^k)$

2. En los casos 3, 8 y 9 (ver detalles en Anexo B), se mantiene el número de reglas seguras, pero aumenta el número de reglas no efectivas.

Son todos los casos donde la revisión genera un valor vacío donde había uno genérico.

En todos estos casos se aplica el patrón ***centro cerrado***.

Podemos distinguir dos situaciones:

- (a) En los casos 8 y 9.

El número de reglas seguras se mantiene en estos casos por ser de la parte derecha de la distribución, donde la anterior propuesta ya presentaba buen comportamiento.

La cardinalidad $n_S^k = 1$ entonces sigue siendo la misma.

La cardinalidad n_e^k pasa de 5 a 3 porque se pierden 2 reglas, al pasar de un valor genérico a uno vacío.

- (b) En el caso 3.

El caso 3 nunca puede tener reglas seguras bajo ninguna formulación porque los rangos de la variable en ambas clases coinciden exactamente. La cardinalidad $n_S^k = 0$ entonces sigue siendo la misma.

La cardinalidad n_e^k pasa de 4 a 2.

En lugar de las 4 reglas que producía la propuesta anterior, se producen aquí únicamente 2 reglas de probabilidad complementarias porque se pierden 2 reglas, al pasar de un valor genérico a uno vacío.

Son los casos que cumplen que $m_{C_i}^k = m_{C_j}^k$

3. En los casos 1 y 2 (ver detalles en Anexo B), disminuye el número de reglas seguras.

En todos estos casos se aplica el patrón ***centro abierto***.

- (a) En los casos 1 y 2.

El sistema de reglas seguro disminuye en 1 una regla, es decir la n_S^k pasa de 2 a 3 por las razones argumentadas antes.

Son los únicos casos en que disminuye el número de reglas seguras. Pero en estos 2 casos particulares, reducir el número de reglas seguras nos entrega un sistema de reglas más compacto, ya que al cambiar el patrón en la construcción de los intervalos, lo que ocurre es que las reglas seguras de antes se fusionan en una única que conjunta sus antecedentes con lo que en realidad formalmente tendríamos sistemas equivalentes desde un punto de vista lógico.

Son los casos que cumplen que $(M_{C_j}^k < m_{C_i}^k) \vee (M_{C_i}^k < m_{C_j}^k)$

La propuesta revisada se comporta mejor, ya sea por producir mayor número de reglas seguras, lo que facilitaría encontrar valores y en consecuencia variables totalmente caracterizadoras o por producir mayor número de reglas no efectivas que revierte en sistemas de reglas más compactos.

7.6 Resumen del capítulo

En este capítulo se presenta en primer lugar una propuesta de tipificación de los sistemas de reglas en función del tipo de reglas que contienen. Los sistemas de reglas, pueden contener hasta 3 tipos de reglas que serán relevantes en la determinación de la calidad del Sistema, estos tipo son:

1. Reglas no efectivas: Una regla $r_{s,c}^k : x_{ik} \in I_s^k \xrightarrow{p_{sc}} i \in C$ es no efectiva si $p_{sc} = 0$
2. Reglas efectivas: Una regla $r_{s,c}^k : x_{ik} \in I_s^k \xrightarrow{p_{sc}} i \in C$ es efectiva si $p_{sc} > 0$
3. Reglas seguras: Una regla $x_{ik} \in I_s^k \xrightarrow{p_{sc}} i \in C$ es segura si $p_{sc} = 1$

Entonces, se propone considerar los siguiente conjuntos de reglas, que se derivan de 1 sola variable X_k :

1. Sistema de reglas completo $\mathcal{R}(X_k, \mathcal{P}_\xi)$, contiene reglas inducidas a partir de X_k
2. Sistema de reglas reducido, $\mathcal{R}^*(X_k, \mathcal{P}_\xi)$, contiene sólo la regla de mayor p_{sc} para cada antecedente I_s^k .
3. Sistemas de reglas de nivel η , $\mathcal{R}^\xi(X_k, \eta)$, poda $\mathcal{R}(X_k, \mathcal{P}_\xi)$ a las reglas con $p_{sc} \geq \eta$.
4. Sistema de reglas efectivas $\mathcal{R}e(X_k, \mathcal{P}_\xi)$, contiene solamente reglas efectivas ($p_{sc} > 0$).
5. Sistema de reglas efectivas reducido $\mathcal{R}e^*(X_k, \mathcal{P}_\xi)$, es $\mathcal{R}^*(X_k, \mathcal{P}_\xi)$ sin reglas no efectivas ($p_{sc} = 0$).
6. Sistema de reglas Seguras $\mathcal{S}(X_k, \mathcal{P}_\xi)$, contiene solamente reglas seguras ($p_{sc} = 1$).

En todos los casos existen los equivalentes de los sistemas globales que se construyen como unión directa de los inducidos para cada variable porque en este contexto se cumple que las reglas tienen siempre antecedentes simples correspondientes a una sola variable y no solapan con las generadas por otras variables.

En segundo lugar se analizan los tipos de valores que puede tomar la variable X_k según el tipo de regla que genera. Los valores de \mathcal{D}^k , ya sean modalidades o intervalos pueden ser de 5 tipos básicos:

1. Valor *totalmente caracterizador* de la clase C . Cubre todo C , es exclusivo y producirá una regla segura para C .
2. Valor *parcialmente caracterizador* de la clase C . Cubre sólo parte de C , es exclusivo de C y producirá una regla segura para C .
3. Valor *caracterizador no propio* de la clase C . Cubre todo C , no es exclusivo y producirá una regla con $p_{sc} \in (0, 1)$.
4. Valor *genérico* de la clase C . Cubre sólo parte de C , no es exclusivo y producirá una regla con $p_{sc} \in (0, 1)$.
5. Valor *vacío*. Los valores vacíos generan siempre reglas no efectivas.

Por otro lado hemos visto que existen tres tipos de intervalos diferentes y si X_K ha sido discretizada por el *BbD* sabemos, además, qué tipo de reglas genera cada uno:

1. El *intervalo propio* (P): Constituye un *valor propio* de la clase C y puede ser caracterizador total o parcial de la misma según si la cubre total o parcialmente. Todo intervalo propio cumple que $p_{sc} = 1$.
2. El *intervalo común* (G): Puede ser un valor genérico o un caracterizador no propio, genera ξ reglas de probabilidades complementarias y cumple siempre que $p_{sc} \in (0, 1)$.
3. El *intervalo vacío* (V): Dan lugar a valores vacíos y por tanto producen ξ reglas no efectivas.

En tercer lugar se proponen criterios de calidad de los sistemas de reglas en general, a partir de sus cardinales y las propiedades que de ellos se deducen.

1. Se cumple siempre que:

- $n_{\mathcal{R}} = n_e + n_0$
- $n_{\mathcal{S}} \leq n_e \leq n_{\mathcal{R}}$
- $n_0 = n_{\mathcal{R}} - n_e$

2. La mejor situación que puede ocurrir es:

- $n_{\mathcal{S}} \rightarrow \xi$ y $n_{\mathcal{S}} \rightarrow n_{\mathcal{R}}$ o bien ($\frac{n_{\mathcal{S}}}{n_{\mathcal{R}}} \rightarrow 1$)

Indirectamente $n_p \rightarrow \xi$ sería pues el mejor caso. De ahí se establece el criterio que:

Entre 2 propuestas es mejor la de mayor n_p

Si hay empate en n_p es mejor la de mayor n_v .

A continuación se analiza la relación entre los 3 tipos de valores (intervalos) de una variable n_p^k , n_g^k y n_v^k y los tipos de reglas que genera cada valor en $\mathcal{R}(\mathcal{P}_\xi)$, considerando que $\text{card } \mathcal{D}^k = n_p^k + n_g^k + n_v^k$.

Primero para una situación genérica:

- El número total de *reglas seguras* de $\mathcal{R}(\mathcal{P}_\xi)$:

$$n_{\mathcal{S}} = \sum_{\forall k} n_p^k$$

- El número total de *reglas efectivas* de $\mathcal{R}(\mathcal{P}_\xi)$:

$$n_e = \sum_{\forall k} n_p^k + \sum_{\forall k} \sum_{\forall s \text{ valor comun}} (E_{ks}) \quad 2 \leq E_{ks} \leq \xi$$

- El número total de *reglas no efectivas* de $\mathcal{R}(\mathcal{P}_\xi)$:

$$n_0 = \sum_{\forall k} \xi \text{ card } \mathcal{D}^k - \sum_{\forall k} n_p^k - \sum_{\forall k} \sum_{\forall s \text{ valor comun}} (E_{sk}) \quad 2 \leq E_{ks} \leq \xi$$

Después para el caso particular en que X_k ha sido discretizada por el *BbD*:

- $\text{card } \mathcal{D}^k = 2\xi - 1$
- $n_{\mathcal{R}} = K(2\xi^2 - \xi)$ reglas de las que $n_0^k = 2\xi^2 - \xi - n_p^k - \sum_{\forall s \text{ valor comun}} (E_{sk})$ con $2 \leq E_{ks} \leq \xi$ son no efectivas.

Y finalmente para la situación más concreta aún y relevante en esta tesis que X_k venga de *BbD* y $\xi = 2$:

- $n_{\mathcal{R}} = 6K$
- $n_{\mathcal{S}}^k = n_p^k \leq 3$

Por último, se presenta la revisión de los patrones establecidos en la propuesta original del *BbD* para particiones binarias. A partir de un análisis por casos se detectan situaciones anómalas:

1. En que había una división artificial del recorrido de la clase.
2. Y situaciones en que se generaban reglas no seguras sólo porque se incluía al límite de un intervalo y se generan reglas seguras donde deberían ser vacías.

Se aplican correcciones para evitar intervalos vacíos o intervalos con un sólo punto conexo a otro intervalo colindante; también para ganar un intervalo propio. Se observa que la propuesta resultante contempla 2 tipos de patrones para construir \mathcal{D}^k a partir de Z^k , ellos son:

Si $(M_{C_j}^k < m_{C_i}^k) \text{ o } (M_{C_i}^k < m_{C_j}^k)$ entonces generar un \mathcal{D}^k centro abierto:

$$\begin{aligned} I_1^{k,\xi} &= [z_1^k, z_2^k] \\ I_2^{k,\xi} &= (z_2^k, z_3^k) \\ I_3^{k,\xi} &= [z_3^k, z_4^k] \end{aligned}$$

sino generar un \mathcal{D}^k centro cerrado:

$$\begin{aligned} I_1^{k,\xi} &= [z_1^k, z_2^k] \\ I_2^{k,\xi} &= [z_2^k, z_3^k] \\ I_3^{k,\xi} &= (z_3^k, z_4^k) \end{aligned}$$

Siendo $m_C^k = \min X_K | C = \min_{i \in C} \{x_{ik}\}$ y $M_C^k = \max X_K | C = \max_{i \in C} \{x_{ik}\}$.

Entonces con la nueva propuesta cada variable genera 2, 3 o 4 reglas efectivas según el número de valores propios y genéricos que tenga la variable:

n_p^k	n_g^k	$n_{\mathcal{R}e}^k$
0		2
1		3
2	0	2
2	1	4

Tabla 7.10: Cardinalidad del sistema de reglas efectivas y número de intervalos propios.

La nueva propuesta mejora la anterior porque en:

1. En todos los casos excepto en 3 aumenta el número de intervalos propios, lo que facilitaría encontrar variables totalmente caracterizadoras (que se identifica con el patrón centro abierto).
2. De los 3 casos restantes en 1 se queda igual el número de intervalos propios n_p pero aumenta el número de valores vacíos de acuerdo con el criterio general definido.
3. En los otros 2 casos restantes tenemos sistemas equivalentes desde el punto de vista lógico a los anteriores pero más compactos porque aumenta n_v .

Las consecuencias inmediatas en lo que se refiere a la estructura de los sistemas de reglas:

1. En 8 de los 13 casos el sistema de reglas contiene más reglas seguras, lo que facilitaría encontrar variables totalmente caracterizadoras..
2. En los demás se mantiene pero crece el número de reglas no efectivas n_0 , lo que implica una mayor compactación de X_k ya que éstas son totalmente prescindibles.
3. Y en otros 2 se obtiene otro sistema equivalente que fusiona reglas seguras de antecedentes contiguos con mayor compactación, el número de reglas efectivas disminuye, lo que implica un aumento del número de reglas no efectivas.

Lo que se corresponde con las siguientes situaciones:

*Si $m_{C_i}^k \in (m_{C_j}^k, M_{C_j}^k) \vee m_{C_j}^k \in (m_{C_i}^k, M_{C_i}^k)$ entonces:
se gana un valor
propio y aumenta n_S*

*Si $m_{C_i}^k = m_{C_j}^k$ entonces:
se pierde un valor genérico
a favor de un vacío.
 n_S se mantiene
y n_0 aumenta*

*Si $(M_{C_j}^k < m_{C_i}^k) \vee (M_{C_i}^k < m_{C_j}^k)$ entonces:
se fusionan 2 propios para
ganar un vacío y aunque se
pierde una regla segura se tiene
un sistema de reglas equiva-
lente más compacto*

Capítulo 8

Ventajas de la jerarquía indexada

Una idea central de la tesis es que la existencia de una jerarquía indexada de clases permite abordar el problema de la interpretación de forma recursiva descendiendo en el dendograma, ello reduce cada iteración a la interpretación de un partición binaria y por ello al análisis que se presenta a continuación hace referencia al caso particular de particiones binarias simplificando el problema de hallar distintivos en las clases.

En una jerarquía binaria, necesariamente $\mathcal{P}_{\xi+1}$ está anidada en \mathcal{P}_ξ , es decir que las dos nuevas clases se desprenden de una y sólo una de las dos clases generadas en la partición anterior.

Esta propiedad se aprovecha en la metodología de caracterización conceptual por condicionamientos sucesivos considerando los pasos que se describen a continuación:

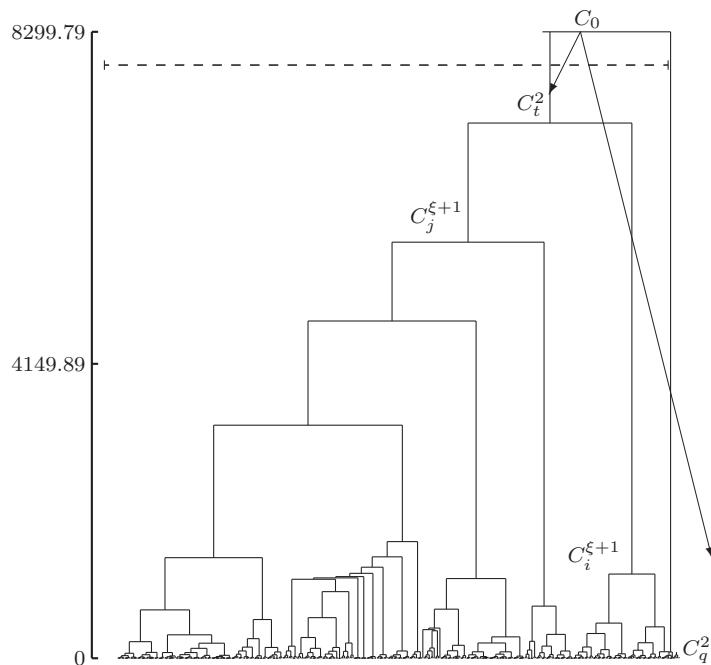


Figura 8.1: Árbol general de clasificación, corte en 2 clases.

1. Comenzar cortando el árbol por el nivel más alto, ver Figura 8.1, de esta manera obtenemos $\xi = 2$ clases. La partición formada por $\mathcal{P}_2 = \{C_t^2, C_q^2\}$.
- $\xi = 2$: Así, observado la Figura 8.1 tenemos que $\mathcal{P}_2 = \{C_t^2, C_q^2\}$. La raíz del árbol C_0 tiene 2 hijos:

$$C_0 \left\{ \begin{array}{l} C_t^2 \\ C_q^2 \end{array} \right.$$

2. Usar el *BbD* revisado (ver sección §7.5) y el *BbIR* (ver sección §5.10) para generar todos los sistemas de reglas $\mathcal{R}(X_k, \mathcal{P}_\xi)$ y construir el sistema de reglas completo global $\mathcal{R}(\mathcal{P}_\xi)$.

3. Determinar los conceptos:

A_t^ξ y A_q^ξ , a partir de $\mathcal{R}(X, \mathcal{P}_\xi)$ que permita distinguir C_t^ξ de C_q^ξ en \mathcal{I} . En el capítulo §9 se discuten algunos criterios para la selección de las mejores reglas de $\mathcal{R}(X, \mathcal{P}_\xi)$ a tal efecto y en el capítulo §10 se explica cómo se selecciona el concepto en cada iteración. En los capítulos §17 y §22 se presenta una aplicación real con detalles.

Hasta aquí lo que hemos hecho es inducir conceptos para C_t^ξ y C_q^ξ que permitan distinguirlas.

4. Bajar un nivel en el árbol a analizar, ver Figura 8.2, aprovechando la jerarquía indexada y ver qué clase es la que se abre. $\xi = \xi + 1$:

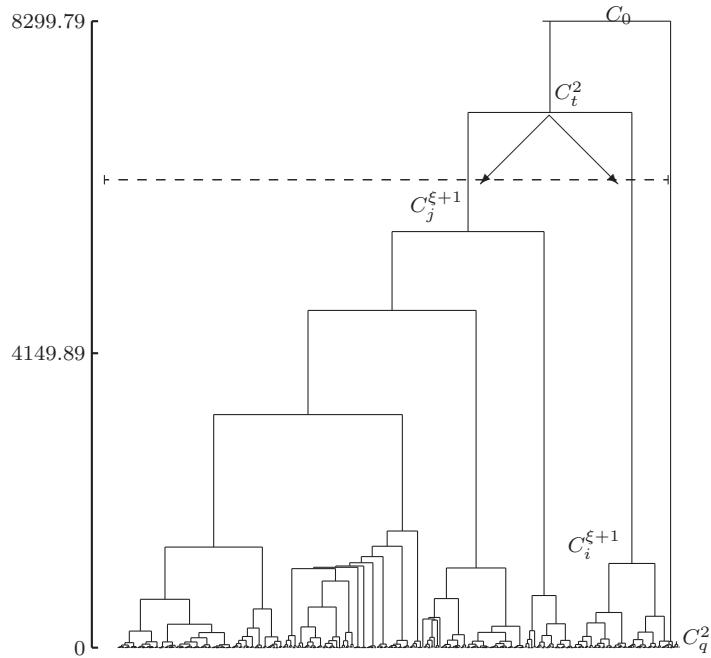


Figura 8.2: Árbol general de clasificación, corte en 3 clases.

5. Necesariamente $\mathcal{P}_{\xi+1}$ está anidada en \mathcal{P}_ξ , es decir que las dos nuevas clases se desprenden de una y sólo una de las dos clases generadas en la partición anterior. Sea:

- $C_i^{\xi+1}$ y $C_j^{\xi+1}$ las clases de $\mathcal{P}_{\xi+1}$ que subdividen una clase C_t^ξ de \mathcal{P}_ξ y
- $C_q^{\xi+1}$, la que ya estaba.

Así, $C_q^{\xi+1} = C_q^\xi$ y

$$C_t^\xi \left\{ \begin{array}{l} C_i^{\xi+1} \\ C_j^{\xi+1} \end{array} \right.$$

Y se cumple que $\mathcal{P}_{\xi+1} = \mathcal{P}_\xi \cup \{C_i^{\xi+1}, C_j^{\xi+1}\} \setminus \{C_t^\xi\}$.

Denominaremos entonces, $\mathcal{P}_{\xi+1}^* = \{C_i^{\xi+1}, C_j^{\xi+1}\}$, donde $\mathcal{P}_{\xi+1}^* \subset \mathcal{P}_{\xi+1}$ y $\mathcal{P}_{\xi+1}^* = \{\mathcal{P}_{\xi+1}\} \setminus \{C_q^{\xi+1}\}$.

En general identificar entre \mathcal{P}_ξ y $\mathcal{P}_{\xi+1}$ qué clase se divide se puede hacer automáticamente a partir de la tabla cruzada de ambas particiones:

En términos generales, para una tabla de contingencia de cualquier par de particiones consecutivas \mathcal{P}_ξ vs $\mathcal{P}_{\xi+1}$ se puede saber automáticamente cuál es la clase que se ha dividido:

$\mathcal{P}_\xi \times \mathcal{P}_{\xi+1}$	$C_i^{\xi+1}$	$C_j^{\xi+1}$	$C_q^{\xi+1}$...	$C_t^{\xi+1}$	
C_q^ξ	0	0	$n_{c_q}^\xi$...	0	0
C_t^ξ	$n_{c_i}^{\xi+1}$	$n_{c_t}^\xi - n_{c_i}^{\xi+1}$	0	0	0	0
\vdots			\vdots			\vdots
			$n_{c_\xi}^\xi$	0	0	0
			0	$n_{c_\xi}^\xi$	0	0
			0	0	$n_{c_\xi}^\xi$	0
C^ξ	0	0	0	0	0	$n_{c_\xi}^\xi$
	n_{c_i}	n_{c_j}	n_{c_q}	...	$n_{c_\xi}^\xi$	n

Tabla 8.1: Tabla de contingencia de \mathcal{P}_ξ vs $\mathcal{P}_{\xi+1}$.

A partir de la tabla de contingencia (ver tabla 8.1) y suponiendo que \mathcal{P}_ξ se ubica en filas, la clase de \mathcal{P}_ξ que se subdivide será la que tenga 2 casillas no nulas (o $\xi - 1$ ceros) en la fila que le corresponde, ya que todas las demás clases de \mathcal{P}_ξ tendrán exactamente una casilla no nula (o ξ ceros) en la fila que le corresponde.

Puesto que en el paso anterior, durante el análisis de \mathcal{P}_ξ , hemos separado

C_t^ξ de $C_q^{\xi+1}$ y $C_t^\xi = C_i^{\xi+1} \cup C_j^{\xi+1}$,

en este punto queda solamente separar:

$C_i^{\xi+1}$ de $C_j^{\xi+1}$,

repitiendo los pasos 2 y 3, para distinguir $C_i^{\xi+1}$ de $C_j^{\xi+1}$ en $\mathcal{I} \setminus C_q^\xi$ o sea en C_t^ξ , obtenemos un $A_i^{*\xi+1}$ y $A_j^{*\xi+1}$ que permite distinguir $C_i^{\xi+1}|C_t^\xi$ de $C_j^{\xi+1}|C_t^\xi$

6. Integrar el conocimiento extraído de la iteración $\xi + 1$ con el de la iteración ξ , permitirá determinar los conceptos finalmente asociados a los elementos de $\mathcal{P}_{\xi+1}$.

Puesto que las características comunes de $C_j^{\xi+1}$ y C_t^ξ , que han de permitir distinguirlas de $C_q^{\xi+1}$ ($= C_q^\xi$), se heredan de la clase C_t^ξ y se han identificado en la iteración anterior, se construirá un:

$$A_i^{\xi+1} = f_i(A_t^\xi, A_i^{*\xi+1})$$

$$A_j^{\xi+1} = f_j(A_t^\xi, A_j^{*\xi+1})$$

que separe ambas clases entre si y también de $C_q^{\xi+1}$.

A su vez, en el capítulo §10 se estudiará con que criterio elegir en cada iteración $A_i^{*\xi+1}$ y $A_j^{*\xi+1}$ y, finalmente, en el capítulo §10 se analizarán distintos criterios de construcción de conceptos que combinan f con los criterios de elección de $A_i^{*\xi+1}$ y $A_j^{*\xi+1}$.

7. Regresar al punto 4 hasta que \mathcal{P}_ξ coincida con la partición definitiva que se quiere interpretar.

La idea general es simplificar el problema de caracterizar las clases de la partición objetivo abordándolo por fragmentos, utilizando para ello las ventajas que presenta contar con una jerarquía indexada binaria. Es decir, la idea fundamental es poder caracterizar, distinguir y separar las clases de 2 en 2 de manera de aprovechar el conocimiento que se puede inducir en cada iteración para generar un concepto final que integre este conocimiento inducido en una interpretación de la partición objetivo.

Así, por iteración de un mecanismo que separa clases de 2 en 2 se logra resolver el problema más complejo de separar varias clases globalmente.

8.1 La hipótesis de mundo cerrado y la forma de f

La hipótesis de mundo cerrado (closed world assumption -CWA) (Reiter 1978a) y (Reiter 1978b) surge de la teoría de la información y de la lógica formal. Esta suposición establece que todo lo que es relevante en el mundo ha sido especificado en el modelo de representación del conocimiento, u observado en la base de datos. Por lo tanto, esto permite asumir de forma segura que un hecho es falso si no se puede inferir que es verdadero. Es decir, permite establecer una equivalencia entre lo falso y lo que no se observa, lo cual no deja de ser bastante temerario si no se tiene mucha seguridad en la representatividad de la muestra observada.

Así, partiendo de la hipótesis que la muestra observada es una representación completa del dominio y estableciendo una hipótesis fuerte de mundo cerrado se puede establecer que:

$$A_i^{\xi+1} = \neg A_q^\xi \wedge A_i^{*\xi+1}$$

y que

$$A_j^{\xi+1} = \neg A_q^\xi \wedge \neg A_i^{*\xi+1}$$

Siendo A_q^ξ el caracterizador de C_q^ξ y por tanto $\neg A_q^\xi$ el caracterizador de C_t^ξ y $A_i^{*\xi+1}$ el caracterizador de $C_i^{\xi+1}$ y por tanto $\neg A_i^{*\xi+1}$ el caracterizador de $C_j^{\xi+1}$ en virtud de dicha hipótesis fuerte de mundo cerrado.

Esto significa que la función f toma formas distintas para i y j y también para el caso en que no se asuma la hipótesis de mundo cerrado, donde se podría construir $A_i^{\xi+1} = A_t^\xi \wedge A_i^{*\xi+1}$, por ejemplo.

En efecto f puede ser la conjunción de un caracterizador y el del padre, o la conjunción del complementario de un hermano y el complementario de un tío según el caso.

Proponemos aquí una formulación unificada de tal modo que f tome siempre la misma forma, la conjuntiva.

$$f(A, A') = A \wedge A'$$

y trasladaremos todas estas diferencias a la formulación de los A_t^ξ , $A_i^{*\xi+1}$ y $A_j^{*\xi+1}$. Es decir construiremos dichos conceptos de forma que siempre:

$$A_i^{\xi+1} = A_t^\xi \wedge A_i^{*\xi+1}$$

y

$$A_j^{\xi+1} = A_t^\xi \wedge A_j^{*\xi+1}$$

y según los casos obtendremos A_t^ξ de los datos o negando A_q^ξ para mantener la homogeneidad formal.

8.2 Propiedades

Con esta formalización, para una partición $\mathcal{P} = \{C_1, C_2, \dots, C_\xi\}$, suponiendo que las clases se abren de la forma que muestra la Figura 8.3

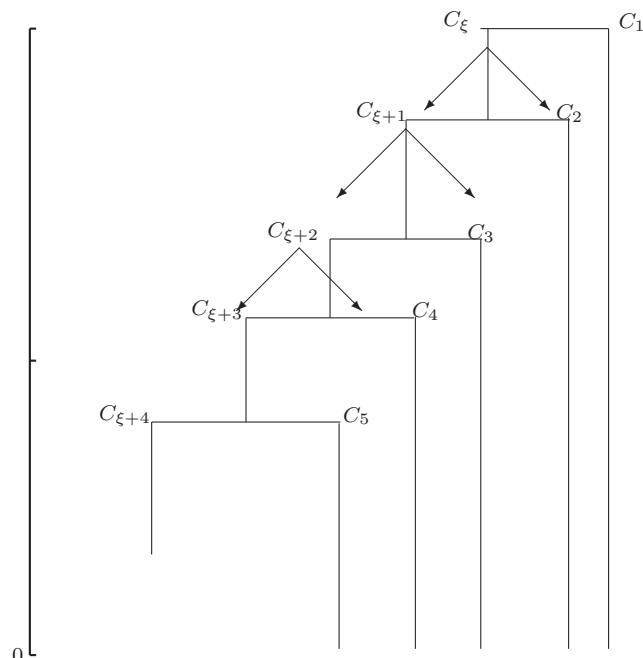


Figura 8.3: Dendrograma.

Como ya se formalizó en la definición del problema de tesis se quiere construir un sistema de conceptos $\mathcal{A}_{\mathcal{P}_\xi} = \{A_1, A_2, A_3, \dots, A_\xi\}$ que describen las clases.

Entonces para la partición $\mathcal{P} = \{C_1, C_2, \dots, C_\xi\}$, se tiene que:

$$\begin{aligned}
 A_1 &= A_1^2 \\
 A_2 &= A_\xi^2 \wedge A_2^3 \\
 A_3 &= A_\xi^2 \wedge A_{\xi+1}^3 \wedge A_3^4 \\
 A_4 &= A_\xi^2 \wedge A_{\xi+1}^3 \wedge A_{\xi+2}^4 \wedge A_4^5 \\
 A_5 &= A_\xi^2 \wedge A_{\xi+1}^3 \wedge A_{\xi+2}^4 \wedge A_{\xi+3}^5 \wedge A_5^6 \\
 A_6 &= A_\xi^2 \wedge A_{\xi+1}^3 \wedge A_{\xi+2}^4 \wedge A_{\xi+3}^5 \wedge A_{\xi+4}^6 \\
 &\vdots \\
 A_\xi &
 \end{aligned}$$

Así:

$$\mathcal{A}_{\mathcal{P}_\xi} = \{ A_1^2, A_\xi^2 \wedge A_2^3, A_\xi^2 \wedge A_{\xi+1}^3 \wedge A_3^4, A_\xi^2 \wedge A_{\xi+1}^3 \wedge A_{\xi+2}^4 \wedge A_4^5, A_\xi^2 \wedge A_{\xi+1}^3 \wedge A_{\xi+2}^4 \wedge A_{\xi+3}^5 \wedge A_5^6, A_\xi^2 \wedge A_{\xi+1}^3 \wedge A_{\xi+2}^4 \wedge A_{\xi+3}^5 \wedge A_{\xi+4}^6 \dots \}$$

Capítulo 9

Criterios de selección de reglas

En este capítulo se presentan los criterios con los que se selecciona la mejor regla en cada iteración de la metodología de interpretación de clasificaciones que se propone en el Capítulo §11 y los criterios para evaluar la base de conocimiento (o sistemas de reglas) que caracteriza las particiones \mathcal{P}_ξ que se quiere interpretar.

Asociar un concepto a cada clase de la forma “ $C : A$ ” es equivalente a construir una regla de la forma $r : A_C(i) \longrightarrow C$ para cada clase. $A_C(i)$ es el resultado de evaluar el antecedente $A_C(i)$ sobre un objeto $i \in \mathcal{I}$. Teniendo en cuenta que existirá cierta incertezza en el modelo, se propone tratar con reglas más genéricas de la forma $r : A_C(i) \xrightarrow{p} C$ donde $p \in [0, 1]$ es la probabilidad con que se cumple r . De este modo las reglas incorporan incertezza bajo una aproximación probabilística.

9.1 Criterios de evaluación para una única regla

Definimos a continuación algunos criterios que en la literatura se usan para evaluar la calidad de una regla sobre una Base de Datos y algunos que proponemos de forma específica para este trabajo.

9.1.1 Soporte ($Sup(r)$)

Dada una regla $r : A_C(i) \xrightarrow{p} C$ el soporte de r es la proporción de objetos de \mathcal{I} que satisfacen el antecedente de la regla, (Liu 2000).

$$Sup(r) = \frac{\text{card}\{i \in \mathcal{I} \text{ tq } A_C(i) = \text{true}\}}{n} \quad (9.1)$$

Mide cuantas veces se activa la regla r en la Base de Datos.

Si el soporte alcanza el 100% significa que todos los objetos de la Base de Datos satisfacen la regla.

9.1.2 Grado de certeza ($p(r)$)

Dada una regla r , el grado de certeza de r es la proporción de objetos del antecedente ($A_C(i) = \text{true}$) que están en C , $\forall C \in \mathcal{P}_\xi$. En (Liu 2000) este mismo concepto se define como *confidence of rule* y en general se calcula como:

$$p(r) = \frac{\text{card}\{i \in C \text{ tq } A_C(i) = \text{true}\}}{\text{card}\{A_C(i) = \text{true}\}} \quad (9.2)$$

donde $r : A_C(i) \xrightarrow{p} C$, $A_C(i)$ es verdadero si i satisface el antecedente $A_C(i)$, sea cuál sea la forma del antecedente de la regla (simple o compuesto).

Permite medir cuántas veces se equivocaría una regla $r : A_C(i) \xrightarrow{p} C$ al asignar un objeto i a una clase C . Si $p(r) = 0$ quiere decir que se equivoca siempre y si $p(r) = 1$ que no se equivoca nunca.

En el caso particular de una variable numérica que ha sido discretizada via que el *BbD*, ver Capítulo §5, el sistema de reglas inducido contiene reglas con antecedente de la forma $A = "x_{ik} \in I_s^{k,\xi}"$. Para una regla $r_{s,c}^k : x_{ik} \in I_s^{k,\xi} \xrightarrow{p_{sc}} i \in C$ en esta sección formalizamos la confianza de la regla $r_{s,c}^k$ que en capítulos anteriores (ver Capítulo §7) ya habíamos mencionado cuando X_k es una variable numérica. Se calcula como (detalles en (Pérez-Bonilla 2005)):

$$p_{sc} = P(C|I^{k,\xi} = I_s^{k,\xi}) = P(i \in C|x_{ik} \in I_s^{k,\xi}) = \frac{\text{card}\{i : x_{ik} \in I_s^{k,\xi} \wedge i \in C\}}{\text{card}\{i : x_{ik} \in I_s^{k,\xi}\}} \quad (9.3)$$

En el caso de que $I_s^{k,\xi}$ sea una modalidad de una variable categórica X_k la confianza se calcula como:

$$p_{sc} = P(C|I^{k,\xi} = I_s^{k,\xi}) = P(i \in C|x_{ik} = I_s^{k,\xi}) = \frac{\text{card}\{i : x_{ik} = I_s^{k,\xi} \wedge i \in C\}}{\text{card}\{i : x_{ik} = I_s^{k,\xi}\}} \quad (9.4)$$

9.1.3 Cobertura relativa ($CovR(r)$)

Dada una regla $r : A_C(i) \xrightarrow{p} C$, la cobertura relativa es la proporción de objetos de C que satisfacen el antecedente de la regla.

$$CovR(r) = \frac{\text{card}\{i \in C \text{ tq } A_C(i) = \text{true}\}}{n_c} \quad (9.5)$$

La Cobertura relativa mide cuántas veces se equivocaría la regla $r : A_C(i) \xrightarrow{p} C$ al describir la clase C con el antecedente $A_C(i)$.

En el caso particular de que X_k sea una variable numérica y el antecedente de la regla tome la forma $"x_{ik} \in I_s^{k,\xi}"$ y se tenga una regla $r_{s,c}^k : x_{ik} \in I_s^{k,\xi} \xrightarrow{p_{sc}} i \in C$ la Cobertura relativa se calcula como:

$$CovR(r) = \frac{\text{card}\{i \in C \text{ tq } x_{ik} \in I_s^{k,\xi}\}}{n_c} \quad (9.6)$$

En el caso de que $I_s^{k,\xi}$ sea una modalidad de una variable categórica X_k la Cobertura relativa se calcula como:

$$CovR(r) = \frac{\text{card}\{i \in C \text{ tq } x_{ik} = I_s^{k,\xi}\}}{n_c} \quad (9.7)$$

Propiedad:

Si $CovR = 100\%$ y $p_{sc} = 1$ la variable es totalmente identificadora de la clase, es decir es una variable totalmente caracterizadora.

En general las mejores reglas tendrán soporte, certeza y cobertura relativa alta, aunque habrá reglas de mucha calidad con otro comportamiento.

9.2 Criterios para la evaluación de un sistema de reglas o base de conocimiento

Estos mismos criterios permitirán después seleccionar, de entre varios sistemas de reglas construidos con distintos criterios, cuál es el mejor. Así, definimos aquí algunos criterios de evaluación global de un sistema de reglas.

Dada una base de conocimiento formada por un conjunto de reglas de la forma $r : A_C(i) \xrightarrow{p} C_j$. Siendo $A_C(i)$ el antecedente de cada regla evaluada sobre un objeto i :

9.2.1 Soporte total ($Sup_T(\mathcal{R})$)

Es el soporte total de la partición que se interpreta y es la suma de los soportes de cada regla compuesta asociada a cada una de las clases que conforman la partición final. Es frecuente utilizarlo en la literatura (Liu 2000) y representa el porcentaje de objetos que activa alguna regla de la base de conocimiento inducida para la partición final.

$$Sup_T(\mathcal{R}) = \sum_{\forall r \in \mathcal{R}} Sup(r) = \sum_{\forall r \in \mathcal{R}} \frac{\text{card}\{i \in \mathcal{I} \text{ tq } A_C(i) = \text{true}\}}{n} \quad (9.8)$$

Cuando se considera un conjunto de reglas \mathcal{R} es frecuente trabajar con el soporte total.

Propiedades:

1. Si $Sup_T(\mathcal{R}) = 100\%$ no hay ningún objeto sin asignar.
2. Si $Sup_T(\mathcal{R}) > 100\%$ hay objetos en inconsistencia satisfaciendo más de una regla. Por construcción serán reglas provenientes de distintas variables, aunque esto no implica necesariamente la inconsistencia.
3. Si $Sup_T(\mathcal{R}) < 100\%$ hay objetos sin asignar.

9.2.2 Certeza o confianza media ($\bar{p}(\mathcal{R})$)

La certeza promedio de la base de conocimiento de un sistema de reglas $\mathcal{R}(\mathcal{P}_\xi)$ será el promedio de los grados de certeza de cada una de las reglas (Liu 2000).

$$\bar{p}(\mathcal{R}) = \frac{\sum_{\forall r \in \mathcal{R}(\mathcal{P}_\xi)} p(r)}{n_{\mathcal{R}}} = \frac{\sum_{\forall r \in \mathcal{R}(\mathcal{P}_\xi)} \frac{\text{card}\{i \in C \text{ tq } A_C(i) = \text{true}\}}{\text{card}\{A_C(i) = \text{true}\}}}{n_{\mathcal{R}}} \quad (9.9)$$

9.2.3 Cobertura global ($CovG_{lobal}(\mathcal{R})$)

Dada una base de conocimiento formada por un conjunto de reglas de la forma $r : A_C(i) \xrightarrow{p} C_j$, la cobertura global es la proporción de objetos de \mathcal{I} que activan correctamente las reglas del sistema de reglas $\mathcal{R}(\mathcal{P}_\xi)$.

$$CovG_{lobal}(\mathcal{R}) = \frac{\sum_{\forall C \in \mathcal{P}_\xi} \text{card}\{i \in C \text{ tq } A_C(i) = \text{true}\} \times n_c}{n} \quad (9.10)$$

En general serán mejores los sistemas de mayor soporte total, mayor cobertura global y mejor certeza media aunque casi nunca se dará este tipo de situaciones ideales. Así habrá que definir heurísticos que hallen los mejores compromisos entre estos parámetros.

9.2.4 Evaluación de sistemas de reglas frente a una partición de referencia

Para el caso de disponer de una partición de referencia \mathcal{P} que indique la clases de cada individuo $i \in \mathcal{I}$, es posible realizar otro tipo de análisis relativo a cuán parecida es la asignación de clases inducida a partir del sistema de reglas y la clase real de referencia de cada objeto; que aporta también información sobre la calidad del sistema de reglas pero en este caso bajo una aproximación supervisada.

Supongamos que un cierto sistema de reglas \mathcal{R} ha inducido una partición sobre \mathcal{I} , $\mathcal{P}_{\mathcal{R}}$. Sea,

$\mathcal{P}_{\mathcal{R}i}$ clase de $\mathcal{P}_{\mathcal{R}}$ a la que pertenece el objeto i inducida por el sistema de reglas y

\mathcal{P}_i = la clase del objeto i en la partición de referencia.

Se requiere una partición de referencia \mathcal{P} , resultado de una Clasificación basada en reglas (ver §3.5) o proporcionada por el experto o cualquier otro método.

Inconsistencia

Denominamos inconsistencia al hecho de que para un mismo objeto hay reglas distintas correspondientes a distintas variables que se activan con un mismo elemento i pero tienen partes derechas diferentes (Pérez-Bonilla and Gibert 2007a), ver Figura 9.1.

Obj	NH4-influent	NH4-2aerobic	O2-1aerobic	...	C*	P
8	0,692	classer353	0,692	classer353	0,688	classer353
9	0,692	classer353	0,692	classer353	0,688	classer353
...
294	0,692	classer353	1,000	classer357	0,688	classer353
...
299	0,692	classer353	1,000	classer357	0,688	classer353

Figura 9.1: Ejemplo de inconsistencia.

Este fenómeno se produce al considerar todas las variables conjuntamente y la formalización es compleja porque detectar las reglas inconsistentes pasa por estudiar interacciones entre sus antecedentes, que refiere a variables distintas aunque un mismo objeto las pueda activar simultáneamente.

Sin embargo la contradicción que involucra satisfacer reglas de consecuentes distintos es clara y representa un problema grave para la interpretación de las clases. En (Pérez-Bonilla and Gibert 2007a) se proponen criterios para resolver éstas inconsistencias o evitarlas y decidir la clase final de i .

Objetos bien asignados

El número de objetos bien asignados se obtiene al comparar la clase a la que pertenece el objeto en la partición de referencia con la clase asignada a partir del sistema de reglas o base de conocimiento inducido. Si ambas clases son iguales el objeto se considera bien asignado.

$$\text{Card}\{i \in \mathcal{I} \text{ tq } \mathcal{P}_{\mathcal{R}i} = \mathcal{P}_i\}$$

Objetos mal asignados

El número de objetos mal asignados se obtiene al comparar la clase a la que pertenece el objeto en la partición de referencia con la clase asignada a partir del sistema de reglas o base de conocimiento inducido. Si ambas clases no son iguales, el objeto se considera mal asignando.

$$\text{Card}\{i \in \mathcal{I} \text{ tq } \mathcal{P}_{\mathcal{R}i} \neq \mathcal{P}_i\}$$

9.3 Resumen del capítulo

Dentro de los criterios de selección de reglas podemos distinguir 2 tipos de criterios:

- los criterios de evaluación para una única regla y
- los criterios para la evaluación de un sistema de reglas o base de conocimiento.

Al asociar un concepto a cada clase de la forma “ $C : A$ ” estamos construyendo una regla de la forma $r : A_C(i) \longrightarrow C$ para cada clase. Donde $A_C(i)$ es el resultado de evaluar el antecedente $A_C(i)$ sobre un objeto $i \in \mathcal{I}$. Teniendo en cuenta que existirá cierta incertezza en el modelo, se propone tratar con reglas más genéricas de la forma $r : A_C(i) \xrightarrow{p} C$ donde $p \in [0, 1]$ es la probabilidad con que se cumple r . De este modo las reglas incorporan incertezza bajo una aproximación probabilística.

De entre de los criterios que en la literatura se usan para evaluar la calidad de una regla sobre una Base de Datos y algunos que proponemos de forma específica para este trabajo tenemos que los mas usados son:

1. Soporte ($Sup(r)$): Mide cuantas veces se activa la regla r en la Base de Datos. Si el soporte alcanza el 100% significa que todos los objetos de la Base de Datos satisfacen la regla.
2. Grado de certeza ($p(r)$): Permite medir cuántas veces se equivocaría una regla r al asignar un objeto i a una clase C . Si $p(r) = 0$ quiere decir que se equivoca siempre y si $p(r) = 1$ que no se equivoca nunca.
3. Cobertura relativa ($CovR(r)$): La Cobertura relativa mide cuántas veces se equivocaría la regla r al describir la clase C con el antecedente $A_C(i)$.

Propiedad:

Si $CovR = 100\%$ y $p_{sc} = 1$ la variable es totalmente identificadora de la clase, es decir es una variable totalmente caracterizadora.

En general las mejores reglas tendrán soporte, certeza y cobertura relativa alta, aunque habrá reglas de mucha calidad con otro comportamiento.

Estos mismos criterios permitirán después seleccionar, de entre varios sistemas de reglas construidos con distintos criterios, cuál es el mejor. Así, definimos aquí algunos criterios de evaluación global de un sistema de reglas.

1. Soporte total ($Sup_T(\mathcal{R})$): Es el soporte total de la partición que se interpreta y es la suma de los soportes de cada regla compuesta asociada a cada una de las clases que conforman la partición final.
 - (a) Si $Sup_T(\mathcal{R}) = 100\%$ no hay ningún objeto sin asignar.
 - (b) Si $Sup_T(\mathcal{R}) > 100\%$ hay objetos en inconsistencia satisfaciendo más de una regla. Por construcción serán reglas provenientes de distintas variables, aunque esto no implica necesariamente la inconsistencia.
 - (c) Si $Sup_T(\mathcal{R}) < 100\%$ hay objetos sin asignar.
2. Certezza o confianza media ($\bar{p}(\mathcal{R})$): La certezza promedio de la base de conocimiento de un sistema de reglas $\mathcal{R}(\mathcal{P}_\xi)$ será el promedio de los grados de certezza de cada una de las reglas (Liu 2000).

3. Cobertura global ($CovG_{global}(\mathcal{R})$): La cobertura global es la proporción de objetos de \mathcal{I} que activan correctamente las reglas de $\mathcal{R}(\mathcal{P}_\xi)$.

En general serán mejores los sistemas de mayor soporte total, mayor cobertura global y mejor certeza media aunque casi nunca se dará este tipo de situaciones ideales. Así habrá que definir heurísticos que hallen los mejores compromisos entre estos parámetros.

Finalmente para el caso de disponer de una partición de referencia \mathcal{P} que indique la clases de cada individuo $i \in \mathcal{I}$, es posible realizar otro tipo de análisis relativo a cuán parecida es la asignación de clases inducida a partir del sistema de reglas y la clase real de referencia de cada objeto; que aporta también información sobre la calidad del sistema de reglas pero en este caso bajo una aproximación supervisada.

Supongamos que un cierto sistema de reglas \mathcal{R} ha inducido una partición sobre \mathcal{I} , $\mathcal{P}_{\mathcal{R}}$. Sea,

$\mathcal{P}_{\mathcal{R}i}$ clase de $\mathcal{P}_{\mathcal{R}}$ a la que pertenece el objeto i inducida por el sistema de reglas y \mathcal{P}_i = la clase del objeto i en la partición de referencia. Se requiere una partición de referencia.

1. Inconsistencia: Es el hecho de que para un mismo objeto hay reglas distintas correspondientes a distintas variables que se activan con un mismo elemento i pero tienen partes derechas diferentes.
2. Objetos bien asignados: Si la clase a la que pertenece el objeto en la partición de referencia con la clase asignada a partir del sistema de reglas son iguales el objeto se considera bien asignado ($Card\{i \in \mathcal{I} \text{ tq } \mathcal{P}_{\mathcal{R}i} = \mathcal{P}_i\}$).
3. Objetos mal asignados: Si la clase a la que pertenece el objeto en la partición de referencia con la clase asignada a partir del sistema de reglas no son iguales el objeto se considera bien asignado ($Card\{i \in \mathcal{I} \text{ tq } \mathcal{P}_{\mathcal{R}i} \neq \mathcal{P}_i\}$).

Capítulo 10

Integración del conocimiento

10.1 Introducción

Hasta ahora hemos visto:

1. Cómo obtener los sistemas de reglas,
2. Cómo aprovechar las ventajas que presenta la jerarquía indexada y
3. Cómo valorar si una regla es mejor que otra.

Lo que nos queda por estudiar es cómo se selecciona un $A_i^{\xi,k}$ cualquiera de entre todos los que componen un sistema de reglas.

Como ya hemos dicho una idea central de la tesis es que la existencia de una jerarquía indexada de clases permite abordar el problema de la interpretación de forma recursiva descendiendo en el dendograma, y por ello en el capítulo §8 de este apartado se muestran las ventajas de la estructura jerárquica. Ello reduce cada iteración a la interpretación de un partición binaria asociando un concepto que distinga cada par de clases y por ello buena parte del trabajo que se presenta a continuación hace referencia al caso particular de particiones binarias.

Utilizando los criterios de calidad definidos en el capítulo §9 se presentan 5 propuestas diferentes para combinar estos criterios con la hipótesis de mundo cerrado (CWA) y determinar los valores de $A_i^{\xi+1} = A_t^\xi \wedge A_i^{*\xi+1}$ y $A_j^{\xi+1} = A_t^\xi \wedge A_j^{*\xi+1}$ que conducirán a la interpretación final de la partición objetivo.

El punto de partida es el que ya se describió en el capítulo §8 donde se tiene $\mathcal{P}_{\xi+1}^* \subset \mathcal{P}_{\xi+1}$ y $\mathcal{P}_{\xi+1}^* = \{\mathcal{P}_{\xi+1}\} \setminus \{C_q^{\xi+1}\}$ y A_t^ξ su caracterizador determinado en la iteración anterior. Así supondremos que se ha inducido $\mathcal{R}(\mathcal{P}_{\xi+1}^*)$ tal y como se especifica en el capítulo §7.

Utilizando 2 de las herramientas de validación de interpretaciones definidas en el capítulo anterior (grado de certeza y cobertura relativa) tenemos que:

1. Por una parte, interesa que el grado de certeza media de la base de conocimiento que interpreta la partición objetivo $\mathbb{R}(\mathcal{P}_\xi) = \{r \mid tq \ r : A \xrightarrow{p(r)} C, \forall C \in \mathcal{P}_\xi\}$ sea máximo con lo cual se necesita que los conceptos asociados a cada par de clases en cada iteración tenga un grado de certeza $p(r)$ lo más grande posible. Es por ello que el primer criterio que utilizamos es trabajar el sistema de reglas $\mathcal{R}^\xi(\eta)$ que contenga sólo reglas seguras, es decir, trabajaremos con $\mathcal{S}(\mathcal{P}_\xi)$ o $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ según sea el caso. Si fuera vacío se elegiría una $\mathcal{R}^\xi(\eta)$ con un umbral alto.

2. Por otro lado interesa que la cobertura global de la base de conocimiento que interpreta la partición objetivo sea lo más grande posible de manera que la proporción de objetos de \mathcal{I} que activen correctamente la base de conocimiento sea máxima. Es por ello que se necesita que los conceptos asociados a cada par clases en cada iteración tenga una cobertura relativa lo más grande posible (en condiciones ideales del 100%).

Entonces considerando todas las reglas de $\mathcal{S}(\mathcal{P}_\xi)$ calculamos todas las coberturas relativas para todas las reglas $r_{s,c}^k : x_{ik} \in I_s^{k,\xi} \xrightarrow{1.0} i \in C \ \forall r_{s,c}^k \in \mathcal{S}(\mathcal{P}_\xi)$ utilizando la expresión (9.6) si X_k es una variable numérica discretizada en intervalos o (9.7) si X_k es una variable categórica ya definidas en el Capítulo §9 y que recordamos a continuación.

$$CovR(r_{s,c}^k) = \frac{\text{card}\{i \in C \text{ tq } x_{ik} \in I_s^{k,\xi}\}}{n_c}$$

ó

$$CovR(r) = \frac{\text{card}\{i \in C \text{ tq } x_{ik} = I_s^{k,\xi}\}}{n_c}$$

De esta manera construimos una tabla con la que se presenta a continuación.

Antecedente	Identificador regla	Cobertura absoluta	Cobertura relativa
$A_i^{\xi,k}$	$r_{s,c}^k$	$\text{card}\{i \in C \text{ tq } x_{ik} \in I_s^{k,\xi}\}$	$CovR(r_{s,c}^k)$
$A_i^{\xi,1}$	$r_{s,c}^1$...	$CovR(r_{s,c}^1)$
$A_i^{\xi,2}$	$r_{s,c}^2$...	$CovR(r_{s,c}^2)$
\vdots	\vdots	...	\vdots
$A_i^{\xi,K}$	$r_{s,c}^K$...	$CovR(r_{s,c}^K)$

Tabla 10.1: Cobertura relativa de $\mathcal{S}(\mathcal{P}_\xi)$.

Será asimismo interesante distinguir el caso en que se busca la mejor regla de todas o se restringe a reglas con una misma clase en el consecuente.

A continuación se presentan 5 formas distintas de construir $A_i^{*\xi+1}$ y $A_j^{*\xi+1}$.

10.2 Best Global concept and Close-World Assumption (BG &CWA)

Consiste en elegir la regla segura de mayor cobertura relativa de todas las del sistema de reglas $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$. Tomar esta regla haciendo CWA y utilizar su negación para la conceptualizar de la clase complementaria:

1. Restringir la búsqueda a la mejor regla de la base de conocimiento $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ para la partición restringida $P_{\xi+1}^* = \{C_i^{\xi+1}, C_j^{\xi+1}\}$.

2. Elegir el concepto ligado a la regla de mayor cobertura relativa de $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$. El consecuente de dicha regla determina qué clase queda caracterizada.

Determinar c, k, s tales que el concepto “ $X_k \in I_s^{k,\xi+1}$ ” tenga $p_{sc} = 1$, $C \in \mathcal{P}_{\xi+1}^*$, y cobertura relativa de la regla $r_{s,c}^k$ correspondiente sea máxima en la base de conocimiento $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$.

Si hay más de una construir \mathcal{K} con los indices de todas ellas.

3. Realizar una hipótesis fuerte de *Mundo cerrado* para describir la clase pareja en función del concepto complementario.
4. Determinar los conceptos $A_i^{*\xi+1}$ y $A_j^{*\xi+1}$ inducidos para $C_i^{\xi+1}$ y $C_j^{\xi+1}$ de la siguiente forma:

- (a) Si la regla identificada es única:

Sea

$$A^{\xi+1,k} = "X_k \in I_s^{k,\xi+1}" \quad y \quad A^{*\xi+1} = A^{\xi+1,k} \quad (10.1)$$

Si el consecuente de $r_{s,c}^k$ era C_i ,

$$A_i^{*\xi+1} = A^{*\xi+1} \quad y \quad A_j^{*\xi+1} = \neg A_i^{*\xi+1} \quad (10.2)$$

si no

$$A_j^{*\xi+1} = A^{*\xi+1} \quad y \quad A_i^{*\xi+1} = \neg A_j^{*\xi+1} \quad (10.3)$$

- (b) Si existe al menos una variable totalmente caracterizadora cada una de ellas da lugar a 2 reglas $r_{s,c}^k$ con $p_{sc} = 1$ y $CovR(r_{s,c}^k) = 100\%$ y todas las otras reglas tienen $CovR(r_{s,c}^k)$ menor.

Sea,

$$\mathcal{K}' = \{k \mid \text{tq } "X_k \text{ es totalmente caracterizada}"\}$$

Entonces;

$$A_i^{*\xi+1} = \bigwedge_{\forall k \in \mathcal{K}'} A_i^{\xi+1,k} \quad (10.4)$$

y

$$A_j^{*\xi+1} = \bigwedge_{\forall k \in \mathcal{K}'} A_j^{\xi+1,k} \quad (10.5)$$

- (c) Si no hay variables totalmente caracterizadoras y hay más de una regla $r_{s,c}^k$ con $p_{sc} = 1$ y cobertura relativa máxima, se consideran todas en la construcción del concepto y se realiza de la siguiente forma:

Sea

$$\mathcal{K}'' = \{k \mid \text{tq } "X_k \in I_s^{k,2}" \text{ corresponde a una regla segura de cobertura máxima}\}$$

Entonces;

$$A^{*\xi+1} = \bigvee_{\forall k \in \mathcal{K}''} A^{\xi+1,k} \quad (10.6)$$

Si el consecuente de $r_{s,c}^k$ era C_i ,

$$A_i^{*\xi+1} = A^{*\xi+1} \quad y \quad A_j^{*\xi+1} = \neg A_i^{*\xi+1} \quad (10.7)$$

si no

$$A_j^{*\xi+1} = A^{*\xi+1} \quad y \quad A_i^{*\xi+1} = \neg A_j^{*\xi+1} \quad (10.8)$$

10.3 Best local concept and no Close-World Assumption (BL &noCWA)

Consiste en elegir de entre todas las reglas de $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ que van a una misma clase, la de mayor cobertura relativa y no realizar la hipótesis de *CWA*. Así, no se utiliza la negación del concepto elegido para definir el concepto de la clase complementaria, si no, que para cada clase se utiliza el concepto seguro de mayor cobertura relativa.

1. Restringir la búsqueda a la mejor regla de la base de conocimiento $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ para la partición restringida $P_{\xi+1}^* = \{C_i^{\xi+1}, C_j^{\xi+1}\}$.
2. Considerar para cada clase $C_i^{\xi+1}$ y $C_j^{\xi+1}$ de $\mathcal{P}_{\xi+1}^*$ un *subsistema* de reglas que reúne las reglas que se dirigen a una misma clase:

$$\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*) = \{r_{s,c}^k : C = C_i \wedge r_{s,c}^k \in \mathcal{S}(\mathcal{P}_{\xi+1}^*)\}$$
 donde $\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*) \subseteq \mathcal{S}(\mathcal{P}_{\xi+1}^*)$
y

$$\mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*) = \{r_{s,c}^k : C = C_j \wedge r_{s,c}^k \in \mathcal{S}(\mathcal{P}_{\xi+1}^*)\}$$
 donde $\mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*) \subseteq \mathcal{S}(\mathcal{P}_{\xi+1}^*)$
3. Elegir el concepto ligado a la regla de mayor cobertura relativa de $\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)$ y el de $\mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)$.

Determinar k_i, s_i tales que el concepto “ $X_{k_i} \in I_{s_i}^{k_i, \xi+1}$ ” tenga $p_{s_i c_i} = 1$ y la cobertura relativa de la regla $r_{s_i, c_i}^{k_i}$ sea máxima en $\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)$ y k_j, s_j tales que el concepto “ $X_{k_j} \in I_{s_j}^{k_j, \xi+1}$ ” tenga $p_{s_j c_j} = 1$ tal que la cobertura relativa de la regla $r_{s_j, c_j}^{k_j}$ sea máxima en $\mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)$. $C_i \neq C_j$ y $C_i, C_j \in \mathcal{P}_{\xi+1}^*$.
Sea,

$\mathcal{K}_i = \{k \mid \text{tq } "X_k \in I_s^{k, \xi+1}" \text{ corresponde a una regla segura de cobertura máxima en } \mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)\}$

$\mathcal{K}_j = \{k \mid \text{tq } "X_k \in I_s^{k, \xi+1}" \text{ corresponde a una regla segura de cobertura máxima en } \mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)\}$.

4. Determinar los conceptos $A_i^{*\xi+1}$ y $A_j^{*\xi+1}$ inducidos para $C_i^{\xi+1}$ y $C_j^{\xi+1}$ de la siguiente forma:

- (a) Si se identifica una única regla por clase:

Sea

$$A_i^{\xi+1, k_i} = "X_{k_i} \in I_{s_i}^{k_i, \xi+1}" \text{ entonces, } A_i^{*\xi+1} = A_i^{\xi+1, k_i} \quad (10.9)$$

y

$$A_j^{\xi+1, k_j} = "X_{k_j} \in I_{s_j}^{k_j, \xi+1}" \text{ entonces, } A_j^{*\xi+1} = A_j^{\xi+1, k_j} \quad (10.10)$$

Para cada clase:

- (b) Si hay más de una regla $r_{s,c}^k$ con $p_{sc} = 1$ y cobertura relativa máxima, se consideran todas en la construcción del concepto y se realiza de forma diferente si la variable es total o parcialmente caracterizadora.

- Si X_k es totalmente caracterizadora (da lugar a reglas $p_{sc} = 1$ y $CovR(r_{s,c}^k) = 100\%$). Se construye de la siguiente forma:

$$A_i^{*\xi+1} = \bigwedge_{\forall k_i \in \mathcal{K}_i} A_i^{\xi+1,k_i} \quad (10.11)$$

y

$$A_j^{*\xi+1} = \bigwedge_{\forall k_j \in \mathcal{K}_j} A_j^{\xi+1,k_j} \quad (10.12)$$

- Si X_k es parcialmente caracterizadora (que da lugar a reglas $p_{sc} = 1$ y $CovR < 100\%$). Se construye de la siguiente forma:

$$A_i^{*\xi+1} = \bigvee_{\forall k_i \in \mathcal{K}_i} A_i^{\xi+1,k_i} \quad (10.13)$$

y

$$A_j^{*\xi+1} = \bigvee_{\forall k_j \in \mathcal{K}_j} A_j^{\xi+1,k_j} \quad (10.14)$$

10.4 Best local concept and Close-World Assumption (BL &CWA)

Consiste en elegir de entre todas las reglas de $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ que van a una misma clase, la de mayor cobertura relativa y realizar una hipótesis fuerte de *Mundo cerrado CWA*.

Así, para cada clase, se utiliza la negación del concepto elegido para definir la otra clase en disyunción lógica (\vee) con el concepto seguro obtenido por la máxima cobertura relativa.

1. Restringir la búsqueda a la mejor regla de la base de conocimiento $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ para la partición restringida $P_{\xi+1}^* = \{C_i^{\xi+1}, C_j^{\xi+1}\}$.
2. Considerar para cada clase $C_i^{\xi+1}$ y $C_j^{\xi+1}$ de $\mathcal{P}_{\xi+1}^*$ un *subsistema* de reglas que reúne las reglas que se dirigen a una misma clase:
 $\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*) \subseteq \mathcal{S}(\mathcal{P}_{\xi+1}^*)$ y $\mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*) \subseteq \mathcal{S}(\mathcal{P}_{\xi+1}^*)$
3. Elegir el concepto ligado a la regla de mayor cobertura relativa de $\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)$ y el de $\mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)$.

Determinar k_i, s_i tales que el concepto “ $X_{k_i} \in I_{s_i}^{k_i, \xi+1}$ ” tenga $p_{s_i c_i} = 1$ y la cobertura relativa de la regla $r_{s_i, c_i}^{k_i}$ sea máxima en $\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)$ y k_j, s_j tales que el concepto “ $X_{k_j} \in I_{s_j}^{k_j, \xi+1}$ ” tenga $p_{s_j c_j} = 1$ tal que la cobertura relativa de la regla $r_{s_j, c_j}^{k_j}$ sea máxima en $\mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)$. $C_i \neq C_j$ y $C_i, C_j \in \mathcal{P}_{\xi+1}^*$.

Sea,

$\mathcal{K}_i = \{k \mid tq \quad "X_k \in I_s^{k, \xi+1}" \text{ corresponde a una regla segura de cobertura máxima en } \mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)\}$.

$\mathcal{K}_j = \{k \mid tq \quad "X_k \in I_s^{k, \xi+1}" \text{ corresponde a una regla segura de cobertura máxima en } \mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)\}$.

4. Realizar una hipótesis fuerte de *Mundo cerrado* para describir la clase pareja en función del concepto complementario
5. Determinar los conceptos $A_i^{*\xi+1}$ y $A_j^{*\xi+1}$ inducidos para $C_i^{\xi+1}$ y $C_j^{\xi+1}$ de la siguiente forma:

- (a) Si se identifica una única regla por clase:

Sea $A_i^{\xi+1,k_i} = "X_{k_i} \in I_{s_i}^{k_i,\xi+1}"$ y
 $A_j^{\xi+1,k_j} = "X_{k_j} \in I_{s_j}^{k_j,\xi+1}"$

Finalmente;

$$A_i^{*\xi+1} = A_i^{\xi+1,k_i} \vee \neg A_j^{\xi+1,k_j} \quad (10.15)$$

$$A_j^{*\xi+1} = A_j^{\xi+1,k_j} \vee \neg A_i^{\xi+1,k_i} \quad (10.16)$$

- (b) Si hay más de una regla $r_{s,c}^k$ con $p_{sc} = 1$ y cobertura relativa máxima, se consideran todas en la construcción del concepto y se realiza de forma diferente si la variable es total o parcialmente caracterizadora.

- Si X_k es totalmente caracterizadora (da lugar a reglas $p_{sc} = 1$ y $CovR(r_{s,c}^k) = 100\%$). Se construye de la siguiente forma:

$$\begin{aligned} A_i^{*\xi+1} &= \bigwedge_{\forall k_i \in \mathcal{K}_i, k_j \in \mathcal{K}_j} (A_i^{\xi+1,k_i} \vee \neg A_j^{\xi+1,k_j}) = \\ &\quad (\bigwedge_{\forall k_i \in \mathcal{K}_i} A_i^{\xi+1,k_i}) \vee (\bigwedge_{\forall k_j \in \mathcal{K}_j} \neg A_j^{\xi+1,k_j}) \end{aligned} \quad (10.17)$$

y

$$\begin{aligned} A_j^{*\xi+1} &= \bigwedge_{\forall k_i \in \mathcal{K}_i, k_j \in \mathcal{K}_j} (A_j^{\xi+1,k_j} \vee \neg A_i^{\xi+1,k_i}) = \\ &\quad (\bigwedge_{\forall k_j \in \mathcal{K}_j} A_j^{\xi+1,k_j}) \vee (\bigwedge_{\forall k_i \in \mathcal{K}_i} \neg A_i^{\xi+1,k_i}) \end{aligned} \quad (10.18)$$

- Si X_k es parcialmente caracterizadora (que da lugar a reglas $p_{sc} = 1$ y $CovR(r_{s,c}^k) < 100\%$). Se construye de la siguiente forma:

$$\begin{aligned} A_i^{*\xi+1} &= \bigvee_{\forall k_i \in \mathcal{K}_i, k_j \in \mathcal{K}_j} (A_i^{\xi+1,k_i} \vee \neg A_j^{\xi+1,k_j}) = \\ &\quad (\bigvee_{\forall k_i \in \mathcal{K}_i} A_i^{\xi+1,k_i}) \vee (\bigvee_{\forall k_j \in \mathcal{K}_j} \neg A_j^{\xi+1,k_j}) \end{aligned} \quad (10.19)$$

y

$$\begin{aligned} A_j^{*\xi+1} &= \bigvee_{\forall k_i \in \mathcal{K}_i, k_j \in \mathcal{K}_j} (A_j^{\xi+1,k_j} \vee \neg A_i^{\xi+1,k_i}) = \\ &\quad (\bigvee_{\forall k_j \in \mathcal{K}_j} A_j^{\xi+1,k_j}) \vee (\bigvee_{\forall k_i \in \mathcal{K}_i} \neg A_i^{\xi+1,k_i}) \end{aligned} \quad (10.20)$$

10.5 Best local concept and partial Close-World Assumption (BL &partial-CWA)

Consiste en elegir de entre todas las reglas de $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ que van a una misma clase, la de mayor cobertura relativa y realizar una hipótesis fuerte de *Mundo cerrado CWA* condicionado.

Es decir, para cada clase, se utiliza la negación del concepto elegido para definir el concepto de la otra clase en disyunción lógica (\vee) con el concepto obtenido por la máxima cobertura relativa, pero cuando la variable referente a la mejor regla coincide en ambas clases (es decir las 2 reglas son inducidas por la misma variable) no se hace uso de la negación del concepto.

1. Restringir la búsqueda a la mejor regla de la base de conocimiento $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ para la partición restringida $P_{\xi+1}^* = \{C_i^{\xi+1}, C_j^{\xi+1}\}$.
 2. Considerar para cada clase $C_i^{\xi+1}$ y $C_j^{\xi+1}$ de $\mathcal{P}_{\xi+1}^*$ un *subsistema* de reglas que reúne las reglas que se dirigen a una misma clase:
- $$\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*) \subseteq \mathcal{S}(\mathcal{P}_{\xi+1}^*) \text{ y } \mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*) \subseteq \mathcal{S}(\mathcal{P}_{\xi+1}^*)$$
3. Elegir el concepto ligado a la regla de mayor cobertura relativa de $\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)$ y de $\mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)$. El consecuente de dicha regla determina qué clase queda caracterizada.

Determinar k_i, s_i tales que el concepto “ $X_{k_i} \in I_{s_i}^{k_i, \xi+1}$ ” tenga $p_{s_i c_i} = 1$ y la cobertura relativa de la regla $r_{s_i, c_i}^{k_i}$ sea máxima en $\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)$ y k_j, s_j tales que el concepto “ $X_{k_j} \in I_{s_j}^{k_j, \xi+1}$ ” tenga $p_{s_j c_j} = 1$ tal que la cobertura relativa de la regla $r_{s_j, c_j}^{k_j}$ sea máxima en $\mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)$. $C_i \neq C_j$ y $C_i, C_j \in \mathcal{P}_{\xi+1}^*$.

Sea,

$\mathcal{K}_i = \{k \mid \text{tq } "X_k \in I_s^{k, \xi+1}" \text{ corresponde a una regla segura de cobertura máxima en } \mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)\}$

$\mathcal{K}_j = \{k \mid \text{tq } "X_k \in I_s^{k, \xi+1}" \text{ corresponde a una regla segura de cobertura máxima en } \mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)\}$

4. Realizar una hipótesis fuerte de *Mundo cerrado* o no, según si la variable que induce la regla de cada clase es o no la misma, para describir la clase pareja en función del concepto complementario.
5. Determinar los conceptos $A_i^{*\xi+1}$ y $A_j^{*\xi+1}$ inducidos para $C_i^{\xi+1}$ y $C_j^{\xi+1}$ de la siguiente forma:
 - (a) Si se identifica una única regla por clase:
Sea $A_i^{\xi+1, k_i} = "X_{k_i} \in I_{s_i}^{k_i, \xi+1}"$ y
 $A_j^{\xi+1, k_j} = "X_{k_j} \in I_{s_j}^{k_j, \xi+1}"$
 - Si $k_i = k_j$

$$A_i^{*\xi+1} = A_i^{\xi+1, k_i} \quad (10.21)$$

$$A_j^{*\xi+1} = A_j^{\xi+1, k_j} \quad (10.22)$$

- Si no

$$A_i^{*\xi+1} = A_i^{\xi+1,k_i} \vee \neg A_j^{\xi+1,k_j} \quad (10.23)$$

$$A_j^{*\xi+1} = A_j^{\xi+1,k_j} \vee \neg A_i^{\xi+1,k_i} \quad (10.24)$$

- (b) Si hay más de una regla $r_{s,c}^k$ con $p_{sc} = 1$ y cobertura relativa máxima, se consideran todas en la construcción del concepto y se realiza de forma diferente si la variable es total o parcialmente caracterizadora.

- Si X_k es totalmente caracterizadora (da lugar a reglas $p_{sc} = 1$ y $CovR(r_{s,c}^k) = 100\%$). Se construye de la siguiente forma:

$$A_i^{*\xi+1} = (\bigwedge_{\forall k_i \in \mathcal{K}_i} A_i^{\xi+1,k_i}) \vee (\bigwedge_{\forall k_j \in \mathcal{K}_j, k_j \neq k_i} \neg A_j^{\xi+1,k_j}) \quad (10.25)$$

y

$$A_j^{*\xi+1} = (\bigwedge_{\forall k_j \in \mathcal{K}_j} A_j^{\xi+1,k_j}) \vee (\bigwedge_{\forall k_i \in \mathcal{K}_i, k_i \neq k_j} \neg A_i^{\xi+1,k_i}) \quad (10.26)$$

- Si X_k es parcialmente caracterizadora (que da lugar a reglas $p_{sc} = 1$ y $CovR(r_{s,c}^k) < 100\%$). Se construye de la siguiente forma:

$$A_i^{*\xi+1} = (\bigvee_{\forall k_i \in \mathcal{K}_i} A_i^{\xi+1,k_i}) \vee (\bigvee_{\forall k_j \in \mathcal{K}_j, k_j \neq k_i} \neg A_j^{\xi+1,k_j}) \quad (10.27)$$

y

$$A_j^{*\xi+1} = (\bigvee_{\forall k_j \in \mathcal{K}_j} A_j^{\xi+1,k_j}) \vee (\bigvee_{\forall k_i \in \mathcal{K}_i, k_i \neq k_j} \neg A_i^{\xi+1,k_i}) \quad (10.28)$$

10.6 Best local-global concept and Close-World Assumption (BL+G & CWA)

Consiste en elegir de entre todas las reglas de $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ que van a una misma clase, la de mayor cobertura relativa (*Best local concept*) y realizar una hipótesis fuerte de *Mundo cerrado CWA* tomando la negación de la mejor regla cuando para ambas clases se tiene la misma variable.

Así, para cada clase, se utiliza la negación del concepto elegido para definir el concepto de la otra clase en disyunción lógica (\vee) con el concepto seguro obtenido por la máxima cobertura relativa, pero cuando la mejor regla coincide en ambas clases (es decir las 2 reglas son inducidas por la misma variable) se elige de entre los 2 el que tiene la cobertura relativa mayor (*Best Global concept*), se trabaja con la negación de éste y se prescinde del otro concepto.

1. Restringir la búsqueda a la mejor regla de la base de conocimiento $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ para la partición restringida $P_{\xi+1}^* = \{C_i^{\xi+1}, C_j^{\xi+1}\}$.

2. Considerar para cada clase $C_i^{\xi+1}$ y $C_j^{\xi+1}$ de $\mathcal{P}_{\xi+1}^*$ un *subsistema* de reglas que reúne las reglas que se dirigen a una misma clase:

$$\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*) \subseteq \mathcal{S}(\mathcal{P}_{\xi+1}^*) \text{ y } \mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*) \subseteq \mathcal{S}(\mathcal{P}_{\xi+1}^*)$$

3. Elegir el concepto ligado a la regla de mayor cobertura relativa de $\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)$ y de $\mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)$ o el concepto de mayor cobertura relativa de $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ según la variable que induce el concepto sea o no la misma.

Determinar k_i, s_i tales que el concepto “ $X_{k_i} \in I_{s_i}^{k_i, \xi+1}$ ” tenga $p_{s_i c_i} = 1$ y la cobertura relativa de la regla $r_{s_i, c_i}^{k_i}$ sea máxima en $\mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)$ y k_j, s_j tales que el concepto “ $X_{k_j} \in I_{s_j}^{k_j, \xi+1}$ ” tenga $p_{s_j c_j} = 1$ tal que la cobertura relativa de la regla $r_{s_j, c_j}^{k_j}$ sea máxima en $\mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)$. $C_i \neq C_j$ y $C_i, C_j \in \mathcal{P}_{\xi+1}^*$.

Sea,

$\mathcal{K}_i = \{k \mid \text{tq } "X_k \in I_s^{k, \xi+1}" \text{ corresponde a una regla segura de cobertura máxima en } \mathcal{S}_{C_i}(\mathcal{P}_{\xi+1}^*)\}$

$\mathcal{K}_j = \{k \mid \text{tq } "X_k \in I_s^{k, \xi+1}" \text{ corresponde a una regla segura de cobertura máxima en } \mathcal{S}_{C_j}(\mathcal{P}_{\xi+1}^*)\}$

4. Realizar una hipótesis fuerte de *Mundo cerrado* para describir la clase pareja en función del concepto complementario según la variable que induce el concepto sea local o global.
5. Determinar los conceptos $A_i^{*\xi+1}$ y $A_j^{*\xi+1}$ inducidos para $C_i^{\xi+1}$ y $C_j^{\xi+1}$ de la siguiente forma:

- (a) Si se identifica una única regla por clase:

Sea $A_i^{\xi+1, k_i} = "X_{k_i} \in I_{s_i}^{k_i, \xi+1}"$ y
 $A_j^{\xi+1, k_j} = "X_{k_j} \in I_{s_j}^{k_j, \xi+1}"$

- Si $k_i = k_j$

– Si $CobR(r_{s_i, c_i}^{k_i}) > CobR(r_{s_j, c_j}^{k_j})$

$$A_i^{*\xi+1} = A_i^{\xi+1, k_i} \quad (10.29)$$

$$A_j^{*\xi+1} = \neg A_i^{\xi+1, k_i} \quad (10.30)$$

– Si $CobR(r_{s_i, c_i}^{k_i}) \leq CobR(r_{s_j, c_j}^{k_j})$

$$A_i^{*\xi+1} = \neg A_j^{\xi+1, k_j} \quad (10.31)$$

$$A_j^{*\xi+1} = A_j^{\xi+1, k_j} \quad (10.32)$$

- Si $k_i \neq k_j$

$$A_i^{*\xi+1} = A_i^{\xi+1, k_i} \vee \neg A_j^{\xi+1, k_j} \quad (10.33)$$

$$A_j^{*\xi+1} = A_j^{\xi+1, k_j} \vee \neg A_i^{\xi+1, k_i} \quad (10.34)$$

- (b) Si hay más de una regla $r_{s,c}^k$ con $p_{sc} = 1$ y cobertura relativa máxima, se consideran todas en la construcción del concepto y se realiza de forma diferente si la variable es total o parcialmente caracterizadora.

- Si X_k es totalmente caracterizadora (da lugar a reglas $p_{sc} = 1$ y $CovR(r_{s,c}^k) = 100\%$). Nunca se dará el caso $CobR(r_{s_i,c_i}^{k_i}) > CobR(r_{s_j,c_j}^{k_j})$ ó $CobR(r_{s_i,c_i}^{k_i}) < CobR(r_{s_j,c_j}^{k_j})$ porque las variables totalmente caracterizadoras siempre tienen $CovR(r_{s,c}^k) = 100\%$, con lo cual se elegirá aleatoriamente como la regla $r_{s,c}^k$ con $CovR(r_{s,c}^k)$ mayor a cualquiera de las 2. Se construye de la siguiente forma:

$$A_i^{*\xi} = \left(\bigwedge_{\forall k_i \in \mathcal{K}} A_i^{\xi, k_i} \right) \quad (10.35)$$

y

$$A_j^{*\xi} = \left(\bigwedge_{\forall k_i \in \mathcal{K}} \neg A_i^{\xi, k_i} \right) \quad (10.36)$$

- Si X_k es parcialmente caracterizadora (que da lugar a reglas $p_{sc} = 1$ y $CovR < 100\%$). Se construye de la siguiente forma:

$$\begin{aligned} A_i^{*\xi} = & \left(\bigvee_{\forall k_i \in \mathcal{K}_i, k_i \notin \mathcal{K}_j} A_i^{\xi, k_i} \right) \vee \left(\bigvee_{\forall k_j \in \mathcal{K}_j, k_j \notin \mathcal{K}_i} \neg A_j^{\xi, k_j} \right) \vee \\ & \left(\bigvee_{\substack{\forall k_i \in \mathcal{K}, \\ CobR(r_{s_i,c_i}^{k_i}) > CobR(r_{s_j,c_j}^{k_j})}} A_i^{\xi, k_i} \right) \vee \left(\bigvee_{\substack{\forall k_j \in \mathcal{K}, \\ CobR(r_{s_i,c_i}^{k_i}) < CobR(r_{s_j,c_j}^{k_j})}} \neg A_j^{\xi, k_j} \right) \end{aligned} \quad (10.37)$$

y

$$\begin{aligned} A_j^{*\xi} = & \left(\bigvee_{\forall k_j \in \mathcal{K}_j, k_j \notin \mathcal{K}_i} A_j^{\xi, k_j} \right) \vee \left(\bigvee_{\forall k_i \in \mathcal{K}_i, k_i \notin \mathcal{K}_j} \neg A_i^{\xi, k_i} \right) \vee \\ & \left(\bigvee_{\substack{\forall k_j \in \mathcal{K}, \\ CobR(r_{s_j,c_j}^{k_j}) > CobR(r_{s_i,c_i}^{k_i})}} A_j^{\xi, k_j} \right) \vee \left(\bigvee_{\substack{\forall k_i \in \mathcal{K}, \\ CobR(r_{s_j,c_j}^{k_j}) < CobR(r_{s_i,c_i}^{k_i})}} \neg A_i^{\xi, k_i} \right) \end{aligned} \quad (10.38)$$

10.7 Conclusión

Estos criterios para la integración del conocimiento se han aplicado a 2 casos de estudio (ver Capítulo 17 y Capítulo 22) y se ha podido concluir que el que mejor se comporta valorando la Cobertura global ($CovG_{global}$), ver capítulo §9, y la proximidad a la interpretación proporcionada por los expertos es el criterio *Best local-global concept and Close-World Assumption*, ver §10.6, porque en términos generales si se observa la interpretación de las clases proporcionada por el experto, la que más se acerca es *Best local-global concept and Close-World Assumption*, básicamente por la forma en que se comportan las variables. Como el objetivo primordial es favorecer la riqueza conceptual de la interpretación o grado de interpretabilidad (y/o utilidad) de las clases formadas, la propuesta *Best local-global concept and Close-World*

Assumption es la mejor, aunque si se considera la capacidad predictiva (permitir para un nuevo objeto (día), predecir la clase (situación típica de la planta) que le corresponde y generar las caracterización e interpretación conceptual correspondiente a esa clase), el hecho que existan inconsistencias genera conflictos, lo que ha se ha estudiado previamente en (Pérez-Bonilla and Gibert 2007a), (Pérez-Bonilla and Gibert 2008a) donde se proponen diversos métodos de resolución de conflictos.

Por ello, en el próximo capítulo se presenta la una formulación definitiva donde se utiliza dicho criterio para conformar la metodología final que se propone.

Capítulo 11

Metodología de caracterización conceptual por condicionamientos sucesivos CCCS

11.1 Introducción

Como ya se formalizó en la definición del problema de tesis se quiere construir un sistema de conceptos $\mathcal{A}_{\mathcal{P}_\xi} = \{A_1, A_2, A_3, \dots, A_\xi\}$ que describen las clases de tal forma que:

- $A, A' \in \mathcal{A}_{\mathcal{P}_\xi} \Rightarrow A \neq A'$
- $\forall i \in \mathcal{I}, A_C(i) = \text{true} , \text{ si } C = C(i, \mathcal{P}_\xi), A_C \in \mathcal{A}_{\mathcal{P}_\xi}$
- $\forall i \in \mathcal{I}, A_C(i) = \text{false}, \text{ si } C \neq C(i, \mathcal{P}_\xi), A_C \in \mathcal{A}_{\mathcal{P}_\xi}$

Siendo; $A_C \in \mathcal{A}_{\mathcal{P}_\xi}$ una expresión booleana que identifica de forma unívoca los elementos de C , A_C función exclusiva de las variables X_1, X_2, \dots, X_K y los valores que éstas toman en cada clase.

$A_C(i)$ la evaluación de A_C sobre las coordenadas x_{ik} del individuo $i \in \mathcal{I}$.

Así,

$$A = [(\langle X \rangle \langle opr \rangle \langle v \rangle) \langle opl \rangle]^*$$

La *metodología de caracterización conceptual por condicionamientos sucesivos CCCS* (Pérez-Bonilla, Gibert, and Vrecko 2008), representa una propuesta metodológica que utiliza las ventajas existentes en una jerarquía indexada τ utilizando la propiedad de ser una estructura jerárquica binaria ($\mathcal{P}_{\xi+1}$ tiene las mismas clases de \mathcal{P}_ξ menos una, que es la que se divide en 2 sub-clases en $\mathcal{P}_{\xi+1}$).

La metodología CCCS pretende generar interpretaciones automáticas de una partición \mathcal{P} perteneciente a una jerarquía indexada τ , $\tau = \{\mathcal{P}_\xi, \xi = 1 : n - 1\}$, donde n es el número de objetos. Habitualmente τ es el resultado de una clasificación jerárquica.

En adelante los sistemas de reglas se inducen para distinguir las clases de una partición final que se desea interpretar \mathcal{P}_ξ y que puede ser el resultado de aplicar un BbIR a todas las variables disponibles o puede tener otro origen cualquiera.

La metodología que proponemos aprovecha la estructura jerárquica de la clasificación objetivo para inducir conceptos iterando sobre las subdivisiones binarias del dendrograma (ver capítulo §8). Por esta razón, se asocia un sistema de reglas en cada iteración a particiones restringidas a 2 clases:

$$\mathcal{P}_{\xi+1}^* = \{C_i^{\xi+1}, C_j^{\xi+1}\} \text{ donde } \mathcal{P}_{\xi+1}^* \subset \mathcal{P}_{\xi+1} \text{ y } \mathcal{P}_{\xi+1}^* = \{\mathcal{P}_{\xi+1}\} \setminus \{C_q^{\xi+1}\}$$

y entonces se considerará:

- $\mathcal{R}(\mathcal{P}_{\xi+1}^*)$ para $\mathcal{P}_{\xi+1}^*$, es decir $\mathcal{R}(\mathcal{P}_{\xi+1}^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_{\xi+1}^*)$.
- Por último definimos: $\mathbb{R}(\mathcal{P}_\xi) = \{r \mid t q \mid r : A \xrightarrow{p(r)} C, \forall C \in \mathcal{P}_\xi\}$ donde $\mathbb{R}(\mathcal{P}_\xi)$ contiene el conjunto de conceptos A que permiten distinguir cada clase de la partición \mathcal{P}_ξ (dichos conceptos serán eventualmente compuestos según sea el caso, ver capítulo §10)

Finalmente la propuesta metodológica definitiva es:

11.2 Propuesta metodológica

La *metodología de caracterización conceptual por condicionamientos sucesivos CCCS* consiste en los siguientes pasos

1. Por el momento $\mathcal{R}(\mathcal{P}_\xi^*) = \emptyset$
2. Comenzar cortando el árbol por el nivel más alto. Haciendo $\xi = 2$ clases obtenemos la partición formada por $\mathcal{P}_2 = \{C_t^2, C_q^2\}$. Supongamos $\mathcal{P}_\xi^* = \{C_i^\xi, C_j^\xi\}$. Para el caso $\xi = 2$, $C_i^\xi = C_t^2$ y $C_j^\xi = C_q^2$.
3. Para todas las variables numéricas de $\mathcal{P}_\xi^* = \{C_i^\xi, C_j^\xi\}$, usar **boxplot based discretization (BbD)** presentado formalmente en la sección §7.5 para encontrar (totales o parciales) valores característicos
 - (a) Calcular *mínimo* (m_C^k) y *máximo* (M_C^k) de X_k en cada clase. Construir $\mathcal{M}^k = \{m_{C_1}^k, \dots, m_{C_\xi}^k, M_{C_1}^k, \dots, M_{C_\xi}^k\}$
 - (b) Construir el conjunto de intervalos $I^{k,\xi}$ inducido por \mathcal{P}_ξ^* sobre X_k , definiendo el intervalo $I_s^{k,\xi}$ entre cada par de valores consecutivos de \mathcal{Z}^k de la siguiente forma:

Si ($M_{C_2}^k < m_{C_1}^k$) o ($M_{C_1}^k < m_{C_2}^k$) entonces generar un \mathcal{D}^k centro abierto:

$$\begin{aligned} I_1^{k,\xi} &= [z_1^k, z_2^k] \\ I_2^{k,\xi} &= (z_2^k, z_3^k) \\ I_3^{k,\xi} &= [z_3^k, z_4^k] \end{aligned}$$

sino generar un \mathcal{D}^k centro cerrado:

$$\begin{aligned} I_1^{k,\xi} &= [z_1^k, z_2^k] \\ I_2^{k,\xi} &= [z_2^k, z_3^k] \\ I_3^{k,\xi} &= (z_3^k, z_4^k) \end{aligned}$$

4. Para todas las variables, numéricas o categóricas de $\mathcal{P}_\xi^* = \{C_i^\xi, C_j^\xi\}$, usar el **boxplot based induction rules (BbIR)**, para inducir el sistema de reglas, $\mathcal{R}(\mathcal{P}_\xi^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_\xi^*)$, que caracteriza a ambas clases.
 - (a) Si la variable es numérica, utilizar $I^{k,\xi} = \{I_1^k, I_2^k, \dots, I_{2\xi-1}^k\}$ obtenido con el *Boxplot based discretization* en el paso anterior.

(b) Construir la tabla $I^{k,\xi} \mid \mathcal{P}_\xi$.d Donde,

$$p_{sc} = P(C|I^{k,\xi} = I_s^{k,\xi}) = P(i \in C|x_{ik} \in I_s^{k,\xi}) = \frac{\text{card}\{i : x_{ik} \in I_s^{k,\xi} \wedge i \in C\}}{\text{card}\{i \in \mathcal{I} : x_{ik} \in I_s^{k,\xi}\}} = \frac{n_{sc}}{\sum_{\forall s} n_{sc}}$$

Si $p_{sc} = 1$, $I_s^{k,\xi}$ es el conjunto de *valores propios* de X_k .

Si $\forall s = 1 : 2\xi - 1, \exists C \in \mathcal{P}_\xi$ tq $p_{sc} = 1$ y $\text{Cov}(r_{s,c}^k) = 100\%$ entonces X_k es una *variable totalmente caracterizadora*.

(c) Para cada celda de la Tabla $I^k \mid \mathcal{P}_\xi$, generar la regla:

$$r_{s,c}^k : Si \ x_{ik} \in I_s^{k,\xi} \xrightarrow{p_{sc}} i \in C$$

5. Construir $\mathcal{R}(\mathcal{P}_\xi^*)$ y $\mathcal{S}(\mathcal{P}_\xi^*)$

6. Restringir la búsqueda a la mejor regla para cada clase de la partición restringida $P_\xi^* = \{C_i^\xi, C_j^\xi\}$ donde $C_i \neq C_j$ y $C_i^\xi, C_j^\xi \in \mathcal{P}_\xi^*$. Considerar los *subsistemas* de reglas:

$$\mathcal{S}_{C_i}(\mathcal{P}_\xi^*) = \{r_{s,c}^k : C = C_i \wedge r_{s,c}^k \in \mathcal{S}(\mathcal{P}_\xi^*)\} \text{ donde } \mathcal{S}_{C_i}(\mathcal{P}_\xi^*) \subseteq \mathcal{S}(\mathcal{P}_\xi^*)$$

y

$$\mathcal{S}_{C_j}(\mathcal{P}_\xi^*) = \{r_{s,c}^k : C = C_j \wedge r_{s,c}^k \in \mathcal{S}(\mathcal{P}_\xi^*)\} \text{ donde } \mathcal{S}_{C_j}(\mathcal{P}_\xi^*) \subseteq \mathcal{S}(\mathcal{P}_\xi^*)$$

7. Elegir el concepto ligado a la regla de mayor cobertura relativa de $\mathcal{S}_{C_i}(\mathcal{P}_\xi^*)$ y de $\mathcal{S}_{C_j}(\mathcal{P}_\xi^*)$:

Determinar:

- k_i, s_i tales que el concepto “ $X_{k_i} \in I_{s_i}^{k_i, \xi}$ ” tenga $p_{s_i c_i^\xi} = 1$ y $\text{CobR}(r_{s_i c_i^\xi}^{k_i}) = \max_{r \in \mathcal{S}_{C_i^\xi}(\mathcal{P}_\xi^*)} \{\text{CobR}(r)\}$ y
- k_j, s_j tales que el concepto “ $X_{k_j} \in I_{s_j}^{k_j, \xi}$ ” tenga $p_{s_j c_j^\xi} = 1$ y $\text{CobR}(r_{s_j c_j^\xi}^{k_j}) = \max_{r \in \mathcal{S}_{C_j^\xi}(\mathcal{P}_\xi^*)} \{\text{CobR}(r)\}$.

Sea,

$$\mathcal{K}_i = \{k \text{ tq } r_{s_i c_i^\xi}^{k_i} \text{ es una regla segura de cobertura máxima en } \mathcal{S}_{C_i}(\mathcal{P}_\xi^*)\}.$$

$$\mathcal{K}_j = \{k \text{ tq } r_{s_j c_j^\xi}^{k_j} \text{ es una regla segura de cobertura máxima en } \mathcal{S}_{C_j}(\mathcal{P}_\xi^*)\}.$$

$\mathcal{K} = \mathcal{K}_i \cap \mathcal{K}_j$ contiene los índices de las variables con máxima $\text{CobR}(r)$ para ambas clases simultáneamente.

8. Realizar una hipótesis fuerte de *Mundo cerrado* para describir la clase pareja en función del concepto complementario según la variable que induce el concepto sea local a una sola clase ($k_i \neq k_j$) o global ($k_i = k_j$).

9. Determinar los conceptos $A_i^{*\xi}$ y $A_j^{*\xi}$ inducidos para C_i^ξ y C_j^ξ de la siguiente forma:

$$\begin{aligned} \text{Sea } A_i^{\xi, k_i} &= “X_{k_i} \in I_{s_i}^{k_i, \xi}” \text{ y} \\ A_j^{\xi, k_j} &= “X_{k_j} \in I_{s_j}^{k_j, \xi}” \end{aligned}$$

(a) Si se identifica una única regla por clase y $k_i \neq k_j$:

$$A_i^{*\xi} = A_i^{\xi, k_i} \vee \neg A_j^{\xi, k_j} \quad (11.1)$$

$$A_j^{*\xi} = A_j^{\xi, k_j} \vee \neg A_i^{\xi, k_i} \quad (11.2)$$

- (b) Si se identifica una única regla por clase, $k_i = k_j$ y $CobR(r_{s_i, c_i}^{k_i}) > CobR(r_{s_j, c_j}^{k_j})$

$$A_i^{*\xi} = A_i^{\xi, k_i} \quad (11.3)$$

$$A_j^{*\xi} = \neg A_i^{\xi, k_i} \quad (11.4)$$

- (c) Si se identifica una única regla por clase, $k_i = k_j$ y $CobR(r_{s_i, c_i}^{k_i}) \leq CobR(r_{s_j, c_j}^{k_j})$

$$A_i^{*\xi} = \neg A_j^{\xi, k_j} \quad (11.5)$$

$$A_j^{*\xi} = A_j^{\xi, k_j} \quad (11.6)$$

- (d) Si $card\mathcal{K}_i > 1$ ó $card\mathcal{K}_j > 1$ y $\exists k \in \{\mathcal{K}_i \cup \mathcal{K}_j\}$ tq $r_{s,c}^k \in \mathcal{S}_{C_i}(\mathcal{P}_\xi^*)$ y $r_{s,c}^k \in \mathcal{S}_{C_j}(\mathcal{P}_\xi^*)$ con $CovR(r_{s,c}^k) = 100\%$ entonces X_k es totalmente caracterizadora y ni en \mathcal{K}_i ni en \mathcal{K}_j habrá ningún elemento con variables parcialmente caracterizadoras. En este caso $\mathcal{K}_i = \mathcal{K}_j = \mathcal{K}_i \cap \mathcal{K}_j$. Nunca se dará el caso $CobR(r_{s_i, c_i}^{k_i}) > CobR(r_{s_j, c_j}^{k_j})$ ó $CobR(r_{s_i, c_i}^{k_i}) < CobR(r_{s_j, c_j}^{k_j})$ porque las variables totalmente caracterizadoras siempre tienen $CovR(r_{s,c}^k) = 100\%$ en ambas clases, con lo cual se elegirá aleatoriamente a $r_{s_i, c_i}^{k_i}$ ó a $r_{s_j, c_j}^{k_j}$ (cuálquiera de las 2).

De existir variable totalmente caracterizadoras, sólo los conceptos inducidos por estas variables se utilizarán en la construcción del concepto compuesto. Se construye de la siguiente forma:

$$A_i^{*\xi} = (\bigwedge_{\forall k_i \in \mathcal{K}} A_i^{\xi, k_i}) \quad (11.7)$$

y

$$A_j^{*\xi} = (\bigwedge_{\forall k_i \in \mathcal{K}} \neg A_i^{\xi, k_i}) \quad (11.8)$$

- (e) Si $card\mathcal{K}_i > 1$ ó $card\mathcal{K}_j > 1$ y $\nexists k \in \{\mathcal{K}_i \cup \mathcal{K}_j\}$ tal que X_k sea totalmente caracterizadora de \mathcal{P}_ξ^* , entonces \mathcal{K}_i y \mathcal{K}_j sólo contiene variables parcialmente caracterizadoras. Si X_k es parcialmente caracterizadora dará lugar a reglas $p_{sc} = 1$ y $CovR(r_{s,c}^k) < 100\%$. En este caso el concepto se construye de la siguiente forma:

- Si $\mathcal{K} = \emptyset$

$$A_i^{*\xi} = (\bigvee_{\forall k_i \in \mathcal{K}_i} A_i^{\xi, k_i}) \vee (\bigvee_{\forall k_j \in \mathcal{K}_j} \neg A_j^{\xi, k_j}) \quad (11.9)$$

y

$$A_j^{*\xi} = (\bigvee_{\forall k_j \in \mathcal{K}_j} A_j^{\xi, k_j}) \vee (\bigvee_{\forall k_i \in \mathcal{K}_i} \neg A_i^{\xi, k_i}) \quad (11.10)$$

- Si $\mathcal{K} \neq \emptyset$

$$\begin{aligned} A_i^{*\xi} = & (\bigvee_{\forall k_i \in \mathcal{K}_i, k_i \notin \mathcal{K}} A_i^{\xi, k_i}) \vee (\bigvee_{\forall k_j \in \mathcal{K}_j, k_j \notin \mathcal{K}} \neg A_j^{\xi, k_j}) \vee \\ & (\bigvee_{\substack{\forall k_i \in \mathcal{K}, \\ CobR(r_{s_i, c_i}^{k_i}) > CobR(r_{s_j, c_j}^{k_j})}} A_i^{\xi, k_i}) \vee (\bigvee_{\substack{\forall k_j \in \mathcal{K}, \\ CobR(r_{s_i, c_i}^{k_i}) < CobR(r_{s_j, c_j}^{k_j})}} \neg A_j^{\xi, k_j}) \end{aligned} \quad (11.11)$$

y

$$A_j^{*\xi} = \left(\bigvee_{\forall k_j \in \mathcal{K}_j, k_j \notin \mathcal{K}_i} A_j^{\xi, k_j} \right) \vee \left(\bigvee_{\forall k_i \in \mathcal{K}_i, k_i \notin \mathcal{K}_j} \neg A_i^{\xi, k_i} \right) \vee \\ \left(\bigvee_{\substack{\forall k_j \in \mathcal{K}, \\ CobR(r_{s_j, c_j}^{k_j}) > CobR(r_{s_i, c_i}^{k_i})}} A_j^{\xi, k_j} \right) \vee \left(\bigvee_{\substack{\forall k_i \in \mathcal{K}, \\ CobR(r_{s_j, c_j}^{k_j}) < CobR(r_{s_i, c_i}^{k_i})}} \neg A_i^{\xi, k_i} \right) \quad (11.12)$$

10. Integrar $A_i^{*\xi}$ y $A_j^{*\xi}$ con el conocimiento de la(s) iteración(es) anterior(es) determinando los conceptos finalmente asociados a los elementos de \mathcal{P}_ξ . Los *conceptos compuestos* para las 2 clases que se separan de \mathcal{P}_ξ^* serán:

- Si $\xi = 2$

$$A_i^2 = A_i^{*2}$$

$$A_j^2 = A_j^{*2}$$

- Si $\xi > 2$

$$A_i^\xi = A_t^{\xi-1} \wedge A_i^{*\xi} \quad (11.13)$$

$$A_j^\xi = A_t^{\xi-1} \wedge A_j^{*\xi} \quad (11.14)$$

11. Construir el sistema de conceptos

$$\begin{aligned} \mathcal{A}_{\mathcal{P}_\xi} = \mathcal{A}_{\mathcal{P}_{\xi-1}} \setminus \{C_t : \mathcal{A}_t\} \\ \cup \quad \{C_i^\xi : A_i^\xi, \\ C_j^\xi : A_j^\xi \\ \}\end{aligned}$$

Asociado a la base de conocimiento:

$$\begin{aligned} \mathbb{R}(\mathcal{P}_\xi) = \mathbb{R}(\mathcal{P}_{\xi-1}) \setminus \{r_t\} \\ \cup \quad \{r_i : A_i^\xi \xrightarrow{p_{sci}} C_i^\xi, \\ r_j : A_j^\xi \xrightarrow{p_{scj}} C_j^\xi \\ \}\end{aligned}$$

12. Bajar un nivel en el árbol a analizar.

Considerar $\xi = \xi + 1$. Necesariamente $\mathcal{P}_{\xi+1}$ está anidada en \mathcal{P}_ξ , es decir que las dos nuevas clases se desprenden de una y sólo una de las dos clases generadas en la partición anterior.

$$\begin{aligned} \exists C_t^\xi \in \mathcal{P}_\xi \text{ y } C_t^\xi \notin \mathcal{P}_{\xi+1} \text{ y} \\ \exists i, j \text{ tq } C_i^{\xi+1} \in \mathcal{P}_{\xi+1} \text{ y } C_j^{\xi+1} \in \mathcal{P}_{\xi+1} \text{ y } C_i^{\xi+1} \cup C_j^{\xi+1} = C_t^\xi \\ \text{El resto de clases se mantienen de } \mathcal{P}_\xi \text{ a } \mathcal{P}_{\xi+1}\end{aligned}$$

Sean $C_i^{\xi+1}$ y $C_j^{\xi+1}$ las clases de $\mathcal{P}_{\xi+1}$ que subdividen una clase C_t^ξ de \mathcal{P}_ξ y $C_q^{\xi+1}$, la (o las) que ya estaba(n). Puesto que en el paso anterior hemos separado conceptualmente $C_i^{\xi+1} \cup C_j^{\xi+1}$ de $C_q^{\xi+1}$, en este punto queda solamente separar $C_i^{\xi+1}$ de $C_j^{\xi+1}$.

13. Volver al **paso 3** y repetir hasta obtener el número de clases deseado (previamente conocido, $\mathcal{P}_\xi = \mathcal{P}$).
14. Finalmente,

$$\mathbb{R}(\mathcal{P}_\xi) = \{r \ t q \ r : A \xrightarrow{p(r)} C \ \forall C \in \mathcal{P}_\xi\}$$

y

$$\mathcal{A}_{\mathcal{P}_\xi} = \{C : \mathcal{A}_C \ \forall C \in \mathcal{P}_\xi\}$$

Parte III

Aplicación y Resultados

Capítulo 12

Introducción

En esta parte se presenta la aplicación de la metodología CCCS a la interpretación de situaciones características en estaciones depuradoras de aguas residuales (EDAR), entorno en el cual la extracción automática de conocimiento tiene un gran interés como apoyo a la toma de decisiones en los procesos de control y supervisión de la planta.

Como ya hemos dicho los objetivos están encaminados a obtener, de un modo u otro, una buena caracterización que permita interpretar las situaciones que se presentan en la EDAR.

En primer lugar en el Capítulo §13 se presenta una descripción general del dominio de estudio y aplicación.

Para los dos casos de estudio, introducidos en los capítulos §14 y §19, se realiza un análisis descriptivo de los datos; el clustering, utilizando la clasificación basada en reglas, incorporando el conocimiento proporcionado por los expertos para obtener la base de conocimiento utilizada en la clasificación y en la interpretación final de la partición objetivo; y por último, se aplica la propuesta metodológica (metodología CCCS) presentada en los capítulos anteriores.

Finalmente a partir de cada base de datos y de la partición de referencia previa de los mismos, se genera una interpretación de las clases usando las variables recomendadas por el experto. Este sistema permitirá generar las caracterización e interpretación conceptual correspondiente a esa clase.

12.1 Acerca de las bases de datos

Como ya se ha dicho se analizan 2 casos de estudio.

EDAR Catalana. En primer lugar una base de datos proveniente de una planta depuradora situada en la costa catalana. Este caso de estudio es con el que se ha trabajado desde el inicio de esta investigación y con el que se han realizado diversas publicaciones hasta integrar el caso de estudio 2. Este caso es, a su vez, el que se utiliza dentro del grupo de investigación como batería de pruebas para las diferentes proyectos que se desarrollan, es por este motivo que se ha decidido presentarlo en primer lugar.

Los datos obedecen a la colaboración existente con la Universidad de Girona y en particular con el Doctor Ignaci Rodriguez Roda y el grupo de investigación LEQUIA, quien realiza estudios con esta EDAR desde hace años.

Se analiza una muestra de 396 observaciones. Estos datos fueron obtenidos en un período de un año y un mes; desde el 1 de Setiembre de 1995 al 30 de Septiembre de 1996, correspondiendo a mediciones medias de cada día.

En el capítulo §14 se presenta el caso de estudio mediante una descripción detallada del funcionamiento de esta planta; en el capítulo §15, el análisis estadístico de los datos

tanto univariante como bivariante (ver detalles en el Anexo C).

EDAR Eslovena. En segundo lugar la descripción de una base de datos proveniente de una planta depuradora de aguas residuales Eslovena. Esta base de datos se integra en el año 2006 producto de la colaboración existente con el Department of Systems and Control Jozef Stefan Institute de Ljubljana Eslovenia y en particular con el Doctor Darko Vrecko, quién realiza investigación con esta estación depuradora de aguas residuales.

La base de datos consta de 365 observaciones (medias diarias) que fueron tomadas desde el 1 de junio de 2005 de el 31 de mayo de 2006. Cada observación incluye mediciones de las 16 variables que son relevantes (según la opinión del experto) para el funcionamiento de la planta piloto.

La descripción detallada del funcionamiento de esta planta se puede ver en el capítulo §19 y la descriptiva estadística tanto univariante como bivariante en el capítulo §20 (ver detalles en el Anexo F).

12.2 Acerca del Clustering jerárquico

En este punto se siguen los siguientes pasos:

En primer lugar se trabaja con la clasificación ascendente jerárquica de los datos, tratando de hallar el clustering más adecuado. A continuación se hace uso del conocimiento proporcionado por los expertos en forma de reglas para realizar una o varias clasificaciones *basada en reglas* (detalles en Capítulo 3, sección §3.5.3) esperando mejorar los resultados. En esta tesis solo se presenta el mejor clustering para cada caso de estudio.

En cuanto a las especificaciones técnicas de cada clustering jerárquico:

- Los valores faltantes son sustituidos automáticamente por KLASS (ver Capítulo 7 y Anexo A): en el caso de variables numéricas por la media, en el caso de variables categóricas por un vector que representa las modalidades de la variable y sus proporciones, según se define y justifica en (Gibert 1996b).
- La métrica es Euclídea ascendente jerárquica clásica Normalizada (Diday and Moreau 1984), al trabajar con variables numéricas.
- El criterio de agregación¹ es el de Ward (Ward 1963).
- No se considera ponderación de los objetos ni de las variables. Algunos de estos aspectos de la clasificación ya han sido descritos en el Capítulo §5.
- A los parámetros de entrada para la clasificación que se mencionan anteriormente (métrica, agregación, etc) agregamos un tipo de *Ponderación Global* referente a la integración de las clases inducidas por las reglas utilizadas por la clasificación basada en reglas, a la jerarquía global.

Para cada clustering jerárquico se dispone de:

- Su histograma de índices de nivel y de su árbol jerárquico (dendrograma), ver detalles en sección §5.3.

¹Es aquel que selecciona en cada paso de la clasificación la fusión de la pareja de puntos, de individuos o subclases, que produzcan la mínima perdida de inercia (ligada a la cantidad de información) entre clases. (Volle 1985)

- Además KLASS da una lista de cortes recomendados, obtenidos por un criterio heurístico que tiene en cuenta la relación entre la variabilidad intra y entre clases.
- Una secuencia de particiones y la partición objetivo con su respectivo análisis descriptivo por clases para cada partición.

Para el Caso de estudio 1, EDAR Cataluña, el mejor clustering puede verse en detalle en publicaciones previas (Pérez-Bonilla 2005) y (Gibert and Pérez-Bonilla 2004a). En el Capítulo 16 de esta tesis se presenta el clustering seleccionado así como los cortes realizados de acuerdo a los criterios proporcionados por los expertos y la partición final que se desea interpretar.

Para el Caso de estudio 2, EDAR Eslovenia, el mejor clustering puede verse en detalle en publicaciones previas (Pérez-Bonilla, Gibert, and Vrecko 2007b) y en (Pérez-Bonilla, Gibert, and Vrecko 2007a). En el Capítulo 21 de esta tesis se presenta el clustering utilizado para probar la metodología CCCS, así como, los cortes realizados de acuerdo a los criterios proporcionados por los expertos para la partición final que se desea interpretar.

Se ha creído oportuno, presentar la descriptiva de las variables para cada una de las clases, mediante histogramas, boxplot múltiples y los estadísticos básicos clásicos. En los anexos D y G aparece el informe automático proporcionado por *Java-KLASS* para cada una de las clasificaciones que aparecen mencionadas en cada uno de los casos de estudio.

12.3 Acerca de la aplicación de la metodología CCCS

Como ya se ha explicado en capítulos anteriores (ver capítulo §10) existen 5 propuestas diferentes para combinar los resultados inducidos a partir del sistema de reglas $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ con los inducidos de la iteración anterior, a partir de $\mathcal{S}(\mathcal{P}_\xi^*)$. Se trata de alternativas, que utilizando la metodología de caracterización conceptual por Condicionamientos sucesivos CCCS, permitirá construir la regla final que asocia los conceptos a cada una de las clases de la partición que se quiere interpretar.

Como ya hemos mencionado en detalle en el Capítulo 10, la propuesta *Best Global concept and Close-World Assumption*, ver sección §10.2, consiste en elegir la regla segura de mayor cobertura relativa de todas las del sistema de reglas $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$. Tomar esta regla haciendo *CWA* y utilizar su negación para la conceptualizar de la clase complementaria.

La propuesta *Best local concept and no Close-World Assumption*, ver sección §10.3, consiste en elegir de entre todas las reglas de $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ que van a una misma clase, la de mayor cobertura relativa y no realizar la hipótesis de *CWA*. Así, no se utiliza la negación del concepto elegido para definir el concepto de la clase complementaria, si no, que para cada clase se utiliza el concepto seguro de mayor cobertura relativa.

La *Best local concept and Close-World Assumption*, ver sección §10.4, en este caso, consiste en elegir de entre todas las reglas de $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ que van a una misma clase, la de mayor cobertura relativa y realizar una hipótesis fuerte de *Mundo cerrado CWA*. Así, para cada clase, se utiliza la negación del concepto elegido para definir la otra clase en disyunción lógica (\vee) con el concepto seguro obtenido por la máxima cobertura relativa.

La *Best local concept and partial Close-World Assumption*, ver sección §10.5, consiste en elegir de entre todas las reglas de $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ que van a una misma clase, la de mayor cobertura relativa y realizar una hipótesis fuerte de *Mundo cerrado CWA* condicionado. Es decir, para cada clase, se utiliza la negación del concepto elegido para definir el concepto de la otra clase en disyunción lógica (\vee) con el concepto obtenido por la máxima cobertura relativa, pero cuando la variable referente a la mejor regla coincide en ambas clases (es decir las 2 reglas son inducidas por la misma variable) no se hace uso de la negación del concepto.

Finalmente la quinta propuesta *Best local-global concept and Close-World Assumption*, ver sección §10.6, que consiste en elegir de entre todas las reglas de $\mathcal{S}(\mathcal{P}_{\xi+1}^*)$ que van a una misma clase, la de mayor cobertura relativa (*Best local concept*) y realizar una hipótesis fuerte de *Mundo cerrado CWA* tomando la negación de la mejor regla cuando para ambas clases se tiene la misma variable. Así, para cada clase, se utiliza la negación del concepto elegido para definir el concepto de la otra clase en disyunción lógica (\vee) con el concepto seguro obtenido por la máxima cobertura relativa, pero cuando la mejor regla coincide en ambas clases (es decir las 2 reglas son inducidas por la misma variable) se elige de entre los 2 el que tiene la cobertura relativa mayor (*Best Global concept*), se trabaja con la negación de éste y se prescinde del otro concepto.

Como ya se ha dicho en la sección §12.1 se presenta en primer lugar el Caso de estudio de la EDAR Cataluña, aunque con esta base de datos sólo se comparan los criterios de integración del conocimiento que se consideran los mejores, a partir de las pruebas realizadas con todos los criterios utilizando el caso de estudio 2, EDAR Eslovenia.

Capítulo 13

Dominio de Aplicación: Estaciones depuradoras de aguas residuales

Antes de empezar con el análisis de datos es importante conocer cierta información sobre el dominio de aplicación en que intentamos trabajar y la importancia de éste, así como la metainformación que pueden tener las variables que entran en este estudio y los métodos y técnicas utilizados para llevar a cabo la aplicación de la Metodología de Caracterización Conceptual por Condicionamientos Sucesivos (CCCS).

13.1 Tratamiento de aguas residuales

En esta tesis trabajaremos el problema de las aguas residuales. Por lo tanto primeramente hemos de definir este término. En (Metcalf and Eddy 2003) se define como: “*Toda combinación de líquidos o aguas que transporten residuos procedentes de residencias, instalaciones públicas y centros comerciales e industrias, a las que, eventualmente, se pueden agregar aguas subterráneas, superficiales y de lluvia*”.

Una vez definido el concepto es el momento de atacar el problema real: la fuerte explotación de recursos debida a la creciente industrialización del mundo, ha agravado el problema de las aguas residuales. Con el paso de los años, no solamente es mayor la cantidad de agua residual generada, sino que también ha aumentado su concentración, así como la cantidad de nuevos compuestos de síntesis que contiene (se estima que hasta 10.000 nuevos productos aparecen cada año).

La naturaleza presenta un complejo ecosistema que interrelaciona la totalidad de cambios que pueden afectar la vida de los seres vivos que la habitan, y que permite superar las pequeñas perturbaciones que modifican este equilibrio natural. Dentro de este ecosistema global, el ciclo hidrológico es el que garantiza la circulación y disponibilidad de las aguas (a parte de las reservas naturales, cada año llegan a la tierra en lluvia o nieve unos 113.000 billones de m^3 de agua, véase (Lean and Hinrichsen 1994)). El hombre, que antiguamente sólo era una pequeña etapa de este ciclo, se ha convertido el factor determinante, y la madre naturaleza se ha visto superada en su lucha por mantener unas mínimas condiciones que favorezcan el desarrollo de la vida.

Por ejemplo, el impresionante volumen de agua residual que producen concentraciones de poblaciones de millones de habitantes como Méjico D.F., Sao Paulo o Nueva York, o sus elevadas concentraciones en extrañas combinaciones de compuestos que caracterizan las corrientes de subproductos de determinados procesos químicos que sostienen la industria moderna, muestran con claridad que la condición humana no parece tener límites ni complejos a la hora de superar la capacidad de autodepuración que presenta el ecosistema, cada vez

más débil y susceptible.

Parece claro que el hombre es el principal responsable de la contaminación de los recursos hidráulicos, pero también es el único que puede actuar sobre estas fuentes contaminantes, para reducir el impacto de estos residuos hasta un límite asimilable por el entorno natural, y facilitando así el restablecimiento del equilibrio natural.

13.1.1 El agua residual: composición

Para poder cuantificar el nivel de contaminación de las aguas residuales y para su posterior normalización es necesario definir una clasificación principal de los diferentes parámetros que intervienen en el proceso de depuración de éstas. Así pues, dividimos estos parámetros en físicos (como podrían ser la temperatura, el color, el olor o la turbiedad) y químicos (como sólidos, materia orgánica, nutrientes, pH, alcalinidad, dureza, cloruros y grasas).

Parámetros físicos

1. Temperatura: cambia en función de la estación del año, pero acostumbra a ser ligeramente superior a la del agua corriente. Tiene efectos sobre la actividad microbiana, la solubilidad de los gases, y la viscosidad.
2. Color: el agua residual presenta un color gris claro, pero se oscurece con el paso de los días o en condiciones de septicidad. Cualquier otro color que presente es producto de la presencia de determinados compuestos (tintes, sangre, cromo, residuos lácteos, etc.).
3. Olor: el agua residual fresca se caracteriza por un olor ligeramente desagradable que nos indica la presencia de aceites y detergentes. Pero cuando envejece aparecen olores fétidos resultantes de la descomposición de productos más complejos.
4. Turbiedad: falta de transparencia debida a la presencia de una amplia variedad de sólidos en suspensión contenida en el agua residual.

Parámetros químicos

1. Sólidos totales:
 - (a) Sedimentables: fracción de sólidos, orgánicos e inorgánicos, que sedimentan en 1 hora en un cono de Imhoff. Representa aproximadamente el barro que se puede eliminar en el tanque de sedimentación (ml/l).
 - (b) Suspensión (SST): fracción de sólidos, orgánicos e inorgánicos, que no se disuelven. Sólo se pueden eliminar por coagulación o filtración (mg/l). Éstos se dividen en fijos y volátiles (minerales y orgánicos) según sea la fracción no combustible o la combustible.
 - (c) Disueltos: fracción de sólidos, orgánicos e inorgánicos, que no es filtrable. Incluye todos aquellos sólidos menores a 1 milímicra. Éstos se dividen en fijos y volátiles (minerales y orgánicos) según sea la fracción no combustible o la combustible, igual que en el caso anterior.
2. Materia orgánica:
 - (a) Materia orgánica biodegradable: representa la fracción orgánica biodegradable presente en el agua residual, y es una medida del oxígeno disuelto que necesitan los microorganismos para consumir esta materia orgánica en 5 días y a 20°C (mg/l).

- (b) Materia orgánica degradable: medida de la fracción orgánica que es degradable por la acción de agentes químicos oxidantes (dicromato de potasio) bajo condiciones de acidez. También se mide por la cantidad estequiométrica de oxígeno disuelto requerido (mg/l).
- (c) Carbón orgánico total: contenido en la materia orgánica. Medido a través de la conversión de este carbón en CO₂ a altas temperaturas y en presencia de un catalizador (mg/l).

3. Nitrógeno total:

- (a) Nitrógeno orgánico: incluye el nitrógeno ligado a las proteínas, a los aminoácidos y a la urea (mg/l).
- (b) Amonio: primer producto de la descomposición del nitrógeno orgánico (mg/l).
- (c) Nitrógeno Kjeldahl: parámetro resultante de la suma de los dos anteriores, el amonio y el nitrógeno orgánico (mg/l).
- (d) Nitritos y nitratos: formas más oxidadas del nitrógeno (mg/l).

4. Fósforo total:

- (a) Orgánico: fracción de fósforo que se encuentra ligada a la materia orgánica (mg/l).
- (b) Inorgánica: fracción inorgánica del fósforo que existe como ortofosfatos y polifosfatos (mg/l).

5. pH: indicativo de la naturaleza básica o ácida del agua residual

- 6. Alcalinidad: debida a la presencia de iones bicarbonato, carbonato e hidróxido en el agua residual. Ésta provoca la resistencia a los cambios de pH (mg de CaCO/l).
- 7. Dureza: debida principalmente a los iones calcio y magnesio disueltos en el agua (mg de CaCO₃/l).
- 8. Cloruros: proporcionan mayor conductividad al agua y aumentan su densidad (mg/l).
- 9. Grasas: fracción de materia orgánica soluble en hexano. Incluye grasas y aceites de origen animal y vegetal (mg/l).

13.1.2 El problema de las aguas residuales

La composición del agua residual supone que su estancamiento pueda producir una serie de consecuencias que dificulten la vida en su entorno.

Por un lado, la elevada presencia de microorganismos patógenos (principalmente procedentes del aparato digestivo humano y animal) favorece la transmisión de enfermedades como la gastroenteritis, el cólera, la disentería, el tifus o la hepatitis A, fenómeno que todavía sucede en zonas subdesarrolladas, donde las precarias condiciones de vida y la nula canalización y tratamiento de sus aguas residuales, provoca constantes y peligrosas epidemias entre los niños (se estima que diariamente mueren hasta 25.000 personas a causa del uso de agua en mal estado, véase (Lean and Hinrichsen 1994)).

El alto contenido de materia orgánica del agua residual facilita la actividad microbiana, bien por vía anaerobia, con la consecuente aparición de olores desagradables, bien por vía aeróbica, hecho que implica una disminución del nivel de oxígeno disuelto en el agua y puede dificultar la vida acuática.

Los sólidos en suspensión de origen inorgánico no son tan problemáticos como los orgánicos, pero en grandes cantidades pueden provocar acumulaciones que dificulten o modifiquen el curso natural de las aguas naturales, con las consecuencias que de ello se pueden derivar.

Los nutrientes presentes en el agua residual facilitan el crecimiento de plantas acuáticas. Cuando este crecimiento es repentino y desmesurado provoca la muerte de peces y plantas.

Los pequeños rastros de metales o compuestos tóxicos, contenidos en el agua residual urbana, también llegar a ser letales si se permite un aumento significativo de su concentración por acumulación.

13.1.3 Descripción general del proceso de depuración

Las grandes áreas urbanas producen gran cantidad de aguas residuales y cuando el medio ambiente está contaminado y la calidad del agua empeora debido a que el proceso residual llega a superar el desempeño de la auto-regulación de las aguas recibidas. En este caso, se deben tomar ciertas medidas previsorias para restaurar el equilibrio del medio ambiente.

Las plantas depuradoras de aguas residuales proporcionan un importante equilibrio entre el medio ambiente y las aguas residuales concentradas de las áreas urbanas. Si estas últimas se liberan de forma descontrolada, se degradaría el medio ambiente, elemento esencial para el bienestar de los seres humanos.

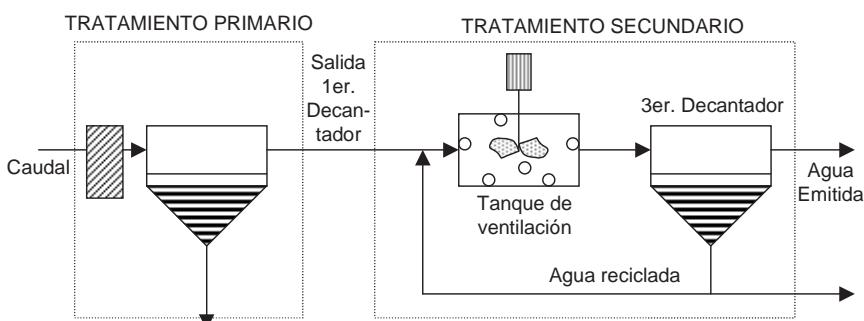


Figura 13.1: Diagrama típico del proceso de tratamiento de aguas residuales.

Para tratar adecuadamente las aguas residuales son necesarias distintas operaciones y procesos unitarios. El diagrama del proceso de una estación depuradora incluye diferentes combinaciones de agentes físicos, químicos y biológicos (ver proceso global en la Figura 13.1). Esta última representa un esquema típico, así como la secuencia lógica de tratamiento, dividida en diferentes fases, las que son resumidas brevemente a continuación (para mayor detalle referirse a (Sànchez-Marrè 1995) y (Rodríguez, D. 1999)).

Una primera etapa, llamada **pretratamiento**, que realiza una primera separación de los sólidos arrastrados por el agua residual que llega al colector. La finalidad del pretratamiento es evitar obstrucciones posteriores, así como eliminar el efecto abrasivo de estos materiales sobre mecanismos como las bombas y válvulas que se utilizan a lo largo del proceso. Esta operación física se acostumbra a realizar mediante una secuencia de rejillas, con diferente apertura y automatismo, pero también existe la posibilidad de incluir una trituradora que reduzca la medida de las partículas. La adición de un desarenador a continuación permite separar las tierras más finas y las grasas o aceites presentes, aprovechando la mayor velocidad de sedimentación de las primeras, y la flotación de las segundas, que se favorece con la aportación de un caudal controlado de aire y el diseño del desarenador.

Una segunda etapa donde el agua se deja reposar unas horas en un tanque de **sedimentación primaria** para que decante la materia orgánica sedimentable, así como el resto de arena o partículas inorgánicas que no han quedado retenidas en el pretratamiento. Los

sólidos sedimentados se envían a una línea de tratamiento específico, la línea de barros, siendo habitual su paso previo por un aparato que separe las pequeñas partículas inorgánicas contenidas. Cuando la carga es bastante elevada, el tiempo de retención es insuficiente, se puede complementar la decantación natural de la materia en suspensión con la adición de coagulantes químicos. Este tratamiento químico es casi obligatorio cuando el agua contenga metales o algún tóxico que pueda estropear el funcionamiento de algún elemento de la planta en fases posteriores.

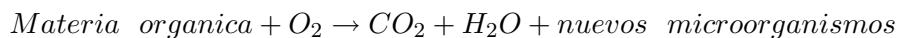
Seguidamente el agua llega a la etapa más importante del proceso. El fundamento de esta etapa no es otro que acelerar el proceso biológico que se daría en la naturaleza, es decir, la **degradación**, por parte de una población multiespecífica de microorganismos, de la materia orgánica disuelta en el agua residual. Esta reacción tiene lugar en unos bioreactores, fuertemente aireados en el caso que el proceso sea aeróbico. Más adelante detallaremos las configuraciones habituales, los mecanismos y la operación que posibilitan este proceso de depuración también llamado **tratamiento secundario**, que actualmente incluye ya la eliminación de nutrientes.

Siguiendo el camino que recorre el agua dentro de la depuradora, la última de las etapas habituales es una **nueva decantación a unos sedimentadores secundarios**. El objetivo es conseguir una buena separación entre el agua tratada y la biomasa (materia orgánica presente en el agua) presente. La zona superior del agua suele ser expulsada directamente; ya tenemos agua depurada que puede incorporarse al río.

Paralelamente, las dos fases de decantación generan una elevada cantidad de sólidos, llamados barros, que precisan un tratamiento específico para reducir su volumen, peso y características. Esta nueva secuencia de procesos se engloba en una nueva línea de tratamiento, la línea de barros, que suele constar de un espesado, una digestión, y una deshidratación final. Este proceso de detalla en la siguiente sección.

El sistema de los lodos activos

El proceso más habitual de tratamiento biológico de las aguas residuales (que también utiliza la Planta de Gerona), es el llamado sistema de lodos activos. Consiste en una oxidación bacteriana del residuo orgánico, seguido de una separación entre los sólidos en suspensión y el agua tratada. Simplificando el proceso, se puede considerar que los microorganismos utilizan el oxígeno presente en el agua para consumir el substrato o alimento, en este caso las moléculas orgánicas biodegradables contenidas en el agua residual. El resultado de este consumo sirve a los microorganismos para mantener sus funciones vitales, a la vez que genera una elevada producción de nuevos individuos. En cierta manera, lo que se consigue es transformar la fracción soluble de materia orgánica en una materia insoluble, lo cual facilita su posterior separación con una simple decantación. Resumiendo:



La mayor parte de los microorganismos separados en el decantador se devuelven al reactor biológico para mantener el nivel de depuración necesario, mientras que una pequeña fracción se aparta diariamente del sistema y se envía a la línea de barros, para evitar un aumento y envejecimiento excesivo de la biomasa presente en el sistema. Estas dos acciones, clave para garantizar el correcto desarrollo del proceso de lodos activos, se llaman recirculación y purga, ver Figura 13.2.

Dentro del mundo de las aguas residuales, existen multitud de nombres para referirse a los microorganismos responsables de la depuración. De este modo, se les puede llamar biomasa, sólidos, licor mezcla (cuando estamos hablando de la mezcla de agua y microorganismos del bioreactor), lodos o barros (cuando la concentración de microorganismos es elevada), y hasta

algunos ingenieros les han llamado bichos. Todos estos nombres se refieren a la población que conforma la micro-plantilla de la depuradora, ya que los microorganismos son los principales trabajadores y verdaderos responsables de la depuración.

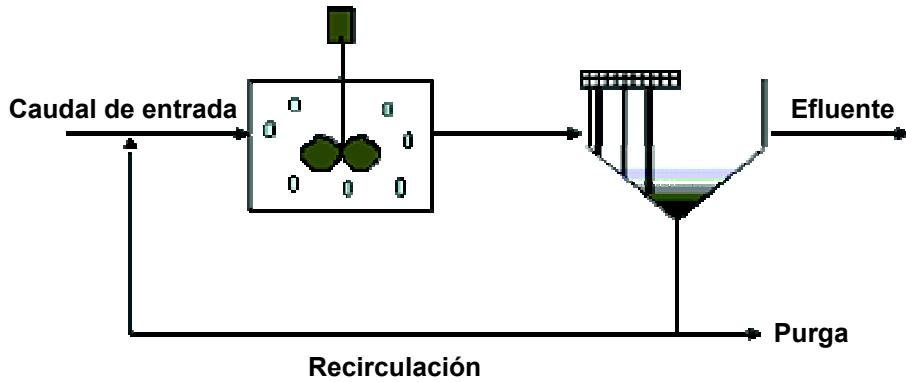


Figura 13.2: Diagrama típico del proceso de lodos activos.

La principal responsabilidad del jefe de planta es conducir este microuniverso que forman los lodos activos, controlando el ambiente en el que se ha de desarrollar la población de microorganismos. El objetivo es construir un ecosistema particular y estable, que elimine los contaminantes del agua, y que decante con facilidad.

Capítulo 14

Caso de Estudio 1: Planta catalana

14.1 Descripción general

Los datos provienen de una Planta Depuradora de aguas residuales de la costa de Catalunya. ver Figura 14.1.



Figura 14.1: Vista aérea de la EDAR Catalana.

Este sección presenta una descripción de la estructura general de esta planta concreta y los mecanismos que en ella se utilizan para detectar el estado de las aguas que se tratan, así como sus mecanismos de control.

Por un lado tenemos la línea de aguas y por otro la línea de barros. Los procesos vinculados a cada una de ellas se han presentado, de forma genérica, en el Capítulo 13, el la sección §13.1.3. La línea de aguas está constituida por una primera depuración previa a la red colectora, un pretratamiento con rejas de gruesos y estrechos, desarenador-desengrasador y un canal donde está instalado el medidor de caudal (canal-Parshall). Seguidamente la decantación primaria (volumen total de $4250\ m^3$) se efectúa en tres decantadores circulares. Viene favorecida con productos químicos para aumentar el rendimiento. La degradación, o digestión aeróbica, ($volumen_T = 8000m^3$) tiene lugar a continuación en tres balsas rectangulares aireadas mediante difusores de microburbujas. Por último la decantación secundaria

($volumen_T = 6000m^3$) formada también por tres decantadores que separan los barros activados del agua ya tratada, que se arroja finalmente al río. La Figura 14.2 esquematiza la estructura de la línea de aguas.

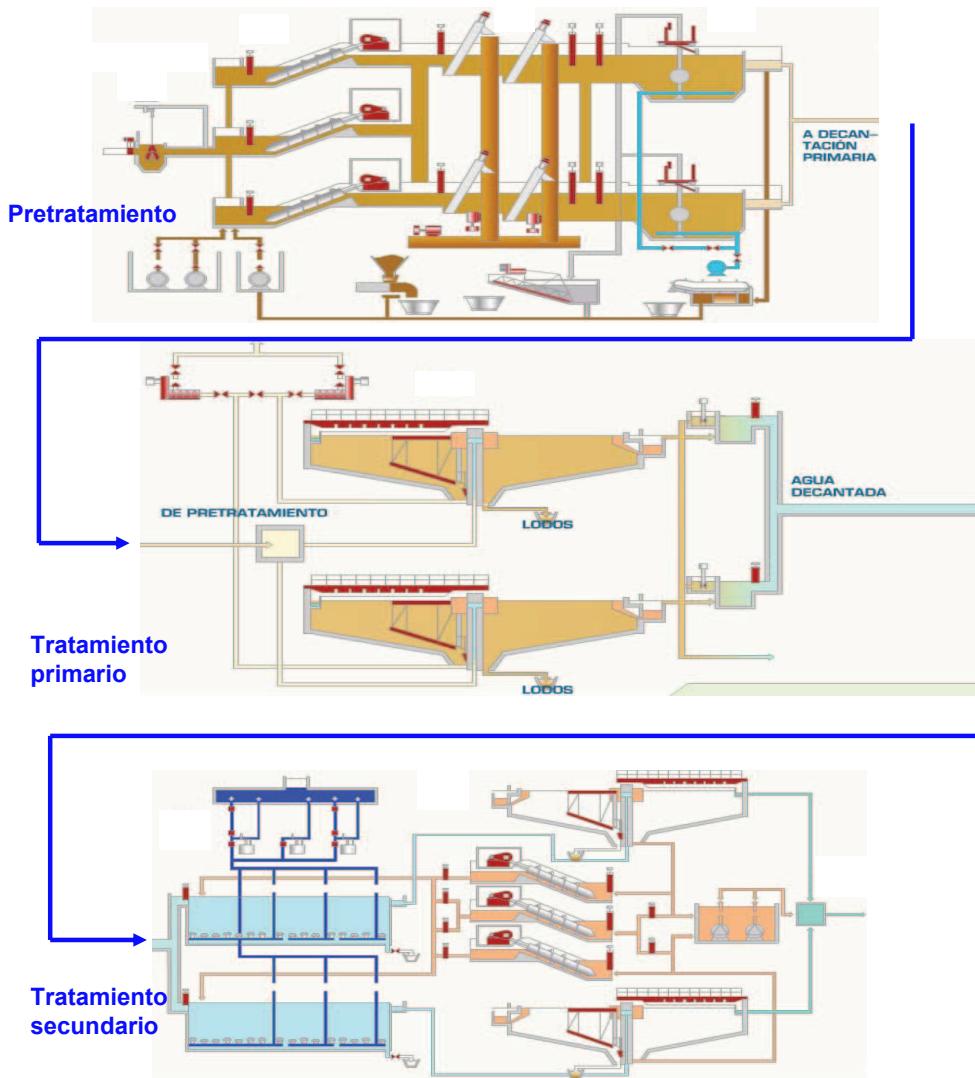


Figura 14.2: Línea de aguas de la planta.

Los barros del fondo de los decantadores primarios y la biomasa en exceso de los secundarios, se envían separadamente hacia la línea de barros, pasando los primeros por unos rollos rotativos y espesándose finalmente en dos espesadores circulares independientes ($volumen_T = 1200m^3$). Seguidamente el barro espeso entra en dos digestores anaeróbios ($volumen_T = 7000m^3$) donde, al mismo tiempo que se degrada la materia orgánica, se forma el biogás utilizado como combustible para mantener la temperatura necesaria en los digestores y para generar parte de la corriente eléctrica utilizada en la planta. Finalmente, los barros digeridos se les une polielectrolito y se pasan por tres filtros donde se elimina el exceso de agua. Todas las aguas excedentes de la línea de barros (muy cargadas), se retornan a la cabecera de la planta, donde se juntan con la corriente de entrada, y reinician el proceso de depuración. La Figura 14.3 representa este proceso.

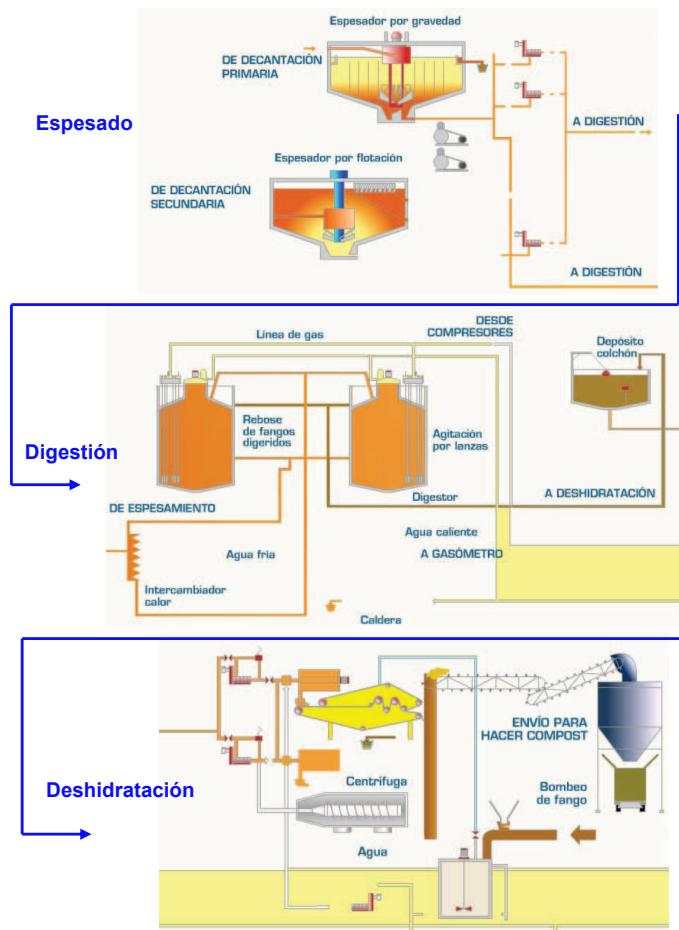


Figura 14.3: Línea de lodos de la planta.

14.2 Seguimiento de la Planta

Con la finalidad de caracterizar el funcionamiento de la Planta depuradora y realizar un seguimiento de ésta, nos interesa disponer del máximo de información y cuanto antes, mejor. Por este motivo desde el edificio de control de la planta se recogen constantemente una serie de medidas de variables como los caudales de entrada a la planta, el flujo en los reactores biológicos de la planta, el de recirculación de barros desde los decantadores secundarios hasta las balsas, el de purga (barros en exceso), el de biogás generado y el de aire añadido a las balsas, así como la conductividad del pH en diferentes puntos de la línea de aguas, la temperatura de los digestores, y finalmente información del estado de los motores y aparatos en general (aunque estos últimos datos no los trataremos en este proyecto y pueden ser objeto de futuros estudios).

Centrándonos en las variables de tipo analítico que afectan la línea de aguas, la estación depuradora de Gerona toma muestras, correspondientes a las últimas 24 horas, en tres puntos de la línea de las aguas (el influente, la corriente de salida de los decantadores primarios o entrada en los bioreactores y el efluente de la estación depuradora como se puede observar en la Figura 14.3) y analizan aquellos parámetros que le permiten realizar un seguimiento más minucioso. La ley determina la prioridad sobre el nivel de sólidos en suspensión y de materia orgánica en el efluente (véase sección §14.4), fijando unos límites de vertido y unos porcentajes de eliminación. Este hecho implica la medida rutinaria de SS (sólidos en suspensión), SSV (sólidos en suspensión volátiles), DQO (materia orgánica químicamente degradable) y DBO

(materia orgánica biodegradable) en los tres puntos citados anteriormente. Aunque la Planta no está obligada a eliminar los nutrientes de su agua, se realiza un seguimiento para saber cuáles son los niveles habituales: NKT (nitrógeno orgánico + amoniaco), NH₄⁺ y fósforo total son analizados tres veces por semana en los tres puntos habituales de recogida de muestras de la línea de aguas.

Con el objetivo de conocer el estado en que operan los bioreactores y qué tipo de ecosistema se está desarrollando, también se analiza el agua que éstos contienen: SSLM (sólidos en suspensión en el licor mezcla), SSVLM (sólidos en suspensión volátiles en el licor mezcla), y V30 (para establecer la sedimentabilidad de la biomasa). También se realizan observaciones microscópicas diarias, a través de las cuales se determina la biodiversidad, la especie predominante, la abundancia relativa de microflagelados y de bacterios filamentosos y el Índice Biótico de Madoni, aunque estas últimas variables no serán tratadas en el estudio actual.

En la línea de barros se establecen tres puntos básicos de caracterización analítica, los espesadores, los digestores anaerobios y el correspondiente a la salida de los filtros banda. En ellos se analiza el pH, ST (sólidos totales) y STV (sólidos totales volátiles). Dentro de los digestores también se estudia la relación entre la acidez volátil y la alcalinidad total, y en cuanto al barro resultante se mira la sequedad y los volátiles.

14.3 Control de la Planta

Aunque hemos visto la amplia caracterización analítica que se lleva en esta Planta, las posibilidades de actuación no son excesivas. Actúan dos sistemas automáticos de control (uno controla el oxígeno disuelto de las balsas biológicas, y otro la temperatura del interior de los digestores anaerobios), mientras que el resto de acciones son tomadas por el jefe de planta, que es quien, en función del estado inferido del estudio de los datos analíticos y del panel de control, decide el caudal de purga y la dosis necesaria de reactivos químicos a añadir. El diseño hidráulico de la arqueta de recirculación ya establece un porcentaje próximo al 100% del caudal recirculado respecto al influente de la planta.

La historia, la heurística y la bibliografía, han llevado a la Planta a fijar unos puntos de consigna que se mantienen constantes siempre que la Planta no sufra ningún tipo de situación imprevista. Estas consignas consisten en mantener el nivel de oxígeno disuelto alrededor de 2 mg/L en las balsas de activación (en el verano esta consigna baja hasta el 1.5 mg/L), el caudal de recirculación próximo al 100% respecto al de entrada (limitado por by-pass hasta los 39.000 m³/d), y el caudal de purga suficiente para mantener edades celulares entre los 4 y los 7 días (aproximadamente unos 500 m³/d). Otras consignas de la planta establecen las dosis aproximadas de sales de hierro añadidas para ayudar a la decantación primaria (50 mg/L), así como la cantidad de calcio y polielectrolito añadidos al barro para favorecer la deshidratación de los lodos digeridos, entre otras cosas.

14.4 Legislación

El **Pla de Sanejament de la Generalitat de Catalunya**, basándose en la directiva del Consejo 91/271/CEE del Diario de las Comunidades Europeas (Rodríguez, D. 1999), insta a la depuración de los principales componentes del agua residual urbana que se arroje.

El cuadro principal de la directiva queda simplificado en la Tabla 14.1, en la que se especifican las cotas permitidas para las principales componentes del agua a la salida de una planta depuradora.

Se establece un límite de hasta 25 mg/l para el parámetro DBO (Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua)), con un por-

centaje mínimo de reducción del 70-90 %, y un límite de hasta 125 mg/l para el parámetro DQO (fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua)), con un porcentaje mínimo de reducción del 75 % mientras que para los SS (Sólidos en suspensión (mg de sólidos por litro de agua)) el límite queda fijado en 35 mg/l con un % mínimo de reducción del 90 %. Por lo que respecta a los vertidos en zonas sensibles (con riesgo de eutrofización), se suma un límite para los nutrientes establecido en 1 mg/l (y un 80 % de reducción) para el fósforo, y en 10 mg/l para el nitrógeno total (con 70-80 % mínimo de eliminación). Estos límites son ligeramente más permisibles en caso que la depuradora trate caudales pequeños (de entre 10.000 y 100.000 habitantes equivalentes).

Parámetro	Concentración	% de reducción ¹
DBO (mg O ₂ /l)	25	70 - 90 %
DQO (mg O ₂ /l)	125	75%
SS (mg/l)	35	90%
Fósforo total (mg-P/l)	2 [1 (>100,000 h.-e.)]	80%
Nitrógeno Total (mg-N/l)	15 [10 (>100,000 h.-e.)]	70 - 80%

Tabla 14.1: Cotas permitidas por la directiva del Consejo 91/271/CEE.

14.5 Presentación de los datos

Como ya se ha dicho anteriormente, los datos del estudio provienen de la una EDAR de catalunya y se deben a la colaboración existente, a raíz de un proyecto de investigación, entre dicha Planta y el equipo de Ingeniería del Conocimiento y Aprendizaje Automático del Departamento de LSI de la Universidad Politécnica de Cataluña, al cual está vinculada Karina Gibert, profesora del departamento de Estadística e Investigación Operativa de la UPC y directora de este proyecto. Se ha contado con todos ellos para el desarrollo del presente trabajo.

Se analiza una muestra de 396 observaciones. Estos datos fueron obtenidos en un periodo de un año y un mes; desde el 1 de Setiembre de 1995 al 30 de Setiembre de 1996, correspondiendo una medición a la media de cada día.

14.6 Descripción de las variables

La descripción de la Planta para cada día consiste en caracterizar el caudal de entrada, el estado de la mezcla después del primer decantador, el caudal de salida y el estado de la mezcla en el reactor biológico. Esta caracterización se hace utilizando principalmente medidas de volumen y resultados de análisis químicos y biológicos, ver Figura 14.4.

A continuación se detallan las variables que se han utilizado en la descripción de los datos. Se indica resumidamente el significado y las unidades de cada una de ellas.

- Variables de entrada:

- Q-E: Caudal de entrada (metros cúbicos de agua por día)
- FE-E: Pretratamiento con hierro (mg de hierro por litro de agua)
- PH-E: pH (unidades de pH)
- SS-E: Sólidos en suspensión (mg de sólidos por litro de agua)
- SSV-E: Sólidos volátiles en suspensión (mg de sólidos por litro de agua)

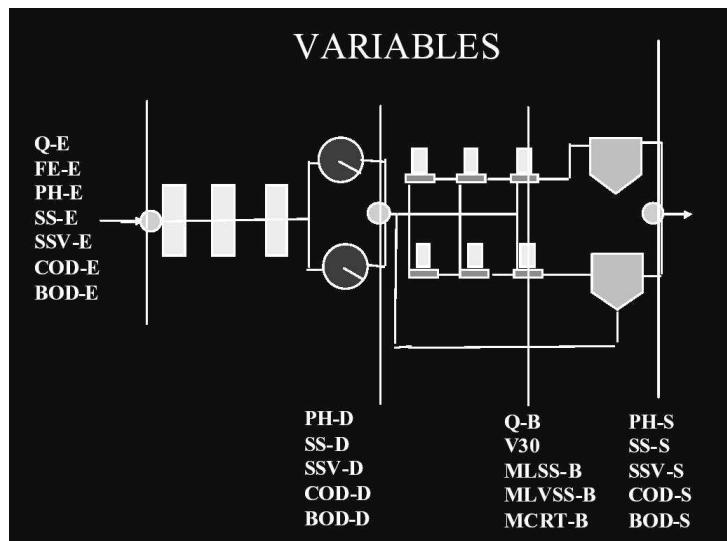


Figura 14.4: Caudal de entrada a la planta versus caudal de entrada al reactor biológico.

- DQO-E: Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua)
- DBO-E: Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua)
- NKT-E: Suma del amonio y el nitrógeno orgánico (mg de nitrógeno por litro de agua)
- NH4-E: Amonio (mg de nitrógeno por litro de agua)
- P-E: Fósforo (mg de fósforo por litro de agua)
- DBO/DQO-E: Cociente de materia orgánica biodegradable en agua residual (sin unidades)
- Variables después de la decantación:
 - PH-D: pH (unidades de pH)
 - SS-D: Sólidos en suspensión (mg de sólidos por litro de agua)
 - SSV-D: Sólidos volátiles en suspensión (mg de sólidos por litro de agua)
 - DQO-D: Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua)
 - DBO-D: Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua)
 - NKT-D: Suma del amonio y el nitrógeno orgánico (mg de nitrógeno por litro de agua)
 - NH4-D: Amonio (mg de nitrógeno por litro de agua)
 - DBO/DQO-D: Cociente de materia orgánica biodegradable en agua residual (sin unidades)
- Variables de salida:
 - PH-S: pH (unidades de pH)

- SS-S: Sólidos en suspensión (mg de sólidos por litro de agua)
- SSV-S: Sólidos volátiles en suspensión (mg de sólidos por litro de agua)
- DQO-S: Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua)
- DBO-S: Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua)
- NKT-S: Suma del amonio y el nitrógeno orgánico (mg de nitrógeno por litro de agua)
- NH4-S: Amonio (mg de nitrógeno por litro de agua)
- P-S: Fósforo (mg de fósforo por litro de agua)
- DBO/DQO-S: Cociente de materia orgánica biodegradable en agua residual (sin unidades)
- Variables del tratamiento biológico:
 - V30-B: Análisis volumétrico 30; calidad de sedimentación de la mezcla (ml por litro de agua)
 - MLSS-B: Sólidos en suspensión del licor mezcla (mg de sólidos por litro de licor mezcla)
 - MLVSS-B: Sólidos volátiles en suspensión del licor mezcla (mg de sólidos por litro de licor mezcla)
 - IM-B: Cociente entre V30 y MLSS-B (ml por gramo)
 - CM1-B: Fracción de materia orgánica degradable (kg) por acción de agentes químicos oxidantes por sólidos en suspensión (kg)
 - CM2-B: Fracción de materia orgánica biodegradable (kg) por sólidos en suspensión (kg)
 - MCRT-B: Edad celular (días)
 - QB-B: Caudal del reactor biológico (metros cúbicos de agua por día)
- Otras variables:
 - QR-G: Caudal de recirculación (metros cúbicos de agua por día)
 - QP-G: Caudal de la purga (metros cúbicos de agua por día)
 - QA-G: Afluencia de aire (metros cúbicos de aire por día)
 - TRH-C: Tiempo de resistencia hidráulico (horas)

De todas estas variables existe un subconjunto de 25 que, según los expertos, son las más informativas y con las que se trabajará en el proceso de clasificación (Gibert and Roda 2000). Éstas corresponden a: Q-E, QB-B, QR-G, QP-G, QA-G, FE-E, PH-E, SS-E, SSV-E, DQO-E, DBO-E, PH-D, SS-D, SSV-D, DQO-D, DBO-D, PH-S, SS-S, SSV-S, DQO-S, DBO-S, V30-B, MLSS-B, MLVSS-B y MCRT-B.

Capítulo 15

Análisis descriptivo de los datos planta catalana

15.1 Introducción

Una vez definidos los parámetros a medir procedemos a una descripción exhaustiva de cada una de las variables disponibles.

El análisis descriptivo nos permitirá, en primera aproximación, hacernos una idea de la composición de la muestra. Este análisis se divide en dos grandes bloques: un primer análisis descriptivo de cada una de las variables (análisis univariante) y un segundo análisis de ciertos parámetros, en algunos casos, con gráficos bivariantes de puntos (Plots), que nos permiten intuir la relación entre dos variables y el sentido de ésta, es decir, si es positivo o negativo (análisis bivariante). Este último análisis es interesante de hacer debido que hay variables que son medidas en cada uno de los tres puntos clave que podemos encontrar en el proceso de depuración (entrada, en el decantador, en el reactor biológico y a la salida de la Planta), ver Figura 15.1 y Figura 14.4.



Figura 15.1: Puntos de medición de variables.

15.2 Análisis univariante

El primer análisis presenta los estadísticos de descripción clásicos (número de valores, número de valores no missing, media, mediana, media truncada, desviación estándar, mínimo, máximo,

1º cuartil y 3º cuartil), un Boxplot y un Histograma, que muestra la distribución de la variable. Finalmente presentamos un resumen de los hallazgos más importantes derivados de este análisis. En este Capítulo sólo se presenta una variable a modo ilustrativo, las 24 variables restantes están en el Anexo C en la sección §C.1.

15.2.1 Q-E. Caudal de entrada (m³/d) Inflow wastewater

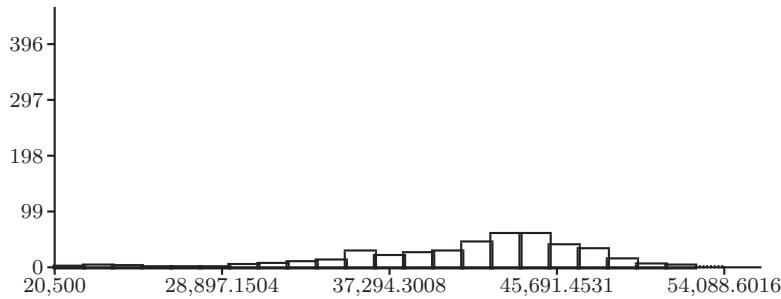


Figura 15.2: Histograma de la variable Q-E.

En el Histograma que se muestra en la Figura 15.2 se observa claramente como este grupo de valores queda al margen del comportamiento normal y cómo el periodo anteriormente comentado convierte la distribución en asimétrica aplanada a la izquierda. Tenemos un caudal de agua de 41.816 metros cúbicos por día en media con una desviación de 5.120 metros cúbicos y con unos valores que van de 20.500 a 54.089 metros cúbicos. Existe un dato missing para esta variable (dato nº: 122: 31-XI-95), aunque éste se repetirá en todas las variables.

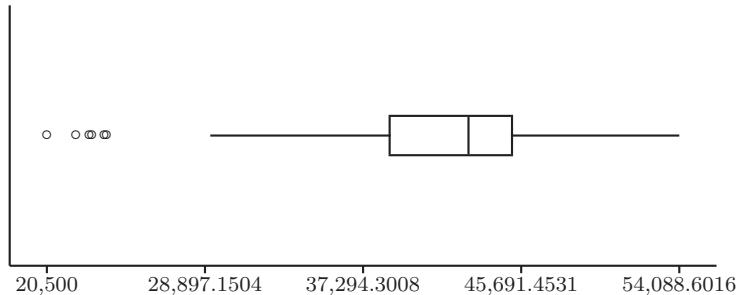


Figura 15.3: Boxplot de la variable Q-E.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	1
Número de observaciones útiles	395
Media	41,816
Mediana	42,892.1016
Primer quartil (Q1)	38,760
Tercer quartil (Q3)	45,147.3008
Mínimo	20,500
Máximo	54,088.6016
Variància	26,152,512
Desviación típica	5,113.9526
Quasi-desviación típica	5,120.4326
Coeficiente de variación	0.1223

15.3 Análisis Bivariante

El segundo análisis nos muestra la evolución de un conjunto de variables que se miden en los puntos clave que podemos encontrar en el proceso de depuración (entrada, después de la decantación y a la salida de la Planta, ver Figura 15.1). Ello contribuye a poner de manifiesto el efecto global del proceso de depuración de aguas residuales. Como en el caso deal análisis univariante sólo presentamos un caso a modo de ejemplo, en el Anexo C en la sección §C.2 se pueden ver otras relaciones entre pares de variables que se han estudiado.

15.3.1 Sólidos en suspensión (SS)

1. Sólidos en suspensión (SS) en la entrada y después del decantador.

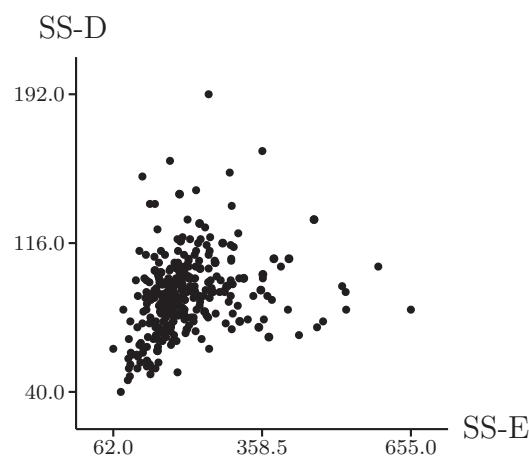


Figura 15.4: Diagrama bivariante para las variables SS-E y SS-D.

Debido al parecido de la variable SS-D con SS-E en cuanto a significado, podría hacernos pensar que existe gran correlación entre estas variables pero, gráficamente se observa que se produce una mayor variabilidad para valores elevados de ambas variables (heterocedasticidad), ver Figura 15.4.

En algunas situaciones, especialmente cuando entran cantidades elevadas de SS, la decantación reduce drásticamente los sólidos en suspensión, pero cuando el agua entra a la planta tiene cantidades bajas o moderadas de SS. En esa situación se reducen siempre los sólidos en suspensión pero existe mayor variabilidad en torno a los valores medios.

Capítulo 16

Clustering planta catalana

16.1 Introducción

Se tiene una familia de ficheros originales *depGI41-396*, estos ficheros son los que contienen la información para realizar la clasificación utilizando Java-KLASS y es la base de datos original de todas las variables medidas en la planta.

La estructura contiene un número total de 41 variables de las cuales 41 variables son numéricas y 0 variables son categóricas y un número total de 396 observaciones.

De todas estas 41 variables existe un subconjunto de 25 que, según los expertos, son las más informativas y son con las que se trabajará en el proceso de clasificación (Gibert and Roda 2000). Éstas corresponden a: Q-E, QB-B, QR-G, QP-G, QA-G, FE-E, PH-E, SS-E, SSV-E, DQO-E, DBO-E, PH-D, SS-D, SSV-D, DQO-D, DBO-D, PH-S, SS-S, SSV-S, DQO-S, DBO-S, V30-B, MLSS-B, MLVSS-B y MCRT-B.

Se considera un subconjunto de datos con esta selección de 25 variables que son las consideradas más relevantes por los expertos (ver la descripción de las variables en sección §14.6).

En este capítulo se presenta los detalles de la mejor clasificación de acuerdo a la recomendación y el conocimiento proporcionado por los expertos. Para más información sobre otras clasificaciones realizadas con esta base de datos ver (Rodríguez, D. 1999) y (Gibert and Pérez-Bonilla 2004a).

Los aspectos técnicos del clustering jerárquico se han presentado previamente en el capítulo §12, en la sección §12.2.

16.2 Base de conocimiento para la clasificación basada en reglas

La base de conocimiento proporcionada por los expertos y que recoge las limitaciones legales expuestas en la sección §14.4, se traduce en las siguientes reglas donde:

- $r_1 : ((\text{and} (\text{>} (\text{SS-S}) 20) (\text{>} (\text{DBO-S}) 35)) \rightarrow P)$
- $r_2 : ((\text{and} (\text{>} (\text{SS-S}) 20) (\text{<} (\text{DBO-S}) 35)) \rightarrow S)$
- $r_3 : ((\text{and} (\text{<} (\text{SS-S}) 20) (\text{>} (\text{DBO-S}) 35)) \rightarrow T)$

En la Figura 16.1 se muestran los *subárboles* inducidos por cada regla.

16.3 Clustering

Así se construye la siguiente familia de ficheros de clasificación que una vez clasificados dan origen a la partición objetivo que se desea interpretar.

16.3.1 Familia de ficheros de clasificación

Se construye una familia de clasificación necesaria para operar con el software Java-KLASS, que incluye 396 observaciones (una observación por día), 25 variables y la base de conocimiento proporcionada por experto para la clasificación basada en reglas, se puede ver la descripción de las reglas y mas detalles de ésta y otras clasificaciones realizadas con esta base de datos en el reporte (Gibert and Pérez-Bonilla 2004a).

1. Nombre de los ficheros de clasificación:
 - (a) depGI25-396.dat
 - (b) depGI25-396.obj
 - (c) depGI25-396.pro
 - (d) depGI25-396.reg
2. Estructura de los ficheros de clasificación:
 - (a) Número de variables utilizadas para clasificar = 25.
 - (b) Número de variables numéricas en la clasificación = 25.
 - (c) Número de variables categóricas en la clasificación = 0.
 - (d) Número de objetos en la clasificación = 396.
 - (e) Descripción de las reglas utilizadas, proporcionadas por el experto (se encuentran el fichero depGI25-396.reg) y se presentan en la sección §16.2.

16.3.2 Clasificación basada en reglas

1. Parámetros de entrada para la Clasificación
 - Métrica utilizada = Euclidea normalizada.
 - Criterio de Clasificación = Ward.
 - Ponderación de objetos = no.
 - Tipo de Ponderación = Global.
2. Resultados:
 - Nombre del Fichero de resultados .his: G1R1EnG.his
 - Porcentaje de missing en los datos = 0,101%.
 - Árbol de clasificación (o dendrograma) [$\tau_{Gi1,R1}^{En,G}$], ver Figura 16.2
 - Se recomienda cortar en (4 3 8 6 5) clases.
 - Partición inducida por las reglas:
 - Seleccionados como clase P = 40 objetos
 - Seleccionados como clase S = 11 objetos
 - Seleccionados como clase Q = 10 objetos

- Seleccionados como clase Residual = 335 objetos

En la Figura 16.1 se muestran los *subárboles* inducidos por cada regla.

3. Cortes realizados:

De acuerdo al conocimiento proporcionado por los expertos y al criterio heurístico que tiene en cuenta la relación entre la variabilidad intra y entre clases que implementa Java-KLASS la mejor partición es la que se obtiene al cortar el árbol en 4 clases, ver Figura 16.2, y es con la que se trabajará para generar la interpretación final utilizando la metodología CCCS.

A partir de lo anterior se procede a cortar el árbol en 4 clases, ver Figura 16.2, y de esta manera obtener la partición de referencia (partición objetivo):

$$\mathcal{P}4_{Gi1,R1}^{En,G} = \{Classer392, Classer389, Classer390, Classer383\}$$

Como ya se ha explicado en capítulos anteriores, la propuesta metodológica aprovecha la estructura jerárquica del clustering y por lo tanto también se presenta, en este capítulo, los cortes en 2 y 3 clases necesarios para aplicar la metodología CCCS.

Los ficheros que incluyen los 3 cortes realizados tienen los siguientes nombres:

- (a) En 2 clases: dani2p.cls (y .par), $P2_{Gi1,R1}^{En,G}$, ver Figura 16.3 y Figura 16.4.
- (b) En 3 clases: dani3p.cls (y .par), $P3_{Gi1,R1}^{En,G}$, ver Figura 16.5 y Figura 16.6.
- (c) En 4 clases: dani4p.cls (y .par), $P4_{Gi1,R1}^{En,G}$, ver Figuras 16.7, 16.8 y Figuras 16.9, 16.10.

16.3.3 Secuencia de particiones

Las particiones (cortes sucesivos) en 2, 3 y 4 clases a partir del dendrograma, ver Figura 16.2, son:

$$Classer394 \left\{ \begin{array}{l} Classer392 \\ Classer393 \left\{ \begin{array}{l} Classer389 \\ Classer391 \left\{ \begin{array}{l} Classer390 \\ Classer383 \end{array} \right. \end{array} \right. \end{array} \right.$$

Cada partición está compuesta de las siguientes clases:

$$\mathcal{P}2_{Gi1,R1}^{En,G} = \{Classer392, Classer393\}.$$

El análisis descriptivo por clases para esta partición se puede ver en la Figura 16.3.

$$\mathcal{P}3_{Gi1,R1}^{En,G} = \{Classer392, Classer389, Classer391\}.$$

El análisis descriptivo por clases para esta partición se puede ver en la Figura 16.5

$$\mathcal{P}4_{Gi1,R1}^{En,G} = \{Classer392, Classer389, Classer390, Classer383\}.$$

El análisis descriptivo por clases para esta partición se puede ver en las Figuras 16.7, 16.8, 16.9 y 16.10.

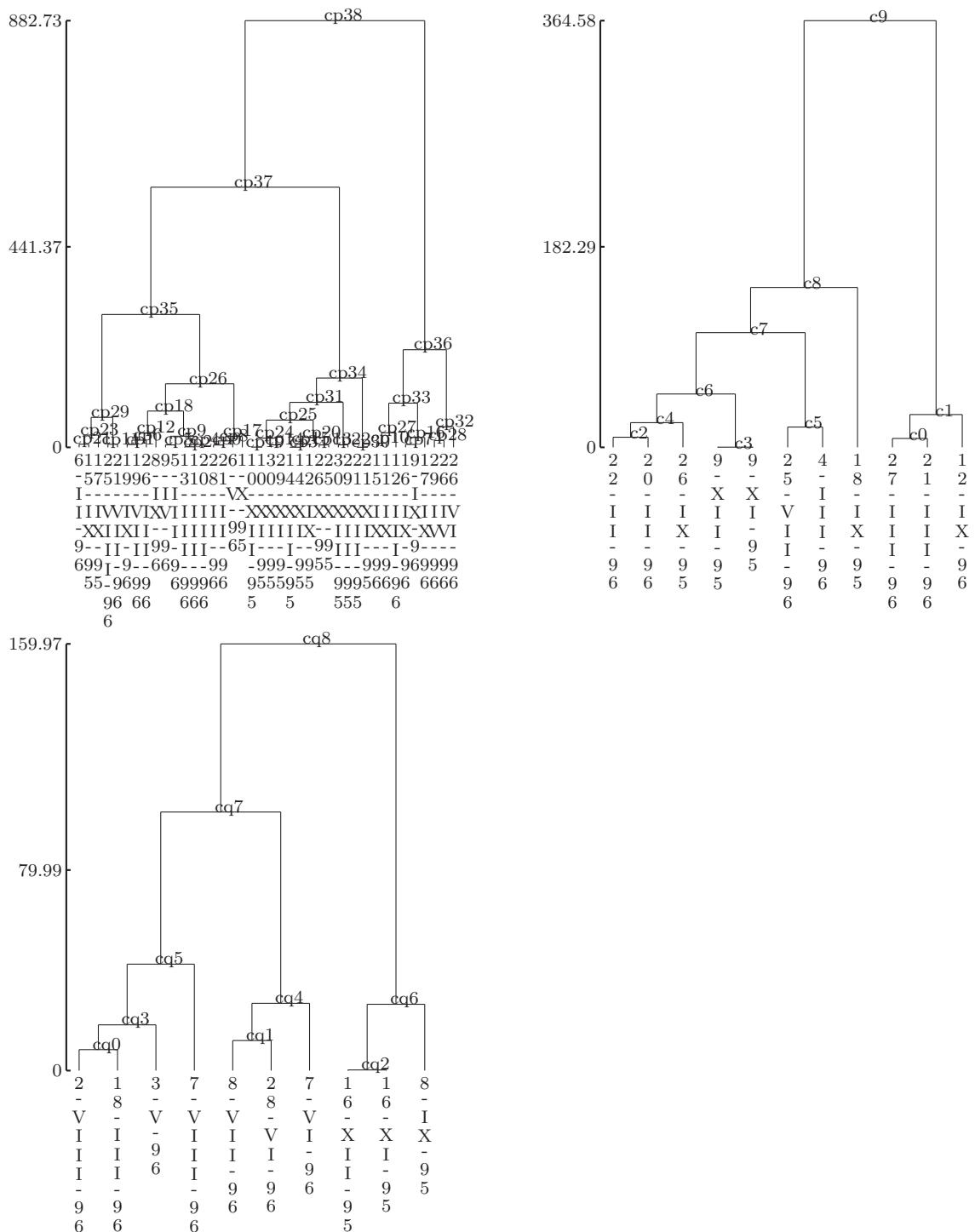


Figura 16.1: Árboles inducidos por las reglas de clasificación (P izq.-arriba, S der.-arriba y Q der.abajo).

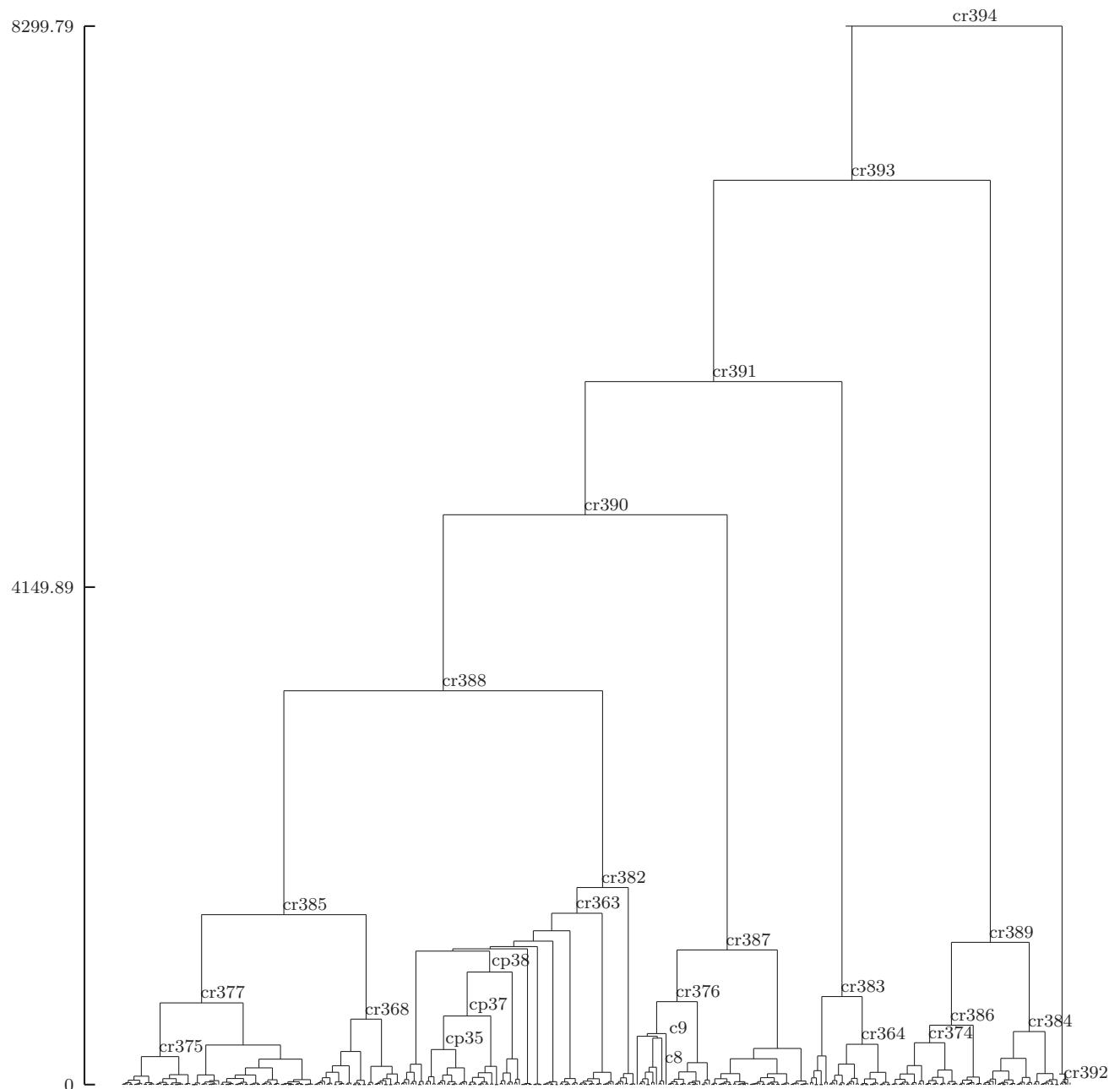
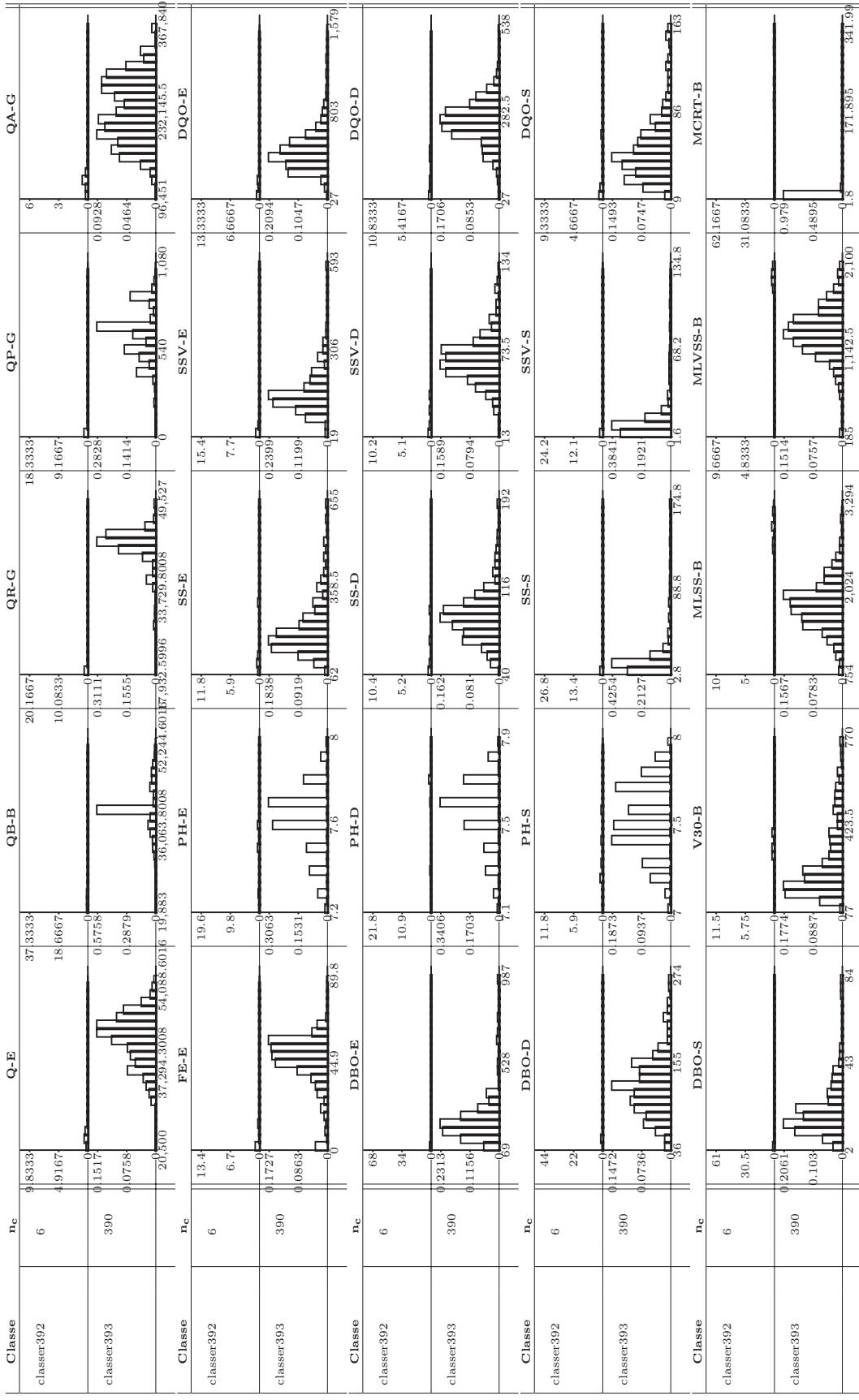
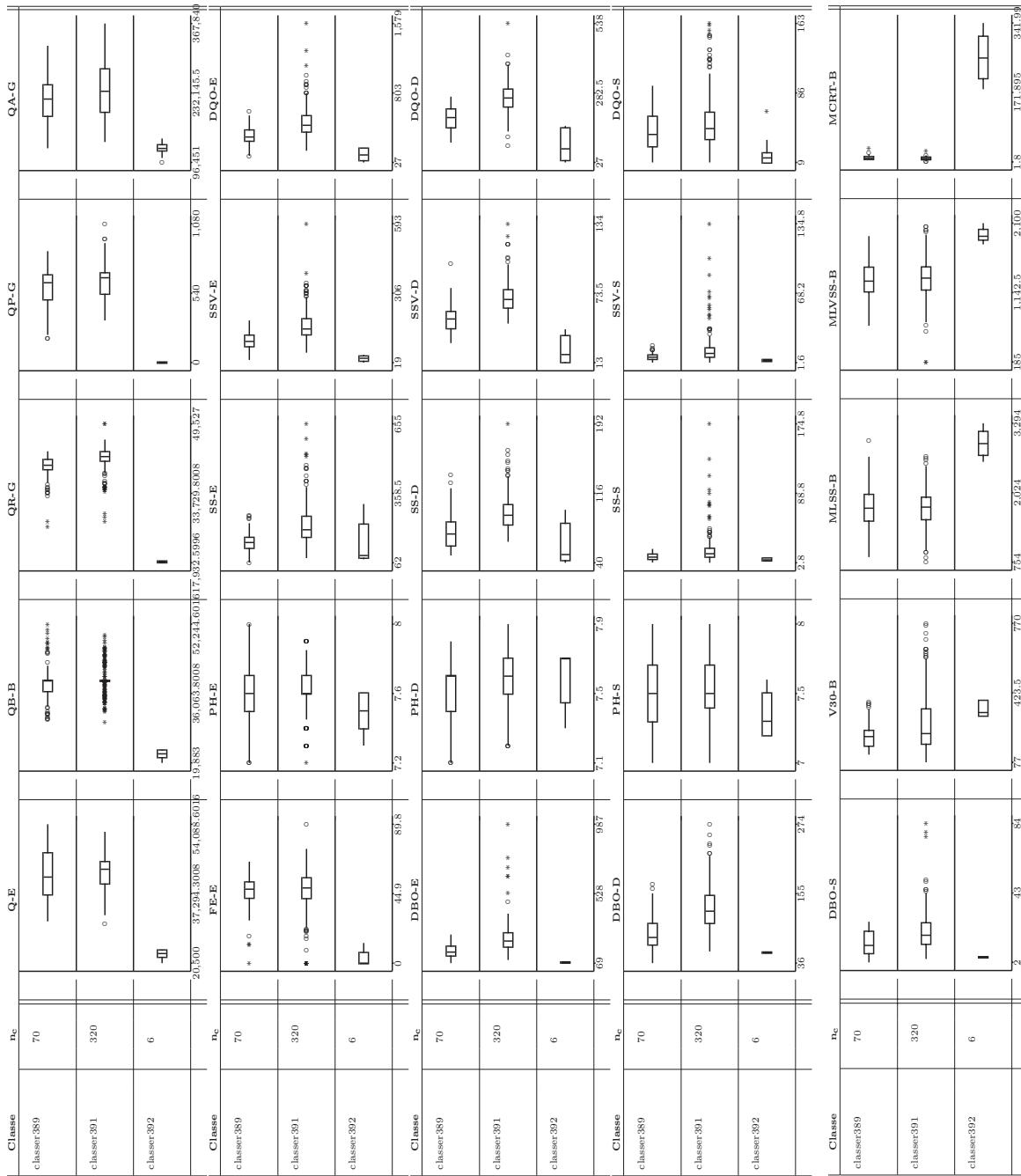


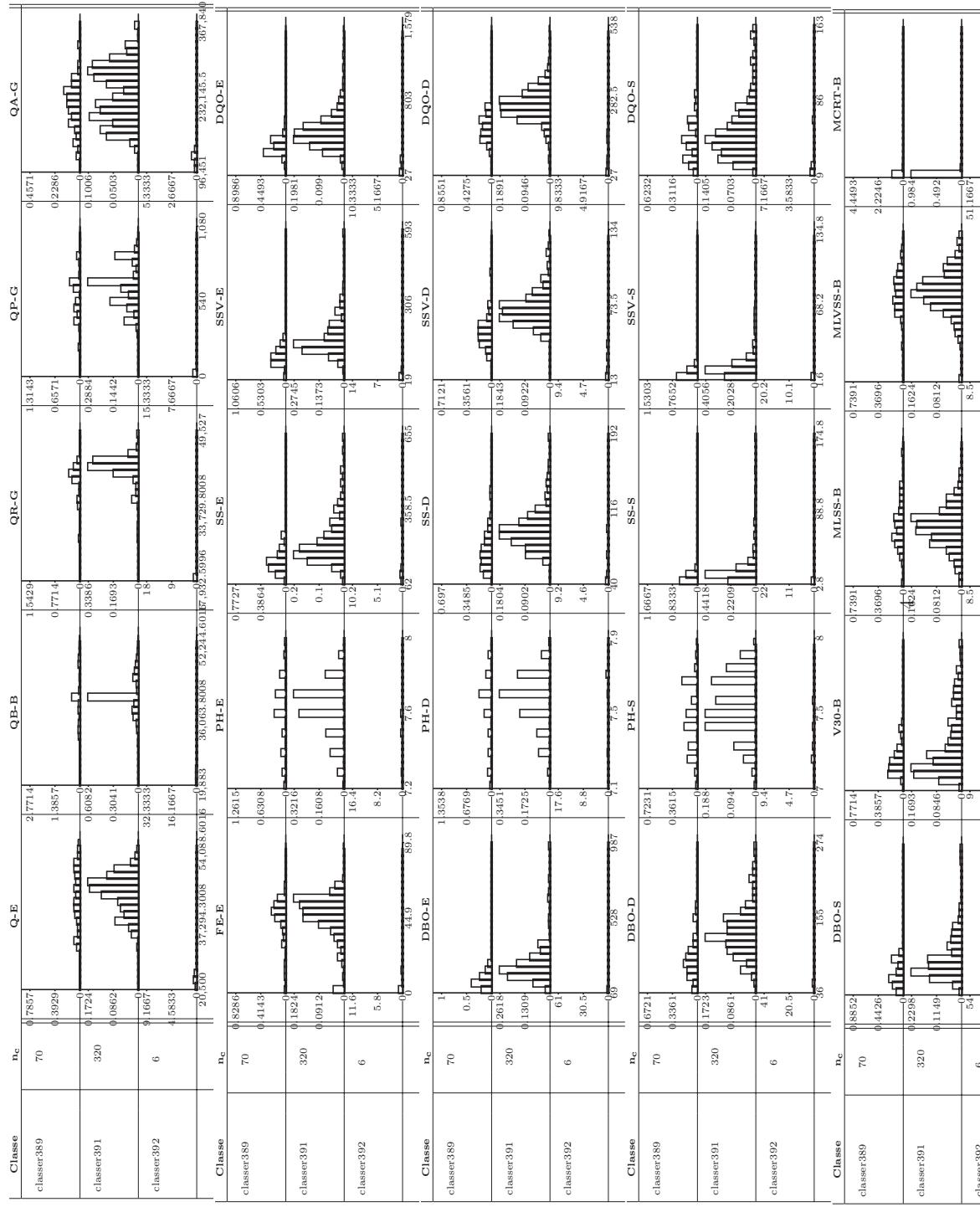
Figura 16.2: CAJ. Árbol general de clasificación con reglas. $[\tau_{Gi2,R1}^{En,G}]$.

$$\text{Classer394} \left\{ \begin{array}{l} \text{Classer392} \\ \text{Classer393} \left\{ \begin{array}{l} \text{Classer389} \\ \text{Classer391} \left\{ \begin{array}{l} \text{Classer390} \\ \text{Classer388} \end{array} \right. \end{array} \right. \end{array} \right.$$

Figura 16.3: Análisis Descriptivo por clases para $[P2_{Gi1,R1}^{En,G}]$.

Figura 16.4: Análisis Descriptivo por clases para $[P2_{Gi1,R1}^{En,G}]$.

Figura 16.5: Análisis Descriptivo por clases para $[P3_{G1,R1}^{En,G}]$.

Figura 16.6: Análisis Descriptivo por clases para $[P3^{En,G}_{Gi1,R1}]$.

16.4 Interpretación validada por el experto

En el estudio realizado para el artículo (Gibert and Roda 2000), se trabaja con varias muestras con una selección de 150 objetos cada una, de esta misma base de datos, y se comprueba que el comportamiento se repite y es muy parecido al nuestro (396 objetos).

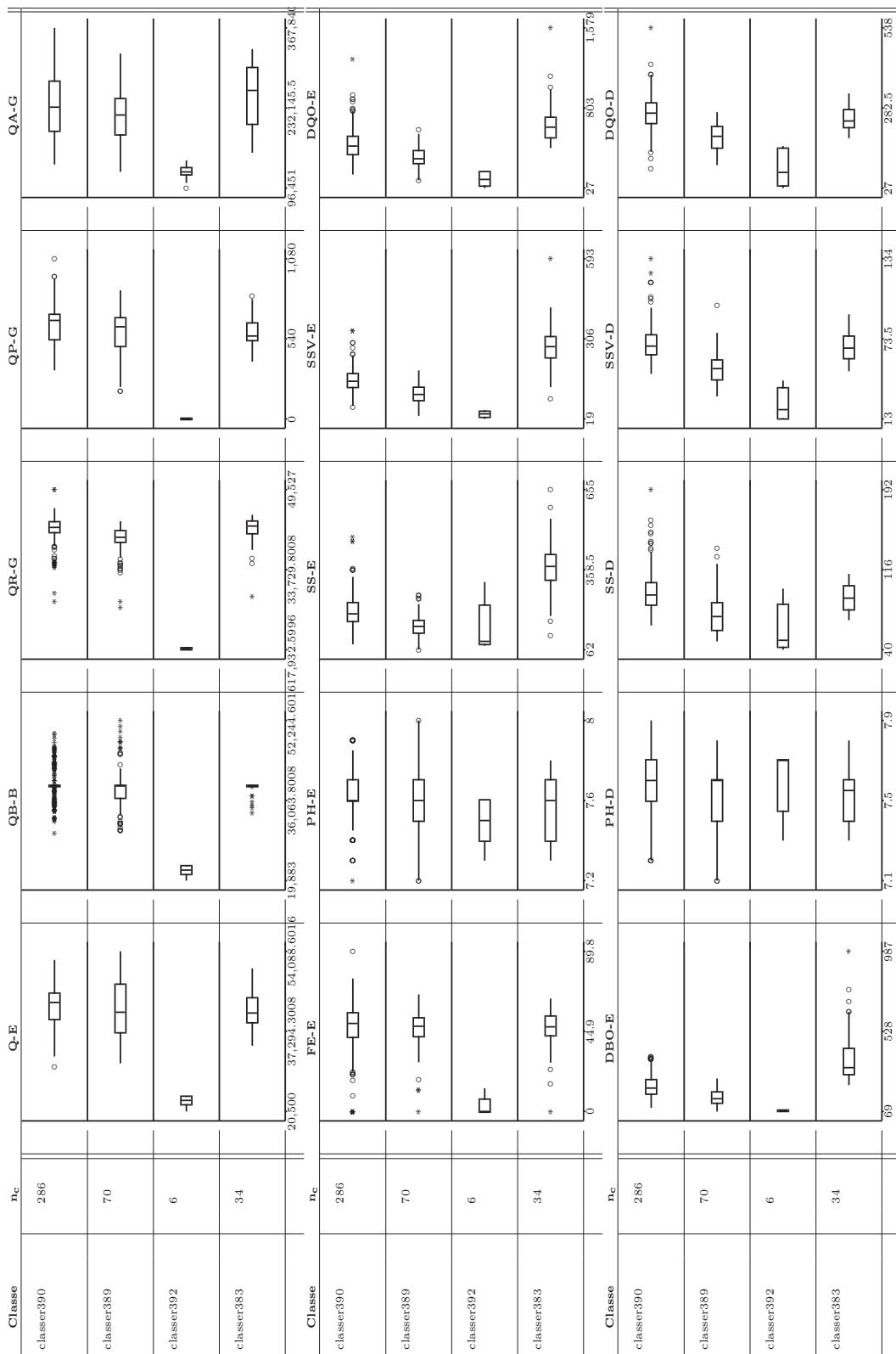
Los expertos han validado que la interpretación publicada en (Gibert and Roda 2000) es válida también para las 4 clases obtenidas con $[P4_{Gi1,R1}^{En,G}]$.

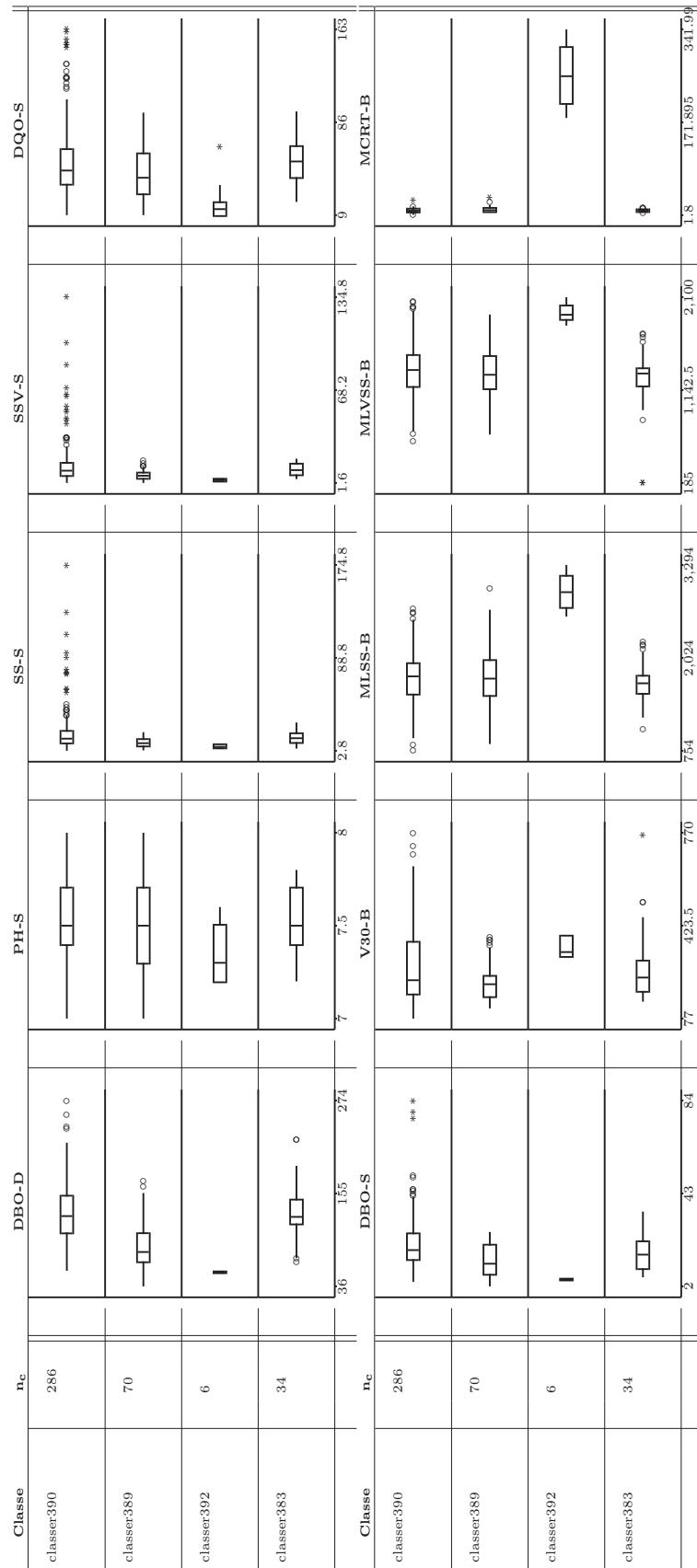
Resumiendo toda la información contenida anteriormente y a juzgar por los expertos que participan en éste estudio, las clases se caracterizan por:

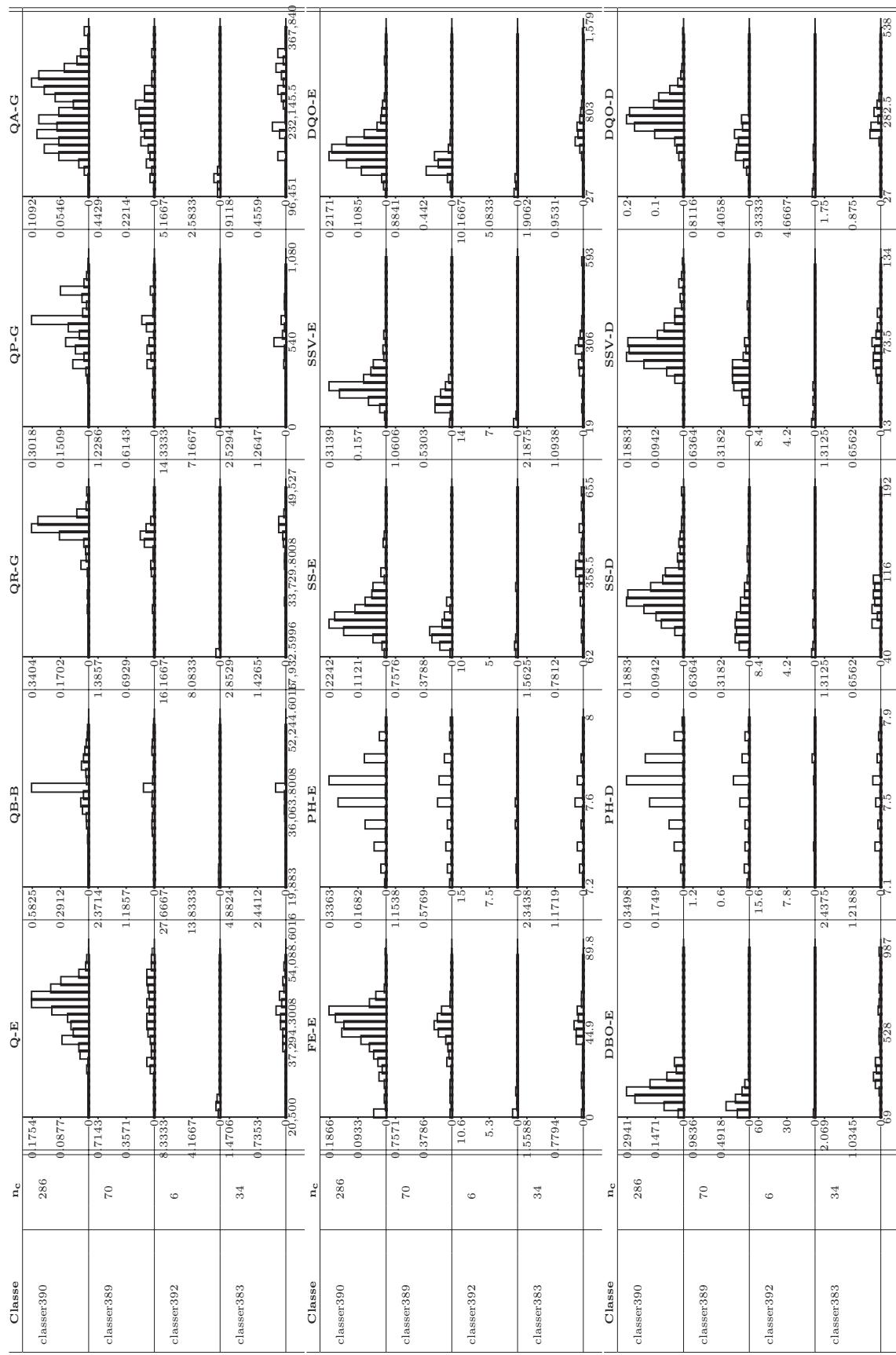
- **Classe392:** Agua que sale más limpia que normalmente (entra con suciedad grado medio) y licor mezcla con más biomasa o microorganismos de lo normal. Valores globalmente intermedios (inferiores en DQO-D, DBO-D, PH-S, SS-S, SSV-S, DQO-S, DBO-S, V30-B, NKT-S y NH4-S, y superiores en MLSS-B y MLSSV-B).

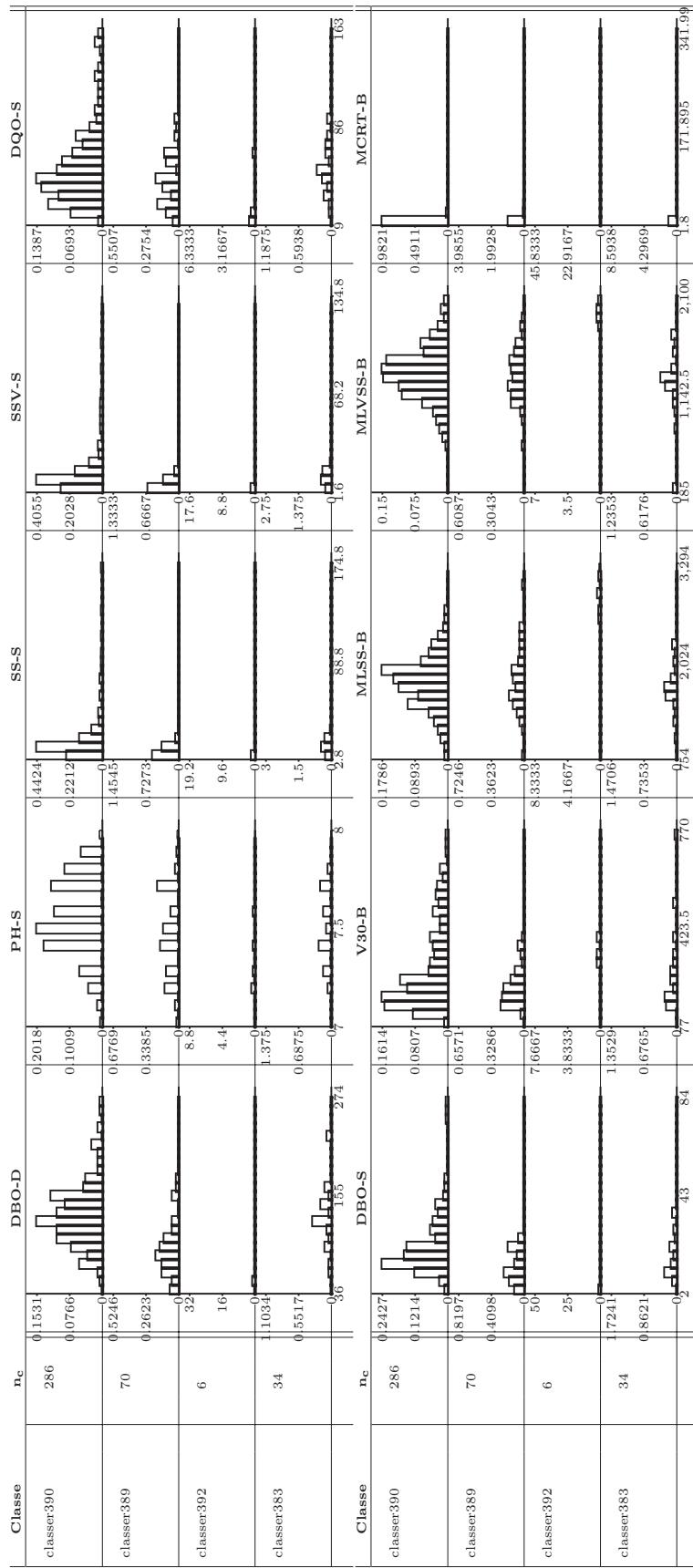
Según el experto corresponde a una clase que se caracteriza por unos elevados rendimientos de depuración (analíticas del agua en la salida y en el punto intermedio con valores bajos) debido a que hay un elevado nivel de biomasa en los reactores. Además, a partir de los últimos gráficos (estudio dinámico) hemos podido observar que esta clase corresponde a agua con nutrientes eliminados básicamente en el reactor biológico.

- **Clase389:** Agua que entra y sale con grado de suciedad medio, así como el licor mezcla aunque tiende a caudales de purga elevados y afluencia de aire baja. Corresponde a valores intermedios para casi todas las variables. Según el experto, cuando se purga más (QP-G elevada) se supone que hay menos biomasa en los reactores y, por lo tanto, es necesario airear menos (QA-G baja). De todas formas se tendría que reflejar en algún otro lugar, con unos MLSSV-B menores o un aumento de FE-E. En este caso, a partir de los últimos gráficos (estudio dinámico) no hemos podido caracterizar esta clase por ningún comportamiento particular de las variables a través del proceso de depuración.
- **Classe383:** Agua que entra más sucia de lo normal. Valores globalmente intermedios (superiores en SS-E, SSV-E, DQO-E y DBO-E). Según el experto existe un choque de carga (materia orgánica) de sólidos en la entrada del proceso. Podrían ser vertidos industriales. Además, a partir de los últimos gráficos (estudio dinámico) hemos podido observar que es una clase de aguas con gran número de partículas en suspensión donde la decantación surte máximo efecto.
- **Classe392:** Es una clase muy diferenciada del resto por tomar valores extremadamente pequeños en casi todas las variables (caudales con valores bajos y agua que entra, sale y está en el proceso poco sucia) menos el índice V30, la suciedad del licor mezcla y la edad celular que son elevados. Según el experto esto se corresponde a un periodo de tempestad. En Gerona, cuando llueve mucho, se cierran las compuertas que pasan por debajo del río para evitar que se rompan las conducciones. Esto provoca que llegue muy poca agua (Q-E baja) y muy diluida. Por contra, el cabezal de planta cierra completamente la purga (QP-G es 0) y el nivel de la biomasa en los reactores (MLSS, MLVSS) y la edad celular (MCRT-B) aumentan mucho. Ésto podrá hacer frente al choque de carga que llegará después de la tormenta. A partir de los últimos gráficos (estudio dinámico) únicamente hemos podido observar que esta clase posee un comportamiento muy diferenciado del resto de clases, tomando valores muy diferenciados del resto.


 Figura 16.7: Análisis Descriptivo por clases para $[P4_{Gi1,R1}^{En,G}]$ -1.

Figura 16.8: Análisis Descriptivo por clases para $[P4_{G1,R1}^{En,G}]$ -2.

Figura 16.9: Análisis Descriptivo por clases para $[P4^{En,G}_{Gi1,R1}]$.

Figura 16.10: Análisis Descriptivo por clases para $[P4_{Gi1,R1}^{En,G}]$.

Capítulo 17

Aplicación, planta catalana

17.1 Interpretación de \mathcal{P}_4 utilizando Best global concept and Close-World Assumption

1. $\xi = 2$: Así, $\mathcal{P}2_{Gi1,R1}^{EnW,G} = \{Classer392, Classer393\}$. La raíz del árbol *Classer394* tiene 2 hijos:

$$Classer394 \left\{ \begin{array}{l} Classer392 \\ Classer393. \end{array} \right.$$

2. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición $\mathcal{P}2_{Gi1,R1}^{EnW,G} = \{Classer392, Classer393\}$ se presenta en el Apéndice E.1
3. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(\mathcal{P}_2) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_2)$, se presenta en el Apéndice E.2
4. Como resultado de los pasos anteriores, se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_2) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_2)$ y se trabaja con $\mathcal{S}(\mathcal{P}_2) = \bigcup_{k=1}^K \mathcal{S}(X_k, \mathcal{P}_2)$.

$$\begin{aligned} \mathcal{S}(\mathcal{P}_2) = \{ & r_{1, classer392}^{Q-E} : x_{Q-E,i} \in [20500.0, 23662.9] \xrightarrow{1.0} classer392, \\ & r_{1, classer392}^{QB-B} : x_{QB-B,i} \in [19883.0, 22891.0] \xrightarrow{1.0} classer392, \\ & r_{1, classer392}^{QR-G} : x_{QR-G,i} \in [17932.6, 18343.5] \xrightarrow{1.0} classer392, \\ & r_{1, classer392}^{QP-G} : x_{QP-G,i} \in [0.0, 0.0] \xrightarrow{1.0} classer392, \\ & r_{1, classer392}^{QA-G} : x_{QA-G,i} \in [96451.0, 124120.0] \xrightarrow{1.0} classer392, \\ & r_{1, classer392}^{SSV-E} : x_{SSV-E,i} \in [19.0, 30.0] \xrightarrow{1.0} classer392, \\ & r_{1, classer392}^{DQO-E} : x_{DQO-E,i} \in [27.0, 100.0] \xrightarrow{1.0} classer392, \\ & r_{2, classer392}^{DBO-E} : x_{DBO-E,i} \in [73.0, 73.0] \xrightarrow{1.0} classer392, \\ & r_{1, classer392}^{SS-D} : x_{SS-D,i} \in [40.0, 48.0] \xrightarrow{1.0} classer392, \\ & r_{1, classer392}^{SSV-D} : x_{SSV-D,i} \in [13.0, 30.0] \xrightarrow{1.0} classer392, \\ & r_{1, classer392}^{DQO-D} : x_{DQO-D,i} \in [27.0, 90.0] \xrightarrow{1.0} classer392, \\ & r_{2, classer392}^{DBO-D} : x_{DBO-D,i} \in [54.0, 54.0] \xrightarrow{1.0} classer392, \\ & r_{3, classer392}^{MLSS-B} : x_{MLSS-B,i} \in (2978.0, 3294.0] \xrightarrow{1.0} classer392, \\ & r_{3, classer392}^{MLVSS-B} : x_{MLVSS-B,i} \in (2054.0, 2100.0] \xrightarrow{1.0} classer392, \\ & r_{3, classer392}^{MCRT-B} : x_{MCRT-B,i} \in [179.8, 341.99] \xrightarrow{1.0} classer392, \\ & r_{3, classer393}^{Q-E} : x_{Q-E,i} \in [29920.0, 54088.6] \xrightarrow{1.0} classer393, \\ & r_{3, classer393}^{QB-B} : x_{QB-B,i} \in [29397.3, 52244.6] \xrightarrow{1.0} classer393, \\ & r_{3, classer393}^{QR-G} : x_{QR-G,i} \in [26218.0, 49527.0] \xrightarrow{1.0} classer393, \end{aligned}$$

$$\begin{aligned}
r_{3,classer393}^{QP-G} : & x_{QP-G,i} \in [188.0, 1080.0] \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{QA-G} : & x_{QA-G,i} \in (143151.0, 367840.0] \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{FE-E} : & x_{FE-E,i} \in (13.0, 89.8] \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{PH-E} : & x_{PH-E,i} \in [7.2, 7.3) \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{PH-E} : & x_{PH-E,i} \in (7.6, 8.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{SS-E} : & x_{SS-E,i} \in [62.0, 77.0) \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{SS-E} : & x_{SS-E,i} \in (313.0, 655.0] \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{SSV-E} : & x_{SSV-E,i} \in (51.0, 593.0] \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{DQO-E} : & x_{DQO-E,i} \in (180.0, 1579.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{DBO-E} : & x_{DBO-E,i} \in [69.0, 73.0) \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{DBO-E} : & x_{DBO-E,i} \in (73.0, 987.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{PH-D} : & x_{PH-D,i} \in [7.1, 7.3) \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{PH-D} : & x_{PH-D,i} \in (7.7, 7.9] \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{SS-D} : & x_{SS-D,i} \in (98.0, 192.0] \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{SSV-D} : & x_{SSV-D,i} \in (42.0, 134.0] \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{DQO-D} : & x_{DQO-D,i} \in (161.0, 538.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{DBO-D} : & x_{DBO-D,i} \in [36.0, 54.0) \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{DBO-D} : & x_{DBO-D,i} \in (54.0, 274.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{PH-S} : & x_{PH-S,i} \in [7.0, 7.2) \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{PH-S} : & x_{PH-S,i} \in (7.6, 8.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{SS-S} : & x_{SS-S,i} \in [2.8, 5.2) \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{SS-S} : & x_{SS-S,i} \in (8.0, 174.8] \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{SSV-S} : & x_{SSV-S,i} \in [1.6, 2.8) \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{SSV-S} : & x_{SSV-S,i} \in (4.0, 134.8] \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{DQO-S} : & x_{DQO-S,i} \in (66.0, 163.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{DBO-S} : & x_{DBO-S,i} \in [2.0, 5.0) \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{DBO-S} : & x_{DBO-S,i} \in (5.0, 84.0] \xrightarrow{1.0} \text{classer393}, \\
r_{3,classer393}^{V30-B} : & x_{V30-B,i} \in (383.0, 770.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{MLSS-B} : & x_{MLSS-B,i} \in [754.0, 2589.0) \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{MLVSS-B} : & x_{MLVSS-B,i} \in [185.0, 1807.0) \xrightarrow{1.0} \text{classer393}, \\
r_{1,classer393}^{MCRT-B} : & x_{MCRT-B,i} \in [1.8, 34.4] \xrightarrow{1.0} \text{classer393} \} ,
\end{aligned}$$

5. El Cuadro 22.1 muestra la cobertura relativa de las reglas de $\mathcal{S}(\mathcal{P}_2)$. Hay más de una regla con cobertura relativa máxima y $p_{sc} = 1$ en $\mathcal{S}(\mathcal{P}_2)$, estas son:

$$\begin{aligned}
r_{1,classer392}^{Q-E}, r_{1,classer392}^{QB-B}, r_{1,classer392}^{QR-G}, r_{1,classer392}^{QP-G}, r_{3,classer392}^{MCRT-B}, r_{3,classer393}^{Q-E}, r_{3,classer393}^{QB-B}, \\
r_{3,classer393}^{QR-G}, r_{3,classer393}^{QP-G}, r_{1,classer393}^{MCRT-B}.
\end{aligned}$$

Con una $CovR(r)=100\%$, lo que significa que nos encontramos con 5 variables totalmente caracterizadoras, estas son; Q-E, QB-B, QR-G, QP-G y MCRT-B:

$$\begin{aligned}
r_{1,classer392}^{Q-E} : & x_{Q-E,i} \in [20500.0, 23662.9] \xrightarrow{1.0} \text{classer392} \\
r_{1,classer392}^{QB-B} : & x_{QB-B,i} \in [19883.0, 22891.0] \xrightarrow{1.0} \text{classer392} \\
r_{1,classer392}^{QR-G} : & x_{QR-G,i} \in [17932.6, 18343.5] \xrightarrow{1.0} \text{classer392} \\
r_{1,classer392}^{QP-G} : & x_{QP-G,i} \in [0.0, 0.0] \xrightarrow{1.0} \text{classer392} \\
r_{3,classer392}^{MCRT-B} : & x_{MCRT-B,i} \in [179.8, 341.99] \xrightarrow{1.0} \text{classer392} \\
r_{3,classer393}^{Q-E} : & x_{Q-E,i} \in [29920.0, 54088.6] \xrightarrow{1.0} \text{classer393} \\
r_{3,classer393}^{QB-B} : & x_{QB-B,i} \in [29397.3, 52244.6] \xrightarrow{1.0} \text{classer393}
\end{aligned}$$

$$\begin{aligned}
r_{3,\text{classer393}}^{QR-G} : x_{QR-G,i} \in [26218.0, 49527.0] &\xrightarrow{1.0} \text{classer393} \\
r_{3,\text{classer393}}^{QP-G} : x_{QP-G,i} \in [188.0, 1080.0] &\xrightarrow{1.0} \text{classer393} \\
r_{1,\text{classer393}}^{MCRT-B} : x_{MCRT-B,i} \in [1.8, 34.4] &\xrightarrow{1.0} \text{classer393}
\end{aligned}$$

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	CovR(r)	Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	CovR(r)
$r_{1,\text{classer392}}^{Q-E}$	6	100%	$r_{3,\text{classer393}}^{SSV-E}$	319	82%
$r_{1,\text{classer392}}^{QB-B}$	6	100%	$r_{3,\text{classer393}}^{DQO-E}$	379	97.4%
$r_{1,\text{classer392}}^{QR-G}$	6	100%	$r_{1,\text{classer393}}^{DBO-E}$	2	0.5%
$r_{1,\text{classer392}}^{QP-G}$	6	100%	$r_{3,\text{classer393}}^{DBO-E}$	292	75.1%
$r_{1,\text{classer392}}^{QA-G}$	3	50%	$r_{1,\text{classer393}}^{PH-D}$	11	2.8%
$r_{1,\text{classer392}}^{SSV-E}$	1	16,7%	$r_{3,\text{classer393}}^{PH-D}$	20	5.1%
$r_{1,\text{classer392}}^{DQO-E}$	3	50%	$r_{3,\text{classer393}}^{SS-D}$	38	9.8%
$r_{2,\text{classer392}}^{DBO-E}$	1	16,7%	$r_{3,\text{classer393}}^{SSV-D}$	305	78.4%
$r_{1,\text{classer392}}^{SS-D}$	2	33,33%	$r_{3,\text{classer393}}^{DQO-D}$	354	93.2%
$r_{1,\text{classer392}}^{SSV-D}$	3	50%	$r_{1,\text{classer393}}^{DBO-D}$	4	1%
$r_{1,\text{classer392}}^{DQO-D}$	3	50%	$r_{3,\text{classer393}}^{DBO-D}$	295	75.8%
$r_{2,\text{classer392}}^{DBO-D}$	1	16,7%	$r_{1,\text{classer393}}^{PH-S}$	7	1.8%
$r_{3,\text{classer392}}^{MLSS-B}$	2	33,33%	$r_{3,\text{classer393}}^{PH-S}$	102	26.2%
$r_{3,\text{classer392}}^{MLVSS-B}$	1	16,7%	$r_{1,\text{classer393}}^{SS-S}$	13	3.3%
$r_{3,\text{classer392}}^{MCRT-B}$	6	100%	$r_{3,\text{classer393}}^{SS-S}$	259	66.6%
$r_{3,\text{classer393}}^{Q-E}$	389	100%	$r_{1,\text{classer393}}^{SSV-S}$	9	2.3%
$r_{3,\text{classer393}}^{QB-B}$	389	100%	$r_{3,\text{classer393}}^{SSV-S}$	294	75.6%
$r_{3,\text{classer393}}^{QR-G}$	389	100%	$r_{3,\text{classer393}}^{DQO-S}$	76	19.5%
$r_{3,\text{classer393}}^{QP-G}$	389	100%	$r_{1,\text{classer393}}^{DBO-S}$	5	1.3 %
$r_{3,\text{classer393}}^{QA-G}$	384	98.7 %	$r_{3,\text{classer393}}^{DBO-S}$	5	1.3%
$r_{3,\text{classer393}}^{FE-E}$	372	95.6%	$r_{3,\text{classer393}}^{V30-B}$	64	16.5%
$r_{1,\text{classer393}}^{PH-E}$	3	0.8%	$r_{1,\text{classer393}}^{MLSS-B}$	379	97.4%
$r_{3,\text{classer393}}^{PH-E}$	148	38%	$r_{1,\text{classer393}}^{MLVSS-B}$	369	94.9%
$r_{1,\text{classer393}}^{SS-E}$	1	0.3%	$r_{1,\text{classer393}}^{MCRT-B}$	389	100%
$r_{3,\text{classer393}}^{SS-E}$	35	9%			

Tabla 17.1: Cobertura relativa de $\mathcal{S}(\mathcal{P}_2)$.

Así,

$$\begin{aligned}
A_{\text{classer392}}^{2,Q-E} &= "x_{Q-E,i} \in [20500.0, 23662.9]" \\
A_{\text{classer392}}^{2,QB-B} &= "x_{QB-B,i} \in [19883.0, 22891.0]" \\
A_{\text{classer392}}^{2,QR-G} &= "x_{QR-G,i} \in [17932.6, 18343.5]" \\
A_{\text{classer392}}^{2,QP-G} &= "x_{QP-G,i} \in [0.0, 0.0]" \\
A_{\text{classer392}}^{2,MCRT-B} &= "x_{MCRT-B,i} \in [179.8, 341.99]" \\
A_{\text{classer393}}^{2,Q-E} &= "x_{Q-E,i} \in [29920.0, 54088.6]" \\
A_{\text{classer393}}^{2,QB-B} &= "x_{QB-B,i} \in [29397.3, 52244.6]" \\
A_{\text{classer393}}^{2,QR-G} &= "x_{QR-G,i} \in [26218.0, 49527.0]" \\
A_{\text{classer393}}^{2,QP-G} &= "x_{QP-G,i} \in [188.0, 1080.0]"
\end{aligned}$$

$$A_{\text{classer393}}^{2,MCRT-B} = "x_{MCRT-B,i} \in [1.8, 34.4]"$$

En este caso hay mas de una reglas (ver ecuaciones (10.4) y (10.5)) y por tanto:

$$\begin{aligned} \bullet \quad & A_{\text{classer392}}^2 = A_{\text{classer392}}^{2,Q-E} \wedge A_{\text{classer392}}^{2,QB-B} \wedge A_{\text{classer392}}^{2,QR-G} \wedge A_{\text{classer392}}^{2,QP-G} \wedge A_{\text{classer392}}^{2,MCRT-B} \\ & = "x_{Q-E,i} \in [20500.0, 23662.9]" \wedge "x_{QB-B,i} \in [19883.0, 22891.0]" \wedge \\ & \quad "x_{QR-G,i} \in [17932.6, 18343.5]" \wedge "x_{QP-G,i} \in [0.0, 0.0]" \wedge \\ & \quad "x_{MCRT-B,i} \in [179.8, 341.99]" \\ \bullet \quad & A_{\text{classer393}}^2 = A_{\text{classer393}}^{2,Q-E} \wedge A_{\text{classer393}}^{2,QB-B} \wedge A_{\text{classer393}}^{2,QR-G} \wedge A_{\text{classer393}}^{2,QP-G} \wedge A_{\text{classer393}}^{2,MCRT-B} \\ & = "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\ & \quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\ & \quad "x_{MCRT-B,i} \in [1.8, 34.4]" \end{aligned}$$

6. Asociando una regla a cada clase,

$$\begin{aligned} \mathbb{R}(\mathcal{P}_2) = \{ & \quad r_{\text{Classer392}} : \quad x_{Q-E,i} \in [20500.0, 23662.9] \wedge \\ & \quad x_{QB-B,i} \in [19883.0, 22891.0] \wedge \\ & \quad x_{QR-G,i} \in [17932.6, 18343.5] \wedge \\ & \quad x_{QP-G,i} \in [0.0, 0.0] \wedge \\ & \quad x_{MCRT-B,i} \in [179.8, 341.99] \quad \xrightarrow{1.0} \text{Classer392}, \end{aligned}$$

$$\begin{aligned} & r_{\text{Classer393}} : \quad x_{Q-E,i} \in [29920.0, 54088.6] \wedge \\ & \quad x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\ & \quad x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\ & \quad x_{QP-G,i} \in [188.0, 1080.0] \wedge \\ & \quad x_{MCRT-B,i} \in [1.8, 34.4] \quad \xrightarrow{1.0} \text{Classer393} \} \end{aligned}$$

Que correspondería a la conceptualization:

- Classer392: “ $Q-E$ es bajo y $QB-B$ es bajo y $QR-G$ es bajo y $QP-G$ es bajo y $MCRT-B$ es alto”
- Classer393: “ $Q-E$ es alto y $QB-B$ es alto y $QR-G$ es alto y $QP-G$ es alto y $MCRT-B$ es bajo”

7. $\xi = 3$: Así, $\mathcal{P}_3^{EnW,G}_{Gi1,R1} = \{\text{Classer392}, \text{Classer389}, \text{Classer391}\}$. La clase *Classer392* es la que se divide en 2 hijos y la clase *Classer393* es la que ya estaba en la partición anterior:

$$\text{Classer393} \left\{ \begin{array}{l} \text{Classer389} \\ \text{Classer391} \end{array} \right.$$

Así $C_i^3 = \text{Classer389}; C_j^3 = \text{Classer391}; C_t^2 = \text{Classer393}$.

8. La discretización realizada con el BBD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_3^* = \{\text{Classer389}, \text{Classer391}\}$, donde $\mathcal{P}_3^* \subset \mathcal{P}_3^{EnW,G}_{Gi1,R1}$, se presenta en el Apéndice E.3.

9. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_3^*)$.
10. Se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_3^*) = \bigcup_{k=1}^k \mathcal{R}(X_k, \mathcal{P}_3^*)$ que se presenta en el Apéndice E.4 y se trabaja con $\mathcal{S}(\mathcal{P}_3^*) = \bigcup_{k=1}^k \mathcal{S}(X_k, \mathcal{P}_3^*)$ que es:

$$\begin{aligned} \mathcal{S}(\mathcal{P}_3^*) = \{ & r_{3, classer389}^{Q-E} : x_{Q-E,i} \in (52255.8, 54088.6] \xrightarrow{1.0} classer389 , \\ & r_{3, classer389}^{QB-B} : x_{QB-B,i} \in (49695.8, 52244.6] \xrightarrow{1.0} classer389 , \\ & r_{3, classer389}^{QR-G} : x_{QR-G,i} \in [26218.0, 27351.0] \xrightarrow{1.0} classer389 , \\ & r_{1, classer389}^{QP-G} : x_{QP-G,i} \in [188.0, 327.6) \xrightarrow{1.0} classer389 , \\ & r_{1, classer389}^{QA-G} : x_{QA-G,i} \in [124120.0, 136371.0) \xrightarrow{1.0} classer389 , \\ & r_{1, classer389}^{PH-E} : x_{PH-E,i} \in (7.9, 8.0] \xrightarrow{1.0} classer389 , \\ & r_{1, classer389}^{SS-E} : x_{SS-E,i} \in [62.0, 82.0) \xrightarrow{1.0} classer389 , \\ & r_{1, classer389}^{SSV-E} : x_{SSV-E,i} \in [30.0, 60.0) \xrightarrow{1.0} classer389 , \\ & r_{1, classer389}^{DQO-E} : x_{DQO-E,i} \in [100.0, 158.0) \xrightarrow{1.0} classer389 , \\ & r_{1, classer389}^{DBO-E} : x_{DBO-E,i} \in [69.0, 90.0) \xrightarrow{1.0} classer389 , \\ & r_{1, classer389}^{SS-D} : x_{SS-D,i} \in [48.0, 63.0) \xrightarrow{1.0} classer389 , \\ & r_{1, classer389}^{SSV-D} : x_{SSV-D,i} \in [30.0, 47.0) \xrightarrow{1.0} classer389 , \\ & r_{1, classer389}^{DBO-D} : x_{DBO-D,i} \in [36.0, 56.0) \xrightarrow{1.0} classer389 , \\ & r_{1, classer389}^{DBO-S} : x_{DBO-S,i} \in [2.0, 4.0) \xrightarrow{1.0} classer389 , \\ & r_{3, classer389}^{MLSS-B} : x_{MLSS-B,i} \in (2696.0, 2978.0] \xrightarrow{1.0} classer389 , \\ & r_{3, classer389}^{MCRT-B} : x_{MCRT-B,i} \in (28.8, 34.4] \xrightarrow{1.0} classer389 , \\ & r_{1, classer391}^{Q-E} : x_{Q-E,i} \in [29920.0, 30592.2) \xrightarrow{1.0} classer391 , \\ & r_{1, classer391}^{QB-B} : x_{QB-B,i} \in [29397.3, 29936.8) \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{QR-G} : x_{QR-G,i} \in (43298.1, 49527.0] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{QP-G} : x_{QP-G,i} \in (866.7, 1080.0] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{QA-G} : x_{QA-G,i} \in (324470.0, 367840.0] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{FE-E} : x_{FE-E,i} \in (65.6, 89.8] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{SS-E} : x_{SS-E,i} \in (266.0, 655.0] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{SSV-E} : x_{SSV-E,i} \in (193.0, 593.0] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{DQO-E} : x_{DQO-E,i} \in (595.0, 1579.0] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{DBO-E} : x_{DBO-E,i} \in (258.0, 987.0] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{PH-D} : x_{PH-D,i} \in (7.8, 7.9] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{SS-D} : x_{SS-D,i} \in (136.0, 192.0] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{SSV-D} : x_{SSV-D,i} \in (99.0, 134.0] \xrightarrow{1.0} classer391 , \\ & r_{1, classer391}^{DQO-D} : x_{DQO-D,i} \in [90.0, 100.0) \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{DQO-D} : x_{DQO-D,i} \in (269.0, 538.0] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{DBO-D} : x_{DBO-D,i} \in (171.0, 274.0] \xrightarrow{1.0} classer391 , \\ & r_{1, classer391}^{SS-S} : x_{SS-S,i} \in [2.8, 3.2) \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{SS-S} : x_{SS-S,i} \in (20.0, 174.8] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{SSV-S} : x_{SSV-S,i} \in (18.0, 134.8] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{DQO-S} : x_{DQO-S,i} \in (94.0, 163.0] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{DBO-S} : x_{DBO-S,i} \in (26.0, 84.0] \xrightarrow{1.0} classer391 , \\ & r_{1, classer391}^{V30-B} : x_{V30-B,i} \in [77.0, 115.0] \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{V30-B} : x_{V30-B,i} \in (380.0, 770.0] \xrightarrow{1.0} classer391 , \\ & r_{1, classer391}^{MLSS-B} : x_{MLSS-B,i} \in [754.0, 846.0) \xrightarrow{1.0} classer391 , \\ & r_{1, classer391}^{MLVSS-B} : x_{MLVSS-B,i} \in [185.0, 684.0) \xrightarrow{1.0} classer391 , \\ & r_{3, classer391}^{MLVSS-B} : x_{MLVSS-B,i} \in (1921.0, 2054.0] \xrightarrow{1.0} classer391 , \\ & r_{1, classer391}^{MCRT-B} : x_{MCRT-B,i} \in [1.8, 6.9] \xrightarrow{1.0} classer391 \} , \end{aligned}$$

11. El Cuadro 22.2 muestra la cobertura relativa de las reglas de $\mathcal{S}(\mathcal{P}_3^*)$. La regla con

mayor cobertura relativa y $p_{sc} = 1$ de $\mathcal{S}(\mathcal{P}_3^*)$ es $r_{3, classer391}^{DQO-D}$, con una cobertura relativa $CovR = 45,63\%$:

$$r_{3, classer391}^{DQO-D} : x_{DQO-D,i} \in (269.0, 538.0] \xrightarrow{1.0} classer391$$

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$	Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3, classer389}^{Q-E}$	3	4,28%	$r_{3, classer391}^{SS-E}$	60	18,75 %
$r_{3, classer389}^{QB-B}$	3	4,28%	$r_{3, classer391}^{SSV-E}$	70	21,875 %
$r_{1, classer389}^{QR-G}$	2	2,85%	$r_{3, classer391}^{DQO-E}$	52	16,25 %
$r_{1, classer389}^{QP-G}$	2	2,85%	$r_{3, classer391}^{DBO-E}$	64	20 %
$r_{1, classer389}^{QA-G}$	2	2,85%	$r_{3, classer391}^{PH-D}$	1	0,3125 %
$r_{3, classer389}^{PH-E}$	1	1,42%	$r_{3, classer391}^{SS-D}$	8	2,5 %
$r_{1, classer389}^{SS-E}$	1	1,42%	$r_{3, classer391}^{SSV-D}$	8	2,5 %
$r_{1, classer389}^{SSV-E}$	2	2,85%	$r_{1, classer391}^{DQO-D}$	1	0,3125 %
$r_{1, classer389}^{DQO-E}$	1	1,42%	$r_{3, classer391}^{DQO-D}$	146	45,625 %
$r_{1, classer389}^{DBO-E}$	5	7,14%	$r_{3, classer391}^{DBO-D}$	25	7,8125 %
$r_{1, classer389}^{SS-D}$	24	34,28%	$r_{1, classer391}^{SS-S}$	1	0,3125 %
$r_{1, classer389}^{SSV-D}$	25	35,71%	$r_{3, classer391}^{SS-S}$	59	18,4375 %
$r_{1, classer389}^{DBO-D}$	5	7,14%	$r_{3, classer391}^{SSV-S}$	42	13,125 %
$r_{1, classer389}^{DBO-S}$	2	2,85%	$r_{3, classer391}^{DQO-S}$	26	8,125 %
$r_{3, classer389}^{MLSS-B}$	1	1,42%	$r_{3, classer391}^{DBO-S}$	55	17,1875 %
$r_{3, classer389}^{MCRT-B}$	1	1,42%	$r_{1, classer391}^{V30-B}$	6	1,875 %
$r_{1, classer391}^{Q-E}$	1	0,3125 %	$r_{3, classer391}^{V30-B}$	64	20 %
$r_{1, classer391}^{QB-B}$	1	0,3125 %	$r_{1, classer391}^{MLSS-B}$	2	0,625 %
$r_{3, classer391}^{QR-G}$	64	20 %	$r_{1, classer391}^{MLVSS-B}$	3	0,9375 %
$r_{3, classer391}^{QP-G}$	36	11,25 %	$r_{3, classer391}^{MLVSS-B}$	6	1,875 %
$r_{3, classer391}^{QA-G}$	8	2,5 %	$r_{1, classer391}^{MCRT-B}$	27	8,4375 %
$r_{3, classer391}^{FE-E}$	4	1,25 %			

Tabla 17.2: Cobertura relativa de $\mathcal{S}(\mathcal{P}_3^*)$.

En este caso no hay empate , ver ecuaciones (10.2) y (10.3), así:

- $A_{classer391}^{*3} = A_{classer391}^{3,DQO-D} = "x_{DQO-D,i} \in (269.0, 538.0]"$
- $A_{classer389}^{*3} = A_{classer389}^{3,DQO-D} = \neg A_{classer391}^{3,DQO-D} = "x_{DQO-D,i} \in [90.0, 269.0]"$

12. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

$$\bullet A_{classer392}^3 = A_{classer392}^2$$

$$A_{classer392}^3 = A_{classer392}^{2,Q-E} \wedge A_{classer392}^{2,QB-B} \wedge A_{classer392}^{2,QR-G} \wedge A_{classer392}^{2,QP-G} \wedge A_{classer393}^{2,MCRT-B} \wedge A_{classer392}^{2,MLSS-B}$$

$$= "x_{Q-E,i} \in [20500.0, 23662.9]" \wedge "x_{QB-B,i} \in [19883.0, 22891.0]" \wedge \\ "x_{QR-G,i} \in [17932.6, 18343.5]" \wedge "x_{QP-G,i} \in [0.0, 0.0]" \wedge \\ "x_{MCRT-B,i} \in [179.8, 341.99]"$$

- $A_{classer389}^3 = A_{classer393}^2 \wedge A_{classer389}^{*3}$

$$\begin{aligned}
A_{classer389}^3 &= A_{classer393}^{2,Q-E} \wedge A_{classer393}^{2,QB-B} \wedge A_{classer393}^{2,QR-G} \wedge A_{classer393}^{2,QP-G} \wedge A_{classer393}^{2,MCRT-B} \wedge A_{classer393}^{3,DQO-D} \\
&= "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\
&\quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\
&\quad "x_{MCRT-B,i} \in [1.8, 34.4]" \wedge "x_{DQO-D,i} \in [90.0, 269.0]"
\end{aligned}$$

- $A_{classer391}^3 = A_{classer393}^2 \wedge A_{classer391}^{*3}$

$$\begin{aligned}
A_{classer391}^3 &= A_{classer393}^{2,Q-E} \wedge A_{classer393}^{2,QB-B} \wedge A_{classer393}^{2,QR-G} \wedge A_{classer393}^{2,QP-G} \wedge A_{classer393}^{2,MCRT-B} \wedge A_{classer393}^{3,DQO-D} \\
&= "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\
&\quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\
&\quad "x_{MCRT-B,i} \in [1.8, 34.4]" \wedge "x_{DQO-D,i} \in (269.0, 538.0]"
\end{aligned}$$

Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$p_{sClasser389} = \frac{\text{card}\{i \in C \mid \text{tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{70}{227} = 30.84\%$$

$$\begin{aligned}
\mathbb{R}(\mathcal{P}_3) = \{ & r_{Classer392} : x_{Q-E,i} \in [20500.0, 23662.9] \wedge \\
& x_{QB-B,i} \in [19883.0, 22891.0] \wedge \\
& x_{QR-G,i} \in [17932.6, 18343.5] \wedge \\
& x_{QP-G,i} \in [0.0, 0.0] \wedge \\
& x_{MCRT-B,i} \in [179.8, 341.99] \quad \xrightarrow{1.0} Classer392 \}
\end{aligned}$$

$$\begin{aligned}
r_{Classer389} : & x_{Q-E,i} \in [29920.0, 54088.6] \wedge \\
& x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\
& x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\
& x_{QP-G,i} \in [188.0, 1080.0] \wedge \\
& x_{MCRT-B,i} \in [1.8, 34.4] \wedge \\
& x_{DQO-D,i} \in [90.0, 269.0] \quad \xrightarrow{0.31} Classer389,
\end{aligned}$$

$$\begin{aligned}
r_{Classer391} : & x_{Q-E,i} \in [29920.0, 54088.6] \wedge \\
& x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\
& x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\
& x_{QP-G,i} \in [188.0, 1080.0] \wedge \\
& x_{MCRT-B,i} \in [1.8, 34.4] \wedge \\
& x_{DQO-D,i} \in (269.0, 538.0) \quad \xrightarrow{1.0} Classer391
\end{aligned}$$

Que correspondería a la conceptualization:

– Classer392: “ $Q-E$ es bajo y $QB-B$ es bajo y $QR-G$ es bajo y $QP-G$ es bajo y $MCRT-B$ es alto”

– Classer389: “ $Q\text{-}E$ es alto y $QB\text{-}B$ es alto y $QR\text{-}G$ es alto y $QP\text{-}G$ es alto y $MCRT\text{-}B$ es bajo y $DQO\text{-}D$ no alto”

– Classer391: “ $Q\text{-}E$ es alto y $QB\text{-}B$ es alto y $QR\text{-}G$ es alto y $QP\text{-}G$ es alto y $MCRT\text{-}B$ es bajo y $DQO\text{-}D$ alto”

13. $\xi = 4$: Así, $\mathcal{P}_4^{EnW,G}_{Gi1,R1} = \{\text{Classer392}, \text{Classer389}, \text{Classer390}, \text{Classer383}\}$. La clase Classer393 de $\mathcal{P}_3^{EnW,G}_{Gi1,R1}$ tiene 2 hijos:

$$\text{Classer391} \left\{ \begin{array}{l} \text{Classer383} \\ \text{Classer390} \end{array} \right.$$

Así $C_i^4 = \text{Classer383}$; $C_j^4 = \text{Classer390}$; $C_t^3 = \text{Classer391}$

14. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_4^* = \{\text{Classer383}, \text{Classer390}\}$, donde $\mathcal{P}_4^* \subseteq \mathcal{P}_4^{EnW,G}_{Gi1,R1}$, se presenta en el Apéndice E.5.
15. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_4^*)$.
16. Se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_4^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_4^*)$ que se presenta en el Apéndice E.6 y $\mathcal{S}(\mathcal{P}_4^*) = \bigcup_{k=1}^K \mathcal{S}(X_k, \mathcal{P}_4^*)$ es:

$$\begin{aligned} \mathcal{S}(\mathcal{P}_4^*) = \{ & r_{3,\text{classer383}}^{SS-E} : x_{SS-E,i} \in (480.0, 655.0] \xrightarrow{1.0} \text{classer383} , \\ & r_{3,\text{classer383}}^{SSV-E} : x_{SSV-E,i} \in (336.0, 593.0] \xrightarrow{1.0} \text{classer383} , \\ & r_{3,\text{classer383}}^{DQO-E} : x_{DQO-E,i} \in (1279.0, 1579.0] \xrightarrow{1.0} \text{classer383} , \\ & r_{3,\text{classer383}}^{DBO-E} : x_{DBO-E,i} \in (382.0, 987.0] \xrightarrow{1.0} \text{classer383} , \\ & r_{1,\text{classer383}}^{MLVSS-B} : x_{MLVSS-B,i} \in [185.0, 611.0) \xrightarrow{1.0} \text{classer383} , \\ & r_{1,\text{classer383}}^{MCRT-B} : x_{MCRT-B,i} \in [1.8, 6.2) \xrightarrow{1.0} \text{classer383} , \\ & r_{1,\text{classer390}}^{Q-E} : x_{Q-E,i} \in [29920.0, 34284.4) \xrightarrow{1.0} \text{classer390} , \\ & r_{3,\text{classer390}}^{Q-E} : x_{Q-E,i} \in (50500.5, 52255.8] \xrightarrow{1.0} \text{classer390} , \\ & r_{1,\text{classer390}}^{QB-B} : x_{QB-B,i} \in [29397.3, 33549.4) \xrightarrow{1.0} \text{classer390} , \\ & r_{3,\text{classer390}}^{QB-B} : x_{QB-B,i} \in (39000.0, 49695.8] \xrightarrow{1.0} \text{classer390} , \\ & r_{1,\text{classer390}}^{QR-G} : x_{QR-G,i} \in [27351.0, 28343.8) \xrightarrow{1.0} \text{classer390} , \\ & r_{3,\text{classer390}}^{QR-G} : x_{QR-G,i} \in (44568.6, 49527.0] \xrightarrow{1.0} \text{classer390} , \\ & r_{1,\text{classer390}}^{QP-G} : x_{QP-G,i} \in [327.6, 385.9) \xrightarrow{1.0} \text{classer390} , \\ & r_{3,\text{classer390}}^{QP-G} : x_{QP-G,i} \in (831.1, 1080.0] \xrightarrow{1.0} \text{classer390} , \\ & r_{1,\text{classer390}}^{QA-G} : x_{QA-G,i} \in [136371.0, 156320.0) \xrightarrow{1.0} \text{classer390} , \\ & r_{3,\text{classer390}}^{QA-G} : x_{QA-G,i} \in (331990.0, 367840.0] \xrightarrow{1.0} \text{classer390} , \\ & r_{3,\text{classer390}}^{FE-E} : x_{FE-E,i} \in (63.3, 89.8] \xrightarrow{1.0} \text{classer390} , \\ & r_{1,\text{classer390}}^{PH-E} : x_{PH-E,i} \in [7.2, 7.3) \xrightarrow{1.0} \text{classer390} , \\ & r_{3,\text{classer390}}^{PH-E} : x_{PH-E,i} \in (7.8, 7.9] \xrightarrow{1.0} \text{classer390} , \\ & r_{1,\text{classer390}}^{SS-E} : x_{SS-E,i} \in [82.0, 114.0) \xrightarrow{1.0} \text{classer390} , \\ & r_{1,\text{classer390}}^{SSV-E} : x_{SSV-E,i} \in [60.0, 92.0) \xrightarrow{1.0} \text{classer390} , \\ & r_{1,\text{classer390}}^{DQO-E} : x_{DQO-E,i} \in [158.0, 414.0) \xrightarrow{1.0} \text{classer390} , \\ & r_{1,\text{classer390}}^{DBO-E} : x_{DBO-E,i} \in [90.0, 220.0) \xrightarrow{1.0} \text{classer390} , \\ & r_{1,\text{classer390}}^{PH-D} : x_{PH-D,i} \in [7.2, 7.3) \xrightarrow{1.0} \text{classer390} , \\ & r_{3,\text{classer390}}^{PH-D} : x_{PH-D,i} \in (7.8, 7.9] \xrightarrow{1.0} \text{classer390} \} \end{aligned}$$

17. El Cuadro 22.3 muestra la cobertura relativa de las reglas de $\mathcal{S}(\mathcal{P}_4^*)$. La regla con mayor cobertura relativa de $\mathcal{S}(\mathcal{P}_4^*)$ es $r_{1,classer390}^{DBO-E}$, con una cobertura relativa $CovR = 44,21\%$:

$$r_{1,classer390}^{DBO-E} : x_{DBO-E,i} \in [90.0, 220.0] \xrightarrow{1.0} classer390$$

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,classer383}^{SSV-E}$	4	11,76%
$r_{3,classer383}^{DQO-E}$	1	2,94%
$r_{3,classer383}^{DBO-E}$	7	20,58%
$r_{1,classer383}^{MLVSS-B}$	2	5,88%
$r_{1,classer383}^{MCRT-B}$	1	2,94%
$r_{1,classer390}^{Q-E}$	9	3,15%
$r_{3,classer390}^{Q-E}$	1	0,35%
$r_{1,classer390}^{QB-B}$	8	2,80%
$r_{3,classer390}^{QB-B}$	60	21,06%
$r_{1,classer390}^{QR-G}$	1	0,35%
$r_{3,classer390}^{QR-G}$	13	4,56%
$r_{1,classer390}^{QP-G}$	7	2,45%
$r_{3,classer390}^{QP-G}$	56	19,64%
$r_{1,classer390}^{QA-G}$	7	2,45%
$r_{3,classer390}^{QA-G}$	2	0,70%
$r_{3,classer390}^{FE-E}$	7	2,45%
$r_{1,classer390}^{PH-E}$	1	0,35%
$r_{3,classer390}^{PH-E}$	8	2,80%
$r_{1,classer390}^{SS-E}$	4	1,40%
$r_{1,classer390}^{SSV-E}$	7	2,45%
$r_{1,classer390}^{DQO-E}$	113	39,64%
$r_{1,classer390}^{DBO-E}$	126	44,21%
$r_{1,classer390}^{PH-D}$	3	1,05%
$r_{3,classer390}^{PH-D}$	1	0,35%

Tabla 17.3: Cobertura relativa de $\mathcal{S}(\mathcal{P}_4^*)$.

En este caso no hay empate, ver ecuación (10.2) y ver ecuación (10.3), y por tanto:

- $A_{classer390}^{*4} = A_{classer390}^{4,DBO-E} = "x_{DBO-E,i} \in [90.0, 220.0]"$

- $A_{classer383}^{*4} = A_{classer383}^{4,DBO-E} = \neg A_{classer390}^{4,DBO-E} = "x_{DBO-E,i} \in [220.0, 987.0]"$

18. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

- $A_{classer392}^4 = A_{classer392}^3 = A_{classer392}^2$

$$A_{classer392}^4 = A_{classer392}^{2,Q-E} \wedge A_{classer392}^{2,QB-B} \wedge A_{classer392}^{2,QR-G} \wedge A_{classer392}^{2,QP-G} \wedge A_{classer392}^{2,MCRT-B}$$

$$\begin{aligned}
&= "x_{Q-E,i} \in [20500.0, 23662.9]" \wedge "x_{QB-B,i} \in [19883.0, 22891.0]" \wedge \\
&\quad "x_{QR-G,i} \in [17932.6, 18343.5]" \wedge "x_{QP-G,i} \in [0.0, 0.0]" \wedge \\
&\quad "x_{MCRT-B,i} \in [179.8, 341.99]"
\end{aligned}$$

- $A_{classer389}^4 = A_{classer389}^3 = A_{classer393}^2 \wedge A_{classer389}^{*3}$
- $$\begin{aligned}
A_{classer389}^4 &= A_{classer393}^{2,Q-E} \wedge A_{classer393}^{2,QB-B} \wedge A_{classer393}^{2,QR-G} \wedge A_{classer393}^{2,QP-G} \wedge A_{classer393}^{2,MCRT-B} \wedge A_{classer393}^{3,DQO-D} \\
&= "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\
&\quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\
&\quad "x_{MCRT-B,i} \in [1.8, 34.4]" \wedge "x_{DQO-D,i} \in [90.0, 269.0]"
\end{aligned}$$
-
- $A_{classer383}^4 = A_{classer391}^3 \wedge A_{classer383}^{*4} = A_{classer393}^2 \wedge A_{classer391}^{*3} \wedge A_{classer383}^{*4}$
- $$\begin{aligned}
A_{classer383}^4 &= A_{classer393}^{2,Q-E} \wedge A_{classer393}^{2,QB-B} \wedge A_{classer393}^{2,QR-G} \wedge A_{classer393}^{2,QP-G} \wedge A_{classer393}^{2,MCRT-B} \wedge \\
&\quad A_{classer391}^{3,DQO-D} \wedge A_{classer383}^{4,DBO-E} \\
&= "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\
&\quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\
&\quad "x_{MCRT-B,i} \in [1.8, 34.4]" \wedge "x_{DQO-D,i} \in (269.0, 538.0]" \wedge \\
&\quad "x_{DBO-E,i} \in [220.0, 987.0]"
\end{aligned}$$
-
- $A_{classer390}^4 = A_{classer391}^3 \wedge A_{classer390}^{*4} = A_{classer393}^2 \wedge A_{classer391}^{*3} \wedge A_{classer390}^{*4}$
- $$\begin{aligned}
A_{classer390}^4 &= A_{classer393}^{2,Q-E} \wedge A_{classer393}^{2,QB-B} \wedge A_{classer393}^{2,QR-G} \wedge A_{classer393}^{2,QP-G} \wedge A_{classer393}^{2,MCRT-B} \wedge \\
&\quad A_{classer391}^{3,DQO-D} \wedge A_{classer390}^{4,DBO-E} \\
&= "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\
&\quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\
&\quad "x_{MCRT-B,i} \in [1.8, 34.4]" \wedge "x_{DQO-D,i} \in (269.0, 538.0]" \wedge \\
&\quad "x_{DBO-E,i} \in [90.0, 220.0]"
\end{aligned}$$

Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$p_{scClasser383} = \frac{\text{card}\{i \in C \mid \text{tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{10}{55} = 18.18\%$$

$$p_{scClasser390} = \frac{\text{card}\{i \in C \mid \text{tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{61}{61} = 100\%$$

$$\begin{aligned}
 \mathbb{R}(\mathcal{P}_4) = \{ & \quad r_{Classer392} : \quad x_{Q-E,i} \in [20500.0, 23662.9] \wedge \\
 & \quad x_{QB-B,i} \in [19883.0, 22891.0] \wedge \\
 & \quad x_{QR-G,i} \in [17932.6, 18343.5] \wedge \\
 & \quad x_{QP-G,i} \in [0.0, 0.0] \wedge \\
 & \quad x_{MCRT-B,i} \in [179.8, 341.99] \quad \xrightarrow{1.0} Classer392 \quad \} \\
 \\
 r_{Classer389} : & \quad x_{Q-E,i} \in [29920.0, 54088.6] \wedge \\
 & \quad x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\
 & \quad x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\
 & \quad x_{QP-G,i} \in [188.0, 1080.0] \wedge \\
 & \quad x_{MCRT-B,i} \in [1.8, 34.4] \wedge \\
 & \quad x_{DQO-D,i} \in [90.0, 269.0] \quad \xrightarrow{0.31} Classer389, \\
 \\
 r_{Classer383} : & \quad x_{Q-E,i} \in [29920.0, 54088.6] \wedge \\
 & \quad x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\
 & \quad x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\
 & \quad x_{QP-G,i} \in [188.0, 1080.0] \wedge \\
 & \quad x_{MCRT-B,i} \in [1.8, 34.4] \wedge \\
 & \quad x_{DQO-D,i} \in (269.0, 538.0] \wedge \\
 & \quad x_{DBO-E,i} \in [220.0, 987.0] \quad \xrightarrow{0.18} Classer383 \\
 \\
 r_{Classer390} : & \quad x_{Q-E,i} \in [29920.0, 54088.6] \wedge \\
 & \quad x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\
 & \quad x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\
 & \quad x_{QP-G,i} \in [188.0, 1080.0] \wedge \\
 & \quad x_{MCRT-B,i} \in [1.8, 34.4] \wedge \\
 & \quad x_{DQO-D,i} \in (269.0, 538.0] \wedge \\
 & \quad x_{DBO-E,i} \in [90.0, 220.0) \quad \xrightarrow{1.0} Classer390 \quad \}
 \end{aligned}$$

17.1.1 Interpretación final:

- Classer392: “ Q -E es bajo y QB -B es bajo y QR -G es bajo y QP -G es bajo y $MCRT$ -B es alto”
- Classer389: “ Q -E es alto y QB -B es alto y QR -G es alto y QP -G es alto y $MCRT$ -B es bajo y DQO -D no alto”
- Classer383: “ Q -E es alto y QB -B es alto y QR -G es alto y QP -G es alto y $MCRT$ -B es bajo y DQO -D alto y DBO -E es no bajo”
- Classer390: “ Q -E es alto y QB -B es alto y QR -G es alto y QP -G es alto y $MCRT$ -B es bajo y DQO -D alto y DBO -E es bajo”

17.1.2 Evaluación de la propuesta

Una vez determinados los conceptos asociados a cada clase se ha procedido a evaluar cual(es) era(n) satisfecho(s) por cada objeto de \mathcal{I} con tal de evaluar la correcta correspondencia entre la muestra y la conceptualización inducida. La tabla 17.4 muestra los resultados de calcular el soporte $Sup(r)$, ver ecuación (9.1), la cobertura relativa $CovR(r)$, ver ecuación (9.5) y la confianza $p(r)$, ver ecuación (9.2), de cada una de las reglas compuestas inducidas para cada clase de la partición final.

	Concec.	$\#\{i \in A_C^\xi\}$	$\#\{i \in A_C^\xi \cap i \in C\}$	n_c	$p(r)$	$Sup(r)$	$CovR(r)$
$r_{Classer392}$	Classer392	6	6	6	100,0%	1,5%	100,0%
$r_{Classer389}$	Classer389	227	70	70	30,8%	57,5%	100,0%
$r_{Classer383}$	Classer383	55	10	34	18,2%	13,9%	29,4%
$r_{Classer390}$	Classer390	61	61	285	100,0%	15,4%	21,4%
<i>Media</i>					62,3%		62,7%
<i>Suma</i>		349	147	395		88,4%	
$CovG_{lobal}(\mathbb{R})$							37,2%

Tabla 17.4: Evaluación: Best global concept and Close-World Assumption.

La evaluación de la regla se hace con respecto al total de objetos de la base de datos, es evidente que no habrá inconsistencias al evaluar cada regla por separado ya que los intervalos que conforman cada regla compuesta, son disjuntos entre una clase y otra debido a la forma en que se construyen los conceptos, ver ecuaciones (10.1), (10.3), (10.2), (10.4), (10.5), (10.6) y (10.7), (10.8).

En cuanto a la **Confianza**(columna $p(r)$), si se observa la Tabla 17.4, la confianza se obtiene dividiendo las celdas de la columna $\#\{i \in A_C^\xi \cap i \in C\}$ por las de la columna $\#\{i \in A_C^\xi\}$. Como hay 2 clases en donde $\#\{i \in A_C^\xi\} > \#\{i \in A_C^\xi \cap i \in C\}$ se puede concluir que hay objetos mal asignados, es decir objetos que no satisfacen el consecuente de la regla, pero que satisfacen el antecedente de esta. Este valor no se considera en el soporte ni en la cobertura relativa, pero queda de manifiesto en la confianza ya que cuando se obtienen confianzas del 100% significa que todos los objetos que satisfacen el antecedente de la regla, también satisfacen simultáneamente el antecedente y el consecuente de la regla y por lo tanto pertenecen a la clase que la regla asigna. Con lo cual el se puede concluir, en este caso, que el número de objetos correctamente asignados por clase viene dado por $\#\{i \in A_C^\xi \cap i \in C\}$ y el número de objetos mal asignados por clases, en este caso, se puede calcular $\#\{i \in A_C^\xi\} - \#\{i \in A_C^\xi \cap i \in C\}$. El porcentaje total de objetos correctamente asignados se puede obtener dividiendo 147 entre 395, 37,2%. Las confianzas en promedio rondan el 60%, lo cual se puede considerar como bueno.

En cuanto al **Soporte**, se tiene que se obtiene dividiendo $\#\{i \in A_C^\xi\}$ entre el total de objetos de la base de datos (365).

Si se considera la suma de los soportes de todas la reglas cuando éste es menor que el 100% significa que hay objetos de la base de datos sin asignar, es decir, que no satisface ninguna regla y cuando es mayor hay inconsistencias, es decir que satisface varias reglas hacia distintas clases.

También es interesante observar que el número de objetos asignados por las reglas compuestas asociadas a cada clase se puede obtener a partir de $\#\{i \in A_C^\xi\}$ y el valor total viene cuantificado por el soporte, ya que si la suma total de los soportes por clase corresponde al porcentaje de objetos asignados, cuando éste es menor que el 100% significa que hay objetos

sin asignar. En la Tabla 17.4 se puede observar que al restar a 395 la suma de la columna $\#\{i \in A_C^\xi\}$ se tiene el número de objetos sin asignar, por tanto hay 46 objetos que no han sido asignados, con lo cual se tiene un promedio del soporte cercano al 90%, esta conclusión se puede realizar sólo en el caso que no ocurran inconsistencias, es decir para soportes totales de la base de conocimiento menores o iguales al 100%.

En cuanto a la **Cobertura relativa**, si se observa la Tabla 17.4, la cobertura relativa se obtiene dividiendo cada celda de la columna $\#\{i \in A_C^\xi \cap i \in C\}$ entre la correspondiente celda de la columna $\#\{C\}$.

En estas tablas, en lugar de representar la cobertura relativa media de cada base de conocimiento, como se hace con la confianza, representamos la media de la cobertura relativa ponderada por el tamaño de cada clase, que coincide con el porcentaje global de objetos de la base de datos que se asignan correctamente, dando una idea mas ajustada de la calidad predictiva de la base de conocimiento inducida.

La relación entre cobertura relativa y confianza es también inversamente proporcional, a medida que se pierde confianza se gana cobertura relativa.

17.2 Interpretación de \mathcal{P}_4 utilizando Best local-global concept and Close-World Assumption:

1. $\xi = 2$: Así, $\mathcal{P}2_{Gi1,R1}^{EnW,G} = \{Classer392, Classer393\}$. La raíz del árbol *Classer394* tiene 2 hijos:

$$Classer394 \left\{ \begin{array}{l} Classer392 \\ Classer393. \end{array} \right.$$

2. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición $\mathcal{P}2_{Gi1,R1}^{EnW,G} = \{Classer392, Classer393\}$ se presenta en el Apéndice E.1
3. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracterizan ambas clases, $\mathcal{R}(\mathcal{P}_2) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_2)$, se presenta en el Apéndice E.2
4. Como resultado de los pasos anteriores, se consideran los siguientes sistemas de reglas con seguras $\mathcal{S}_{Classer392}(\mathcal{P}_2) = \bigcup_{k=1}^K \mathcal{S}_{Classer392}(X_k, \mathcal{P}_2)$ donde $\mathcal{S}_{Classer392}(\mathcal{P}_2) \subseteq \mathcal{S}(\mathcal{P}_2)$ y $\mathcal{S}_{Classer393}(\mathcal{P}_2) = \bigcup_{k=1}^K \mathcal{S}_{Classer393}(X_k, \mathcal{P}_2)$ donde $\mathcal{S}_{Classer393}(\mathcal{P}_2) \subseteq \mathcal{S}(\mathcal{P}_2)$:

$$\mathcal{S}_{Classer392}(\mathcal{P}_2) = \{ \begin{array}{l} r_{1, classer392}^{Q-E} : x_{Q-E,i} \in [20500.0, 23662.9] \xrightarrow{1.0} classer392 , \\ r_{1, classer392}^{QB-B} : x_{QB-B,i} \in [19883.0, 22891.0] \xrightarrow{1.0} classer392 , \\ r_{1, classer392}^{QR-G} : x_{QR-G,i} \in [17932.6, 18343.5] \xrightarrow{1.0} classer392 , \\ r_{1, classer392}^{QP-G} : x_{QP-G,i} \in [0.0, 0.0] \xrightarrow{1.0} classer392 , \\ r_{1, classer392}^{QA-G} : x_{QA-G,i} \in [96451.0, 124120.0] \xrightarrow{1.0} classer392 , \\ r_{1, classer392}^{SSV-E} : x_{SSV-E,i} \in [19.0, 30.0] \xrightarrow{1.0} classer392 , \\ r_{1, classer392}^{DQO-E} : x_{DQO-E,i} \in [27.0, 100.0] \xrightarrow{1.0} classer392 , \\ r_{2, classer392}^{DBO-E} : x_{DBO-E,i} \in [73.0, 73.0] \xrightarrow{1.0} classer392 , \\ r_{1, classer392}^{SS-D} : x_{SS-D,i} \in [40.0, 48.0] \xrightarrow{1.0} classer392 , \\ r_{1, classer392}^{SSV-D} : x_{SSV-D,i} \in [13.0, 30.0] \xrightarrow{1.0} classer392 , \\ r_{1, classer392}^{DQO-D} : x_{DQO-D,i} \in [27.0, 90.0] \xrightarrow{1.0} classer392 , \\ r_{2, classer392}^{DBO-D} : x_{DBO-D,i} \in [54.0, 54.0] \xrightarrow{1.0} classer392 , \\ r_{3, classer392}^{MLSS-B} : x_{MLSS-B,i} \in [2978.0, 32940.0] \xrightarrow{1.0} classer392 , \\ r_{3, classer392}^{MLVSS-B} : x_{MLVSS-B,i} \in [2054.0, 2100.0] \xrightarrow{1.0} classer392 , \\ r_{3, classer392}^{MCRT-B} : x_{MCRT-B,i} \in [179.8, 341.99] \xrightarrow{1.0} classer392 \end{array} \}$$

y $\mathcal{S}_{Classer393}(\mathcal{P}_2)$:

$$\mathcal{S}_{Classer393}(\mathcal{P}_2) = \{ \begin{array}{l} r_{3, classer393}^{SS-E} : x_{SS-E,i} \in (313.0, 655.0] \xrightarrow{1.0} classer393 , \\ r_{3, classer393}^{SSV-E} : x_{SSV-E,i} \in (51.0, 593.0] \xrightarrow{1.0} classer393 , \\ r_{3, classer393}^{DQO-E} : x_{DQO-E,i} \in (180.0, 1579.0] \xrightarrow{1.0} classer393 , \\ r_{1, classer393}^{DBO-E} : x_{DBO-E,i} \in [69.0, 73.0] \xrightarrow{1.0} classer393 , \\ r_{3, classer393}^{DBO-E} : x_{DBO-E,i} \in (73.0, 987.0] \xrightarrow{1.0} classer393 , \\ r_{1, classer393}^{PH-D} : x_{PH-D,i} \in [7.1, 7.3] \xrightarrow{1.0} classer393 , \\ r_{3, classer393}^{PH-D} : x_{PH-D,i} \in (7.7, 7.9] \xrightarrow{1.0} classer393 , \\ r_{3, classer393}^{SS-D} : x_{SS-D,i} \in (98.0, 192.0] \xrightarrow{1.0} classer393 , \\ r_{3, classer393}^{SSV-D} : x_{SSV-D,i} \in (42.0, 134.0] \xrightarrow{1.0} classer393 , \\ r_{3, classer393}^{DQO-D} : x_{DQO-D,i} \in (161.0, 538.0] \xrightarrow{1.0} classer393 , \\ r_{1, classer393}^{DBO-D} : x_{DBO-D,i} \in [36.0, 54.0] \xrightarrow{1.0} classer393 , \\ r_{3, classer393}^{DBO-D} : x_{DBO-D,i} \in (54.0, 274.0] \xrightarrow{1.0} classer393 , \\ r_{1, classer393}^{PH-S} : x_{PH-S,i} \in [7.0, 7.2] \xrightarrow{1.0} classer393 , \end{array} \}$$

$$\begin{aligned}
r_{3, \text{classer393}}^{PH-S} : x_{PH-S,i} \in (7.6, 8.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{SS-S} : x_{SS-S,i} \in [2.8, 5.2) &\xrightarrow{1.0} \text{classer393}, \\
r_{3, \text{classer393}}^{SS-S} : x_{SS-S,i} \in (8.0, 174.8] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{SSV-S} : x_{SSV-S,i} \in [1.6, 2.8) &\xrightarrow{1.0} \text{classer393}, \\
r_{3, \text{classer393}}^{SSV-S} : x_{SSV-S,i} \in (4.0, 134.8] &\xrightarrow{1.0} \text{classer393}, \\
r_{3, \text{classer393}}^{DQO-S} : x_{DQO-S,i} \in (66.0, 163.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{DBO-S} : x_{DBO-S,i} \in [2.0, 5.0) &\xrightarrow{1.0} \text{classer393}, \\
r_{3, \text{classer393}}^{DBO-S} : x_{DBO-S,i} \in (5.0, 84.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{3, \text{classer393}}^{V30-B} : x_{V30-B,i} \in (383.0, 770.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{MLSS-B} : x_{MLSS-B,i} \in [754.0, 2589.0) &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{MLVSS-B} : x_{MLVSS-B,i} \in [185.0, 1807.0) &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{MCRT-B} : x_{MCRT-B,i} \in [1.8, 34.4] &\xrightarrow{1.0} \text{classer393} \quad \}
\end{aligned}$$

5. Los Cuadros 22.26 y 22.27 muestran la cobertura relativa de $\mathcal{S}_{\text{classer392}}(\mathcal{P}_2)$ y $\mathcal{S}_{\text{Classer393}}(\mathcal{P}_2)$ respectivamente. Hay más de una regla con cobertura relativa máxima

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1, \text{classer392}}^{Q-E}$	6	100%
$r_{1, \text{classer392}}^{QB-B}$	6	100%
$r_{1, \text{classer392}}^{QR-G}$	6	100%
$r_{1, \text{classer392}}^{QP-G}$	6	100%
$r_{1, \text{classer392}}^{QA-G}$	3	50%
$r_{1, \text{classer392}}^{SSV-E}$	1	16,7%
$r_{1, \text{classer392}}^{DQO-E}$	3	50%
$r_{2, \text{classer392}}^{DBO-E}$	1	16,7%
$r_{1, \text{classer392}}^{SS-D}$	2	33,33%
$r_{1, \text{classer392}}^{SSV-D}$	3	50%
$r_{1, \text{classer392}}^{DQO-D}$	3	50%
$r_{2, \text{classer392}}^{DBO-D}$	1	16,7%
$r_{3, \text{classer392}}^{MLSS-B}$	2	33,33%
$r_{3, \text{classer392}}^{MLVSS-B}$	1	16,7%
$r_{3, \text{classer392}}^{MCRT-B}$	6	100%
$r_{3, \text{classer393}}^{Q-E}$	389	100%

Tabla 17.5: Cobertura relativa de $\mathcal{S}_{\text{Classer392}}(\mathcal{P}_2)$.

y $p_{sc} = 1$ en $\mathcal{S}_{\text{Classer392}}(\mathcal{P}_2)$, estas son:

$$r_{1, \text{classer392}}^{Q-E}, r_{1, \text{classer392}}^{QB-B}, r_{1, \text{classer392}}^{QR-G}, r_{1, \text{classer392}}^{QP-G}, r_{3, \text{classer392}}^{MCRT-B}$$

Con una $CovR(r)=100\%$, lo que significa que nos encontramos con 5 variables totalmente caracterizadoras, estas son; Q-E, QB-B, QR-G, QP-G y MCRT-B:

$$\begin{aligned}
r_{1, \text{classer392}}^{Q-E} : x_{Q-E,i} \in [20500.0, 23662.9] &\xrightarrow{1.0} \text{classer392} \\
r_{1, \text{classer392}}^{QB-B} : x_{QB-B,i} \in [19883.0, 22891.0] &\xrightarrow{1.0} \text{classer392} \\
r_{1, \text{classer392}}^{QR-G} : x_{QR-G,i} \in [17932.6, 18343.5] &\xrightarrow{1.0} \text{classer392} \\
r_{1, \text{classer392}}^{QP-G} : x_{QP-G,i} \in [0.0, 0.0] &\xrightarrow{1.0} \text{classer392} \\
r_{3, \text{classer392}}^{MCRT-B} : x_{MCRT-B,i} \in [179.8, 341.99] &\xrightarrow{1.0} \text{classer392}
\end{aligned}$$

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3, classer393}^{Q-E}$	389	100%
$r_{3, classer393}^{QB-B}$	389	100%
$r_{3, classer393}^{QR-G}$	389	100%
$r_{3, classer393}^{QP-G}$	389	100%
$r_{3, classer393}^{QA-G}$	384	98.7 %
$r_{3, classer393}^{FE-E}$	372	95.6%
$r_{1, classer393}^{PH-E}$	3	0.8%
$r_{3, classer393}^{PH-E}$	148	38%
$r_{1, classer393}^{SS-E}$	1	0.3%
$r_{3, classer393}^{SS-E}$	35	9%
$r_{3, classer393}^{SSV-E}$	319	82%
$r_{3, classer393}^{DQO-E}$	379	97.4%
$r_{1, classer393}^{DBO-E}$	2	0.5%
$r_{3, classer393}^{DBO-E}$	292	75.1%
$r_{1, classer393}^{PH-D}$	11	2.8%
$r_{3, classer393}^{PH-D}$	20	5.1%
$r_{3, classer393}^{SS-D}$	38	9.8%
$r_{3, classer393}^{SSV-D}$	305	78.4%
$r_{3, classer393}^{DQO-D}$	354	93.2%
$r_{1, classer393}^{DBO-D}$	4	1%
$r_{3, classer393}^{DBO-D}$	295	75.8%
$r_{1, classer393}^{PH-S}$	7	1.8%
$r_{3, classer393}^{PH-S}$	102	26.2%
$r_{1, classer393}^{SS-S}$	13	3.3%
$r_{3, classer393}^{SS-S}$	259	66.6%
$r_{1, classer393}^{SSV-S}$	9	2.3%
$r_{3, classer393}^{SSV-S}$	294	75.6%
$r_{3, classer393}^{DQO-S}$	76	19.5%
$r_{1, classer393}^{DBO-S}$	5	1.3 %
$r_{3, classer393}^{DBO-S}$	5	1.3%
$r_{3, classer393}^{V30-B}$	64	16.5%
$r_{1, classer393}^{MLSS-B}$	379	97.4%
$r_{1, classer393}^{MLVSS-B}$	369	94.9%
$r_{1, classer393}^{MCRT-B}$	389	100%

Tabla 17.6: Cobertura relativa de $\mathcal{S}_{Classer393}(\mathcal{P}_2)$.

Con lo cual también hay más de una regla con cobertura relativa máxima y $p_{sc} = 1$ en $\mathcal{S}_{Classer393}(\mathcal{P}_2)$, estas son:

$$r_{3, classer393}^{Q-E}, r_{3, classer393}^{QB-B}, r_{3, classer393}^{QR-G}, r_{3, classer393}^{QP-G}, r_{1, classer393}^{MCRT-B}.$$

Con una $CovR(r)=100\%$, lo que valida que nos encontramos con 5 variables totalmente caracterizadoras, estas son; Q-E, QB-B,QR-G, QP-G y MCRT-B:

$$\begin{aligned}
r_{3, \text{classer393}}^{Q-E} : x_{Q-E,i} \in [29920.0, 54088.6] &\xrightarrow{1.0} \text{classer393} \\
r_{3, \text{classer393}}^{QB-B} : x_{QB-B,i} \in [29397.3, 52244.6] &\xrightarrow{1.0} \text{classer393} \\
r_{3, \text{classer393}}^{QR-G} : x_{QR-G,i} \in [26218.0, 49527.0] &\xrightarrow{1.0} \text{classer393} \\
r_{3, \text{classer393}}^{QP-G} : x_{QP-G,i} \in [188.0, 1080.0] &\xrightarrow{1.0} \text{classer393} \\
r_{1, \text{classer393}}^{MCRT-B} : x_{MCRT-B,i} \in [1.8, 34.4] &\xrightarrow{1.0} \text{classer393}
\end{aligned}$$

Así se tiene que:

$$\begin{aligned}
A_{\text{classer392}}^{2,Q-E} &= "x_{Q-E,i} \in [20500.0, 23662.9]" \\
A_{\text{classer392}}^{2,QB-B} &= "x_{QB-B,i} \in [19883.0, 22891.0]" \\
A_{\text{classer392}}^{2,QR-G} &= "x_{QR-G,i} \in [17932.6, 18343.5]" \\
A_{\text{classer392}}^{2,QP-G} &= "x_{QP-G,i} \in [0.0, 0.0]" \\
A_{\text{classer392}}^{2,MCRT-B} &= "x_{MCRT-B,i} \in [179.8, 341.99]" \\
A_{\text{classer393}}^{2,Q-E} &= "x_{Q-E,i} \in [29920.0, 54088.6]" \\
A_{\text{classer393}}^{2,QB-B} &= "x_{QB-B,i} \in [29397.3, 52244.6]" \\
A_{\text{classer393}}^{2,QR-G} &= "x_{QR-G,i} \in [26218.0, 49527.0]" \\
A_{\text{classer393}}^{2,QP-G} &= "x_{QP-G,i} \in [188.0, 1080.0]" \\
A_{\text{classer393}}^{2,MCRT-B} &= "x_{MCRT-B,i} \in [1.8, 34.4]"
\end{aligned}$$

En este caso hay mas de una regla, por tanto:

$$\begin{aligned}
\bullet A_{\text{classer392}}^2 &= A_{\text{classer392}}^{2,Q-E} \wedge A_{\text{classer392}}^{2,QB-B} \wedge A_{\text{classer392}}^{2,QR-G} \wedge A_{\text{classer392}}^{2,QP-G} \wedge A_{\text{classer392}}^{2,MCRT-B} \\
&= "x_{Q-E,i} \in [20500.0, 23662.9]" \wedge "x_{QB-B,i} \in [19883.0, 22891.0]" \wedge \\
&\quad "x_{QR-G,i} \in [17932.6, 18343.5]" \wedge "x_{QP-G,i} \in [0.0, 0.0]" \wedge \\
&\quad "x_{MCRT-B,i} \in [179.8, 341.99]" \\
\bullet A_{\text{classer393}}^2 &= A_{\text{classer393}}^{2,Q-E} \wedge A_{\text{classer393}}^{2,QB-B} \wedge A_{\text{classer393}}^{2,QR-G} \wedge A_{\text{classer393}}^{2,QP-G} \wedge A_{\text{classer393}}^{2,MCRT-B} \\
&= "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\
&\quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\
&\quad "x_{MCRT-B,i} \in [1.8, 34.4]"
\end{aligned}$$

6. Asociando una regla a cada clase,

$$\begin{aligned}
\mathbb{R}(\mathcal{P}_2) \{ & r_{\text{Classer392}} : x_{Q-E,i} \in [20500.0, 23662.9] \wedge \\
& x_{QB-B,i} \in [19883.0, 22891.0] \wedge \\
& x_{QR-G,i} \in [17932.6, 18343.5] \wedge \\
& x_{QP-G,i} \in [0.0, 0.0] \wedge \\
& x_{MCRT-B,i} \in [179.8, 341.99] \xrightarrow{1.0} \text{Classer392},
\end{aligned}$$

$$\begin{aligned}
r_{\text{Classer393}} : & x_{Q-E,i} \in [29920.0, 54088.6] \wedge \\
& x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\
& x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\
& x_{QP-G,i} \in [188.0, 1080.0] \wedge \\
& x_{MCRT-B,i} \in [1.8, 34.4] \xrightarrow{1.0} \text{Classer393} \}
\end{aligned}$$

Que correspondería a la conceptualization:

– Classer392: “ $Q\text{-}E$ es bajo y $QB\text{-}B$ es bajo y $QR\text{-}G$ es bajo y $QP\text{-}G$ es bajo y $MCRT\text{-}B$ es alto”

– Classer393: “ $Q\text{-}E$ es alto y $QB\text{-}B$ es alto y $QR\text{-}G$ es alto y $QP\text{-}G$ es alto y $MCRT\text{-}B$ es bajo”

7. $\xi = 3$: Así, $\mathcal{P}_3^{EnW,G} = \{\text{Classer392}, \text{Classer389}, \text{Classer391}\}$. La clase *Classer392* es la que se divide en 2 hijos y la clase *Classer393* es la que ya estaba en la partición anterior:

$$\text{Classer393} \left\{ \begin{array}{l} \text{Classer389} \\ \text{Classer391} \end{array} \right.$$

Así $C_i^3 = \text{Classer389}$; $C_j^3 = \text{Classer391}$; $C_t^2 = \text{Classer393}$.

8. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_3^* = \{\text{Classer389}, \text{Classer391}\}$, donde $\mathcal{P}_3^* \subset \mathcal{P}_3^{EnW,G}$, se presenta en el Apéndice E.3.
9. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_3^*)$.
10. Como resultado de los pasos anteriores, se considera una único sistema de reglas $\mathcal{R}(\mathcal{P}_3^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_3^*)$ que se se presenta en el Apéndice E.4 y se trabaja con el sistema de reglas seguras $\mathcal{S}_{\text{Classer389}}(\mathcal{P}_3^*) = \bigcup_{k=1}^K \mathcal{S}_{\text{Classer389}}(X_k, \mathcal{P}_3^*)$ donde $\mathcal{S}_{\text{Classer389}}(\mathcal{P}_3^*) \subseteq \mathcal{S}(\mathcal{P}_3^*)$ la cual es:

$$\mathcal{S}_{\text{Classer389}}(\mathcal{P}_3^*) = \{ \begin{array}{ll} r_{3,\text{classer389}}^{Q-E} : x_{Q-E,i} \in (52255.8, 54088.6] & \xrightarrow{1.0} \text{classer389} , \\ r_{3,\text{classer389}}^{QB-B} : x_{QB-B,i} \in (49695.8, 52244.6] & \xrightarrow{1.0} \text{classer389} , \\ r_{1,\text{classer389}}^{QR-G} : x_{QR-G,i} \in [26218.0, 27351.0) & \xrightarrow{1.0} \text{classer389} , \\ r_{1,\text{classer389}}^{QP-G} : x_{QP-G,i} \in [188.0, 327.6) & \xrightarrow{1.0} \text{classer389} , \\ r_{1,\text{classer389}}^{QA-G} : x_{QA-G,i} \in [124120.0, 136371.0) & \xrightarrow{1.0} \text{classer389} , \\ r_{3,\text{classer389}}^{PH-E} : x_{PH-E,i} \in (7.9, 8.0] & \xrightarrow{1.0} \text{classer389} , \\ r_{1,\text{classer389}}^{SS-E} : x_{SS-E,i} \in [62.0, 82.0) & \xrightarrow{1.0} \text{classer389} , \\ r_{1,\text{classer389}}^{SSV-E} : x_{SSV-E,i} \in [30.0, 60.0) & \xrightarrow{1.0} \text{classer389} , \\ r_{1,\text{classer389}}^{DQO-E} : x_{DQO-E,i} \in [100.0, 158.0) & \xrightarrow{1.0} \text{classer389} , \\ r_{1,\text{classer389}}^{DBO-E} : x_{DBO-E,i} \in [69.0, 90.0) & \xrightarrow{1.0} \text{classer389} , \\ r_{1,\text{classer389}}^{SS-D} : x_{SS-D,i} \in [48.0, 63.0) & \xrightarrow{1.0} \text{classer389} , \\ r_{1,\text{classer389}}^{SSV-D} : x_{SSV-D,i} \in [30.0, 47.0) & \xrightarrow{1.0} \text{classer389} , \\ r_{1,\text{classer389}}^{DBO-D} : x_{DBO-D,i} \in [36.0, 56.0) & \xrightarrow{1.0} \text{classer389} , \\ r_{1,\text{classer389}}^{DBO-S} : x_{DBO-S,i} \in [2.0, 4.0) & \xrightarrow{1.0} \text{classer389} , \\ r_{3,\text{classer389}}^{MLSS-B} : x_{MLSS-B,i} \in (2696.0, 2978.0] & \xrightarrow{1.0} \text{classer389} , \\ r_{3,\text{classer389}}^{MCRT-B} : x_{MCRT-B,i} \in (28.8, 34.4] & \xrightarrow{1.0} \text{classer389} \end{array} \}$$

y $\mathcal{S}_{\text{Classer391}}(\mathcal{P}_3^*) \subseteq \mathcal{S}(\mathcal{P}_3^*)$ la cual es:

$$\begin{aligned} \mathcal{S}_{\text{Classer391}}(\mathcal{P}_3^*) = \{ & r_{1,\text{classer391}}^{Q-E} : x_{Q-E,i} \in [29920.0, 30592.2] \xrightarrow{1.0} \text{classer391}, \\ & r_{1,\text{classer391}}^{QB-B} : x_{QB-B,i} \in [29397.3, 29936.8] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{QR-G} : x_{QR-G,i} \in (43298.1, 49527.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{QP-G} : x_{QP-G,i} \in (866.7, 1080.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{QA-G} : x_{QA-G,i} \in (324470.0, 367840.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{FE-E} : x_{FE-E,i} \in (65.6, 89.8] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{SS-E} : x_{SS-E,i} \in (266.0, 655.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{SSV-E} : x_{SSV-E,i} \in (193.0, 593.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{DQO-E} : x_{DQO-E,i} \in (595.0, 1579.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{DBO-E} : x_{DBO-E,i} \in (258.0, 987.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{PH-D} : x_{PH-D,i} \in (7.8, 7.9] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{SS-D} : x_{SS-D,i} \in (136.0, 192.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{SSV-D} : x_{SSV-D,i} \in (99.0, 134.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{1,\text{classer391}}^{DQO-D} : x_{DQO-D,i} \in [90.0, 100.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{DQO-D} : x_{DQO-D,i} \in (269.0, 538.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{DBO-D} : x_{DBO-D,i} \in (171.0, 274.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{1,\text{classer391}}^{SS-S} : x_{SS-S,i} \in [2.8, 3.2] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{SS-S} : x_{SS-S,i} \in (20.0, 174.8] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{SSV-S} : x_{SSV-S,i} \in (18.0, 134.8] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{DQO-S} : x_{DQO-S,i} \in (94.0, 163.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{DBO-S} : x_{DBO-S,i} \in (26.0, 84.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{1,\text{classer391}}^{V30-B} : x_{V30-B,i} \in [77.0, 115.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{V30-B} : x_{V30-B,i} \in (380.0, 770.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{1,\text{classer391}}^{MLSS-B} : x_{MLSS-B,i} \in [754.0, 846.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{1,\text{classer391}}^{MLVSS-B} : x_{MLVSS-B,i} \in [185.0, 684.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{3,\text{classer391}}^{MLVSS-B} : x_{MLVSS-B,i} \in (1921.0, 2054.0] \xrightarrow{1.0} \text{classer391}, \\ & r_{1,\text{classer391}}^{MCRT-B} : x_{MCRT-B,i} \in [1.8, 6.9] \xrightarrow{1.0} \text{classer391} \} \end{aligned}$$

11. Los Cuadros 22.28 y 22.29 muestran la coberturas relativas de las reglas de $\mathcal{S}_{\text{Classer389}}(\mathcal{P}_3^*)$ y $\mathcal{S}_{\text{Classer391}}(\mathcal{P}_3^*)$ respectivamente.

La regla con mayor cobertura relativa y $p_{sc} = 1$ de $\mathcal{S}_{\text{Classer391}}(\mathcal{P}_3^*)$ es $r_{3,\text{classer391}}^{DQO-D}$, con una cobertura relativa $CovR = 45,63\%$ y la regla con mayor cobertura relativa de $\mathcal{S}_{\text{Classer389}}(\mathcal{P}_3^*)$ es $r_{1,\text{classer389}}^{SSV-D}$, con una cobertura relativa $CovR = 35,71\%$:

$$r_{3,\text{classer391}}^{DQO-D} : x_{DQO-D,i} \in (269.0, 538.0] \xrightarrow{1.0} \text{classer391}$$

$$r_{1,\text{classer389}}^{SSV-D} : x_{SSV-D,i} \in [30.0, 47.0] \xrightarrow{1.0} \text{classer389}$$

Así,

$$\bullet A_{\text{classer391}}^{3,DQO-D} = "x_{DQO-D,i} \in (269.0, 538.0)"$$

$$\bullet A_{\text{classer389}}^{3,SSV-D} = "x_{SSV-D,i} \in [30.0, 47.0]"$$

En este caso no hay empate y por tanto:

$$\bullet A_{\text{classer391}}^{*3} = A_{\text{classer391}}^{3,DQO-D} \vee \neg A_{\text{classer389}}^{3,SSV-D} \\ = "x_{DQO-D,i} \in (269.0, 538.0)" \vee "x_{SSV-D,i} \in [47.0, 134.0]"$$

$$\bullet A_{\text{classer389}}^{*3} = A_{\text{classer389}}^{3,SSV-D} \vee \neg A_{\text{classer391}}^{3,DQO-D} \\ = "x_{SSV-D,i} \in [30.0, 47.0]" \vee "x_{DQO-D,i} \in [90.0, 269.0]"$$

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3, classer389}^{Q-E}$	3	4,28%
$r_{3, classer389}^{QB-B}$	3	4,28%
$r_{1, classer389}^{QR-G}$	2	2,85%
$r_{1, classer389}^{QP-G}$	2	2,85%
$r_{1, classer389}^{QA-G}$	2	2,85%
$r_{3, classer389}^{PH-E}$	1	1,42%
$r_{1, classer389}^{SS-E}$	1	1,42%
$r_{1, classer389}^{SSV-E}$	2	2,85%
$r_{1, classer389}^{DQO-E}$	1	1,42%
$r_{1, classer389}^{DBO-E}$	5	7,14%
$r_{1, classer389}^{SS-D}$	24	34,28%
$r_{1, classer389}^{SSV-D}$	25	35,71%
$r_{1, classer389}^{DBO-D}$	5	7,14%
$r_{1, classer389}^{DBO-S}$	2	2,85%
$r_{3, classer389}^{MLSS-B}$	1	1,42%
$r_{3, classer389}^{MCRT-B}$	1	1,42%

Tabla 17.7: Cobertura relativa de $\mathcal{S}_{Classer389}(\mathcal{P}_3^*)$.

12. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

- $A_{classer392}^3 = A_{classer392}^2$

$$\begin{aligned}
 A_{classer392}^3 &= A_{classer392}^{2,Q-E} \wedge A_{classer392}^{2,QB-B} \wedge A_{classer392}^{2,QR-G} \wedge A_{classer392}^{2,QP-G} \wedge A_{classer393}^{2,MCRT-B} \\
 &= "x_{Q-E,i} \in [20500.0, 23662.9]" \wedge "x_{QB-B,i} \in [19883.0, 22891.0]" \wedge \\
 &\quad "x_{QR-G,i} \in [17932.6, 18343.5]" \wedge "x_{QP-G,i} \in [0.0, 0.0]" \wedge \\
 &\quad "x_{MCRT-B,i} \in [179.8, 341.99]"
 \end{aligned}$$

- $A_{classer389}^3 = A_{classer393}^2 \wedge A_{classer389}^{*3}$

$$\begin{aligned}
 A_{classer389}^3 &= A_{classer393}^{2,Q-E} \wedge A_{classer393}^{2,QB-B} \wedge A_{classer393}^{2,QR-G} \wedge A_{classer393}^{2,QP-G} \wedge A_{classer393}^{2,MCRT-B} \wedge \\
 &\quad A_{classer389}^{3,SSV-D} \vee \neg A_{classer391}^{3,DQO-D} \\
 &= "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\
 &\quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\
 &\quad "x_{MCRT-B,i} \in [1.8, 34.4]" \wedge ("x_{SSV-D,i} \in [30.0, 47.0]) \vee \\
 &\quad "x_{DQO-D,i} \in [90.0, 269.0]")
 \end{aligned}$$

- $A_{classer391}^3 = A_{classer393}^2 \wedge A_{classer391}^{*3}$

$$\begin{aligned}
 A_{classer391}^3 &= A_{classer393}^{2,Q-E} \wedge A_{classer393}^{2,QB-B} \wedge A_{classer393}^{2,QR-G} \wedge A_{classer393}^{2,QP-G} \wedge A_{classer393}^{2,MCRT-B} \wedge \\
 &\quad A_{classer391}^{3,DQO-D} \vee \neg A_{classer389}^{3,SSV-D}
 \end{aligned}$$

$$\begin{aligned}
&= "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\
&\quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\
&\quad "x_{MCRT-B,i} \in [1.8, 34.4]" \wedge ("x_{SSV-D,i} \in [47.0, 134.0]" \vee \\
&\quad "x_{DQO-D,i} \in (269.0, 538.0]")
\end{aligned}$$

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,classer391}^{Q-E}$	1	0,3125 %
$r_{1,classer391}^{QB-B}$	1	0,3125 %
$r_{3,classer391}^{QR-G}$	64	20 %
$r_{3,classer391}^{QP-G}$	36	11,25 %
$r_{3,classer391}^{QA-G}$	8	2,5 %
$r_{3,classer391}^{FE-E}$	4	1,25 %
$r_{3,classer391}^{SS-E}$	60	18,75 %
$r_{3,classer391}^{SSV-E}$	70	21,875 %
$r_{3,classer391}^{DQO-E}$	52	16,25 %
$r_{3,classer391}^{DBO-E}$	64	20 %
$r_{3,classer391}^{PH-D}$	1	0,3125 %
$r_{3,classer391}^{SS-D}$	8	2,5 %
$r_{3,classer391}^{SSV-D}$	8	2,5 %
$r_{1,classer391}^{DQO-D}$	1	0,3125 %
$r_{3,classer391}^{DQO-D}$	146	45,625 %
$r_{3,classer391}^{DBO-D}$	25	7,8125 %
$r_{1,classer391}^{SS-S}$	1	0,3125 %
$r_{3,classer391}^{SS-S}$	59	18,4375 %
$r_{3,classer391}^{SSV-S}$	42	13,125 %
$r_{3,classer391}^{DQO-S}$	26	8,125 %
$r_{3,classer391}^{DBO-S}$	55	17,1875 %
$r_{1,classer391}^{V30-B}$	6	1,875 %
$r_{3,classer391}^{V30-B}$	64	20 %
$r_{1,classer391}^{MLSS-B}$	2	0,625 %
$r_{1,classer391}^{MLVSS-B}$	3	0,9375 %
$r_{3,classer391}^{MLVSS-B}$	6	1,875 %
$r_{1,classer391}^{MCRT-B}$	27	8,4375 %

Tabla 17.8: Cobertura relativa de $\mathcal{S}_{Classer391}(\mathcal{P}_3^*)$.

Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$\begin{aligned}
p_{sClasser389} &= \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{70}{227} = 30.84\% \\
p_{sClasser391} &= \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{273}{320} = 85.31\%
\end{aligned}$$

$$\begin{aligned}
\mathbb{R}(\mathcal{P}_3) = \{ & \quad r_{Classer392} : \quad x_{Q-E,i} \in [20500.0, 23662.9] \wedge \\
& \quad x_{QB-B,i} \in [19883.0, 22891.0] \wedge \\
& \quad x_{QR-G,i} \in [17932.6, 18343.5] \wedge \\
& \quad x_{QP-G,i} \in [0.0, 0.0] \wedge \\
& \quad x_{MCRT-B,i} \in [179.8, 341.99] \quad \xrightarrow{1.0} Classer392 \quad \} \\
\\
r_{Classer389} : & \quad x_{Q-E,i} \in [29920.0, 54088.6] \wedge \\
& \quad x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\
& \quad x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\
& \quad x_{QP-G,i} \in [188.0, 1080.0] \wedge \\
& \quad x_{MCRT-B,i} \in [1.8, 34.4] \wedge \\
& \quad x_{SSV-D,i} \in [30.0, 47.0] \vee \\
& \quad x_{DQO-D,i} \in [90.0, 269.0] \quad \xrightarrow{0.31} Classer389, \\
\\
r_{Classer391} : & \quad x_{Q-E,i} \in [29920.0, 54088.6] \wedge \\
& \quad x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\
& \quad x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\
& \quad x_{QP-G,i} \in [188.0, 1080.0] \wedge \\
& \quad x_{MCRT-B,i} \in [1.8, 34.4] \wedge \\
& \quad x_{SSV-D,i} \in [47.0, 134.0] \vee \\
& \quad x_{DQO-D,i} \in (269.0, 538.0] \quad \xrightarrow{0.85} Classer391 \quad \}
\end{aligned}$$

Que correspondería a la conceptualización:

- Classer392: “ $Q-E$ es bajo y $QB-B$ es bajo y $QR-G$ es bajo y $QP-G$ es bajo y $MCRT-B$ es alto”
- Classer389: “($Q-E$ es alto y $QB-B$ es alto y $QR-G$ es alto y $QP-G$ es alto y $MCRT-B$ es bajo) y ($SSV-D$ es bajo o $DQO-D$ no alto)”
- Classer391: “($Q-E$ es alto y $QB-B$ es alto y $QR-G$ es alto y $QP-G$ es alto y $MCRT-B$ es bajo) y ($SSV-D$ no es bajo o $DQO-D$ alto)”

13. $\xi = 4$: Así, $\mathcal{P}^{EnW,G}_{Gi1,R1} = \{Classer392, Classer389, Classer390, Classer383\}$. La clase $Classer393$ de $\mathcal{P}^{EnW,G}_{Gi1,R1}$ tiene 2 hijos:

$$Classer391 \left\{ \begin{array}{l} Classer383 \\ Classer390 \end{array} \right.$$

Así $C_i^4 = Classer383$; $C_j^4 = Classer390$; $C_t^3 = Classer391$

14. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_4^* = \{Classer383, Classer390\}$, donde $\mathcal{P}_4^* \subseteq \mathcal{P}^{EnW,G}_{Gi1,R1}$, se presenta en el Apéndice E.5.
15. Como resultado de los pasos anteriores, se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_4^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_4^*)$ que se presenta en el Apéndice E.6 y se trabaja con los sistemas de reglas seguros:

$$\mathcal{S}_{Classer383}(\mathcal{P}_4^*) = \bigcup_{k=1}^K \mathcal{S}_{Classer383}(X_k, \mathcal{P}_4^*) \text{ donde } \mathcal{S}_{Classer383}(\mathcal{P}_4^*) \subseteq \mathcal{S}(\mathcal{P}_4^*) \text{ el cual es:}$$

$$\mathcal{S}_{Classer383}(\mathcal{P}_4^*) = \{ \begin{array}{l} r_{3, classer383}^{SS-E} : x_{SS-E,i} \in (480.0, 655.0] \xrightarrow{1.0} classer383 , \\ r_{3, classer383}^{SSV-E} : x_{SSV-E,i} \in (336.0, 593.0] \xrightarrow{1.0} classer383 , \\ r_{3, classer383}^{DQO-E} : x_{DQO-E,i} \in (1279.0, 1579.0] \xrightarrow{1.0} classer383 , \\ r_{3, classer383}^{DBO-E} : x_{DBO-E,i} \in (382.0, 987.0] \xrightarrow{1.0} classer383 , \\ r_{1, classer383}^{MLVSS-B} : x_{MLVSS-B,i} \in [185.0, 611.0] \xrightarrow{1.0} classer383 , \\ r_{1, classer383}^{MCRT-B} : x_{MCRT-B,i} \in [1.8, 6.2] \xrightarrow{1.0} classer383 \end{array} \}$$

y con $\mathcal{S}_{Classer390}(\mathcal{P}_4^*) = \bigcup_{k=1}^k \mathcal{S}_{Classer390}(X_k, \mathcal{P}_4^*)$ donde $\mathcal{S}_{Classer390}(\mathcal{P}_4^*) \subseteq \mathcal{S}(\mathcal{P}_4^*)$ el cual es:

$$\mathcal{S}_{Classer390}(\mathcal{P}_4^*) = \{ \begin{array}{l} r_{1, classer390}^{Q-E} : x_{Q-E,i} \in [29920.0, 34284.4] \xrightarrow{1.0} classer390 , \\ r_{3, classer390}^{Q-E} : x_{Q-E,i} \in (50500.5, 52255.8] \xrightarrow{1.0} classer390 , \\ r_{1, classer390}^{QB-B} : x_{QB-B,i} \in [29397.3, 33549.4] \xrightarrow{1.0} classer390 , \\ r_{3, classer390}^{QB-B} : x_{QB-B,i} \in (39000.0, 49695.8] \xrightarrow{1.0} classer390 , \\ r_{1, classer390}^{QR-G} : x_{QR-G,i} \in [27351.0, 28343.8] \xrightarrow{1.0} classer390 , \\ r_{3, classer390}^{QR-G} : x_{QR-G,i} \in (44568.6, 49527.0] \xrightarrow{1.0} classer390 , \\ r_{1, classer390}^{QP-G} : x_{QP-G,i} \in [327.6, 385.9] \xrightarrow{1.0} classer390 , \\ r_{3, classer390}^{QP-G} : x_{QP-G,i} \in (831.1, 1080.0] \xrightarrow{1.0} classer390 , \\ r_{1, classer390}^{QA-G} : x_{QA-G,i} \in [136371.0, 156320.0] \xrightarrow{1.0} classer390 , \\ r_{3, classer390}^{QA-G} : x_{QA-G,i} \in (331990.0, 367840.0] \xrightarrow{1.0} classer390 , \\ r_{3, classer390}^{FE-E} : x_{FE-E,i} \in (63.3, 89.8] \xrightarrow{1.0} classer390 , \\ r_{1, classer390}^{PH-E} : x_{PH-E,i} \in [7.2, 7.3] \xrightarrow{1.0} classer390 , \\ r_{3, classer390}^{PH-E} : x_{PH-E,i} \in (7.8, 7.9] \xrightarrow{1.0} classer390 , \\ r_{1, classer390}^{SS-E} : x_{SS-E,i} \in [82.0, 114.0] \xrightarrow{1.0} classer390 , \\ r_{1, classer390}^{SSV-E} : x_{SSV-E,i} \in [60.0, 92.0] \xrightarrow{1.0} classer390 , \\ r_{1, classer390}^{DQO-E} : x_{DQO-E,i} \in [158.0, 414.0] \xrightarrow{1.0} classer390 , \\ r_{1, classer390}^{DBO-E} : x_{DBO-E,i} \in [90.0, 220.0] \xrightarrow{1.0} classer390 , \\ r_{1, classer390}^{PH-D} : x_{PH-D,i} \in [7.2, 7.3] \xrightarrow{1.0} classer390 , \\ r_{3, classer390}^{PH-D} : x_{PH-D,i} \in (7.8, 7.9] \xrightarrow{1.0} classer390 \end{array} \}$$

16. Los Cuadros 17.9 y 17.10 muestran la cobertura relativa de las reglas de $\mathcal{S}_{Classer383}(\mathcal{P}_4^*)$ y $\mathcal{S}_{Classer390}(\mathcal{P}_4^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3, classer383}^{SSV-E}$	4	11,76%
$r_{3, classer383}^{DQO-E}$	1	2,94%
$r_{3, classer383}^{DBO-E}$	7	20,58%
$r_{1, classer383}^{MLVSS-B}$	2	5,88%
$r_{1, classer383}^{MCRT-B}$	1	2,94%

Tabla 17.9: Cobertura relativa de $\mathcal{S}_{Classer383}(\mathcal{P}_4^*)$.

La regla con mayor cobertura relativa de $\mathcal{S}_{Classer390}(\mathcal{P}_4^*)$ es $r_{1, classer390}^{DBO-E}$, con una cobertura relativa $CovR = 44,21\%$

Y la regla con mayor cobertura relativa de $\mathcal{S}_{Classer383}(\mathcal{P}_4^*)$ es $r_{3, classer383}^{DBO-E}$, con una cobertura relativa $CovR = 20,58\%$.

Así;

$$r_{1, \text{classer}390}^{DBO-E} : x_{DBO-E,i} \in [90.0, 220.0] \xrightarrow{1.0} \text{classer}390$$

$$r_{3, \text{classer}383}^{DBO-E} : x_{DBO-E,i} \in (382.0, 987.0] \xrightarrow{1.0} \text{classer}383$$

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1, \text{classer}390}^{Q-E}$	9	3,15%
$r_{3, \text{classer}390}^{Q-E}$	1	0,35%
$r_{1, \text{classer}390}^{QB-B}$	8	2,80%
$r_{3, \text{classer}390}^{QB-B}$	60	21,06%
$r_{1, \text{classer}390}^{QR-G}$	1	0,35%
$r_{3, \text{classer}390}^{QR-G}$	13	4,56%
$r_{1, \text{classer}390}^{QP-G}$	7	2,45%
$r_{3, \text{classer}390}^{QP-G}$	56	19,64%
$r_{1, \text{classer}390}^{QA-G}$	7	2,45%
$r_{3, \text{classer}390}^{QA-G}$	2	0,70%
$r_{3, \text{classer}390}^{FE-E}$	7	2,45%
$r_{1, \text{classer}390}^{PH-E}$	1	0,35%
$r_{3, \text{classer}390}^{PH-E}$	8	2,80%
$r_{1, \text{classer}390}^{SS-E}$	4	1,40%
$r_{1, \text{classer}390}^{SSV-E}$	7	2,45%
$r_{1, \text{classer}390}^{DQO-E}$	113	39,64%
$r_{1, \text{classer}390}^{DBO-E}$	126	44,21%
$r_{1, \text{classer}390}^{PH-D}$	3	1,05%
$r_{3, \text{classer}390}^{PH-D}$	1	0,35%

Tabla 17.10: Cobertura relativa de $\mathcal{S}_{\text{Classer}390}(\mathcal{P}_4^*)$.

Como la variable es la misma. La regla con mayor cobertura relativa de $\mathcal{S}(\mathcal{P}_4^*)$ es $r_{1, \text{classer}390}^{DBO-E}$, con una cobertura relativa $CovR = 44,21\%$.

Por tanto:

- $A_{\text{classer}390}^{*4} = A_{\text{classer}390}^{4,DBO-E} = "x_{DBO-E,i} \in [90.0, 220.0]"$
- $A_{\text{classer}383}^{*4} = A_{\text{classer}383}^{4,DBO-E} = -A_{\text{classer}390}^{4,DBO-E} = "x_{DBO-E,i} \in [220.0, 987.0]"$

17. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

- $A_{\text{classer}392}^4 = A_{\text{classer}392}^3 = A_{\text{classer}392}^2$

$$A_{\text{classer}392}^4 = A_{\text{classer}392}^{2,Q-E} \wedge A_{\text{classer}392}^{2,QB-B} \wedge A_{\text{classer}392}^{2,QR-G} \wedge A_{\text{classer}392}^{2,QP-G} \wedge A_{\text{classer}392}^{2,MCRT-B}$$

$$= "x_{Q-E,i} \in [20500.0, 23662.9]" \wedge "x_{QB-B,i} \in [19883.0, 22891.0]" \wedge$$

$$"x_{QR-G,i} \in [17932.6, 18343.5]" \wedge "x_{QP-G,i} \in [0.0, 0.0]" \wedge$$

$$"x_{MCRT-B,i} \in [179.8, 341.99]"$$
- $A_{\text{classer}389}^3 = A_{\text{classer}393}^2 \wedge A_{\text{classer}389}^{*3}$

$$A_{\text{classer}389}^3 = A_{\text{classer}393}^{2,Q-E} \wedge A_{\text{classer}393}^{2,QB-B} \wedge A_{\text{classer}393}^{2,QR-G} \wedge A_{\text{classer}393}^{2,QP-G} \wedge A_{\text{classer}393}^{2,MCRT-B} \wedge$$

- $$\begin{aligned}
& A_{classer389}^{3,SSV-D} \vee \neg A_{classer391}^{3,DQO-D} \\
& = "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\
& \quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\
& \quad "x_{MCRT-B,i} \in [1.8, 34.4]" \wedge ("x_{SSV-D,i} \in [30.0, 47.0]" \vee \\
& \quad "x_{DQO-D,i} \in [90.0, 269.0]")
\end{aligned}$$
- $A_{classer383}^4 = A_{classer391}^3 \wedge A_{classer383}^{*4} = A_{classer393}^2 \wedge A_{classer391}^{*3} \wedge A_{classer383}^{*4}$
$$\begin{aligned}
& A_{classer383}^4 = A_{classer393}^{2,Q-E} \wedge A_{classer393}^{2,QB-B} \wedge A_{classer393}^{2,QR-G} \wedge A_{classer393}^{2,QP-G} \wedge A_{classer393}^{2,MCRT-B} \wedge \\
& A_{classer391}^{3,SSV-D} \vee A_{classer391}^{3,DQO-D} \wedge A_{classer383}^{4,DBO-E} \\
& = "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\
& \quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\
& \quad "x_{MCRT-B,i} \in [1.8, 34.4]" \wedge ("x_{SSV-D,i} \in [47.0, 134.0]" \vee \\
& \quad "x_{DQO-D,i} \in (269.0, 538.0]") \wedge "x_{DBO-E,i} \in [220.0, 987.0]"
\end{aligned}$$
 - $A_{classer390}^4 = A_{classer391}^3 \wedge A_{classer390}^{*4} = A_{classer393}^2 \wedge A_{classer391}^{*3} \wedge A_{classer390}^{*4}$
$$\begin{aligned}
& A_{classer390}^4 = A_{classer393}^{2,Q-E} \wedge A_{classer393}^{2,QB-B} \wedge A_{classer393}^{2,QR-G} \wedge A_{classer393}^{2,QP-G} \wedge A_{classer393}^{2,MCRT-B} \wedge \\
& A_{classer391}^{3,SSV-D} \vee A_{classer391}^{3,DQO-D} \wedge A_{classer390}^{4,DBO-E} \\
& = "x_{Q-E,i} \in [29920.0, 54088.6]" \wedge "x_{QB-B,i} \in [29397.3, 52244.6]" \wedge \\
& \quad "x_{QR-G,i} \in [26218.0, 49527.0]" \wedge "x_{QP-G,i} \in [188.0, 1080.0]" \wedge \\
& \quad "x_{MCRT-B,i} \in [1.8, 34.4]" \wedge ("x_{SSV-D,i} \in [47.0, 134.0]" \vee \\
& \quad "x_{DQO-D,i} \in (269.0, 538.0)") \wedge "x_{DBO-E,i} \in [90.0, 220.0]"
\end{aligned}$$

Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$p_{scClasser383} = \frac{\text{card}\{i \in C \mid t_q \wedge A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{27}{108} = 25\%$$

$$p_{scClasser390} = \frac{\text{card}\{i \in C \mid t_q \wedge A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{70}{227} = 30.84\%$$

$$\begin{aligned}
 \mathbb{R}(\mathcal{P}_4) = \{ & \quad r_{Classer392} : \quad x_{QE,i} \in [20500.0, 23662.9] \wedge \\
 & \quad x_{QB-B,i} \in [19883.0, 22891.0] \wedge \\
 & \quad x_{QR-G,i} \in [17932.6, 18343.5] \wedge \\
 & \quad x_{QP-G,i} \in [0.0, 0.0] \wedge \\
 & \quad x_{MCRT-B,i} \in [179.8, 341.99] \quad \xrightarrow{1.0} Classer392 \quad \} \\
 \\
 r_{Classer389} : & \quad x_{QE,i} \in [29920.0, 54088.6] \wedge \\
 & \quad x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\
 & \quad x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\
 & \quad x_{QP-G,i} \in [188.0, 1080.0] \wedge \\
 & \quad x_{MCRT-B,i} \in [1.8, 34.4] \wedge \\
 & \quad x_{SSV-D,i} \in [30.0, 47.0] \vee \\
 & \quad x_{DQO-D,i} \in [90.0, 269.0] \quad \xrightarrow{0.31} Classer389, \\
 \\
 r_{Classer383} : & \quad x_{QE,i} \in [29920.0, 54088.6] \wedge \\
 & \quad x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\
 & \quad x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\
 & \quad x_{QP-G,i} \in [188.0, 1080.0] \wedge \\
 & \quad x_{MCRT-B,i} \in [1.8, 34.4] \wedge \\
 & \quad x_{SSV-D,i} \in [47.0, 134.0] \vee \\
 & \quad x_{DQO-D,i} \in (269.0, 538.0] \wedge \\
 & \quad x_{DBO-E,i} \in [220.0, 987.0] \quad \xrightarrow{0.25} Classer383 \\
 \\
 r_{Classer390} : & \quad x_{QE,i} \in [29920.0, 54088.6] \wedge \\
 & \quad x_{QB-B,i} \in [29397.3, 52244.6] \wedge \\
 & \quad x_{QR-G,i} \in [26218.0, 49527.0] \wedge \\
 & \quad x_{QP-G,i} \in [188.0, 1080.0] \wedge \\
 & \quad x_{MCRT-B,i} \in [1.8, 34.4] \wedge \\
 & \quad x_{SSV-D,i} \in [47.0, 134.0] \vee \\
 & \quad x_{DQO-D,i} \in (269.0, 538.0] \wedge \\
 & \quad x_{DBO-E,i} \in [90.0, 220.0] \quad \xrightarrow{0.83} Classer390 \quad \}
 \end{aligned}$$

17.2.1 Interpretación final:

- Classer392: “*QE* es bajo y *QB-B* es bajo y *QR-G* es bajo y *QP-G* es bajo y *MCRT-B* es alto”
- Classer389: “(*QE* es alto y *QB-B* es alto y *QR-G* es alto y *QP-G* es alto y *MCRT-B* es bajo) y (*SSV-D* es bajo o *DQO-D* no alto)”
- Classer383: “(*QE* es alto y *QB-B* es alto y *QR-G* es alto y *QP-G* es alto y *MCRT-B* es bajo) y (*SSV-D* no es bajo o *DQO-D* alto) y *DBO-E* no es bajo”
- Classer390: “(*QE* es alto y *QB-B* es alto y *QR-G* es alto y *QP-G* es alto y *MCRT-B* es bajo) y (*SSV-D* no es bajo o *DQO-D* alto) y *DBO-E* es bajo”

17.2.2 Evaluación de la propuesta

Una vez determinados los conceptos asociados a cada clase se ha procedido a evaluar cual(es) era(n) satisfecho(s) por cada objeto de \mathcal{I} con tal de evaluar la correcta correspondencia entre la muestra y la conceptualización inducida. La tabla 17.11 muestra los resultados de calcular el soporte $Sup(r)$, ver ecuación (9.1), la cobertura relativa $CovR(r)$, ver ecuación (9.5) y la confianza $p(r)$, ver ecuación (9.2), de cada una de las reglas compuestas inducidas para cada clase de la partición final.

La evaluación de la regla se hace con respecto al total de objetos de la base de datos, es evidente que habrá inconsistencias al evaluar cada regla por separado ya que los intervalos que conforman cada reglas compuesta, no son disjuntos entre una clase y otro debido a la forma en que se construyen los conceptos, ver ecuaciones (11.3), (11.4), (11.5), (11.6) y (11.1).

En cuanto a la **Confianza**(columna $p(r)$), si se observa la Tabla 17.11, la confianza se obtiene dividiendo las celdas de la columna $\#\{i \in A_C^\xi \cap i \in C\}$ por las de la columna $\#\{i \in A_C^\xi\}$. Como hay 3 clases en donde $\#\{i \in A_C^\xi\} > \#\{i \in A_C^\xi \cap i \in C\}$ se puede concluir que hay objetos mal asignados, es decir objetos que no satisfacen el consecuente de la regla, pero que si satisfacen el antecedente de esta. Este valor no se considera en el soporte ni en la cobertura relativa, pero queda de manifiesto en la confianza ya que cuando se obtienen confianzas del 100% significa que todos los objetos que satisfacen el antecedente de la regla, también satisfacen simultáneamente el antecedente y el consecuente de la regla y por lo tanto pertenecen a la clase que la regla asigna. Con lo cual el se puede concluir, en este caso, que el número de objetos correctamente asignados por clase viene dado por $\#\{i \in A_C^\xi \cap i \in C\}$ y el número de objetos mal asignados por clases, en este caso, se puede calcular $\#\{i \in A_C^\xi\} - \#\{i \in A_C^\xi \cap i \in C\}$. El porcentaje total de objetos correctamente asignados se puede obtener dividiendo 233 entre 395, 59%. Las confianzas en promedio rondan el 60%, lo cual se puede considerar como bueno.

	Concec.	$\#\{i \in A_C^\xi\}$	$\#\{i \in A_C^\xi \cap i \in C\}$	n_c	$p(r)$	$Sup(r)$	$CovR(r)$
$r_{Classer392}$	Classer392	6	6	6	100,0%	1,5%	100,0%
$r_{Classer389}$	Classer389	227	70	70	30,8%	57,5%	100,0%
$r_{Classer383}$	Classer383	108	27	34	25,0%	27,3%	79,4%
$r_{Classer390}$	Classer390	156	130	285	83,3%	39,5%	45,6%
<i>Media</i>					59,8%		81,3%
<i>Suma</i>		497	233	395		125,8%	
<i>CovGlobal(\mathbb{R})</i>							59%

Tabla 17.11: Evaluación: Best local-global concept and Close-World Assumption.

En cuanto al **Soporte**, se tiene que se obtiene dividiendo $\#\{i \in A_C^\xi\}$ entre el total de objetos de la base de datos (395). Si se considera la suma de los soportes de todas la reglas, es decir, el soporte de la base de conocimiento, cuando éste es menor que el 100% significa que hay objetos de la base de datos sin asignar, es decir, que no satisface ninguna regla y cuando es mayor hay inconsistencias, es decir que satisface varias reglas hacia distintas clases. También es interesante observar que el número de objetos asignados por las reglas compuestas asociadas a cada clase se puede obtener a partir de $\#\{i \in A_C^\xi\}$.

En cuanto a la **Cobertura relativa**, si se observa la Tabla 17.4, la cobertura relativa se obtiene dividiendo cada celda de la columna $\#\{i \in A_C^\xi \cap i \in C\}$ entre la correspondiente celda de la columna $\#\{C\}$.

En estas tablas, en lugar de representar la cobertura relativa media de cada base de

conocimiento, como se hace con la confianza, representamos la media de la cobertura relativa ponderada por el tamaño de cada clase, que coincide con el porcentaje global de objetos de la base de datos que se asignan correctamente, dando una idea mas ajustada de la calidad predictiva de la base de conocimiento inducida.

Capítulo 18

Análisis y resultados, planta catalana

18.1 Interpretación validada por el experto

Como ya hemos dicho anteriormente, los expertos han validado que la interpretación publicada en (Gibert and Roda 2000) es válida también para las 4 clases obtenidas con $[P4_{Gi1,R1}^{En,G}]$.

En esta sección recordamos la interpretación que a juzgar por los expertos que participan en éste estudio, es la que mejor caracteriza a las 4 clases provenientes de la partición objetivo y la que nos servirá para validar y comparar las interpretaciones obtenidas utilizando las propuestas Best global concept and Close-World Assumption y Best local-global concept and Close-World Assumption.

- **Classer392:** Agua que sale más limpia que lo normal (entra con suciedad grado medio) y licor mezcla con más biomasa o microorganismos de lo normal. Valores globalmente intermedios (inferiores en DQO-D, DBO-D, PH-S, SS-S, SSV-S, DQO-S, DBO-S, V30-B, NKT-S y NH4-S, y superiores en MLSS-B y MLSSV-B).

Según el experto corresponde a una clase que se caracteriza por unos elevados rendimientos de depuración (analíticas del agua en la salida y en el punto intermedio con valores bajos) debido a que hay un elevado nivel de biomasa en los reactores. Además, a partir de los últimos gráficos (estudio dinámico) hemos podido observar que esta clase corresponde a agua con nutrientes eliminados básicamente en el reactor biológico.

- **Claser389:** Agua que entra y sale con grado de suciedad medio, así como el licor mezcla aunque tiende a caudales de purga elevados y afluencia de aire baja. Corresponde a valores intermedios para casi todas las variables. Según el experto, cuando se purga más (QP-G elevada) se supone que hay menos biomasa en los reactores y, por lo tanto, es necesario airear menos (QA-G baja). De todas formas se tendría que reflejar en algún otro lugar, con unos MLSSV-B menores o un aumento de FE-E. En este caso, a partir de los últimos gráficos (estudio dinámico) no hemos podido caracterizar esta clase por ningún comportamiento particular de las variables a través del proceso de depuración.
- **Classer383:** Agua que entra más sucia de lo normal. Valores globalmente intermedios (superiores en SS-E, SSV-E, DQO-E y DBO-E). Según el experto existe un choque de carga (materia orgánica) de sólidos en la entrada del proceso. Podrían ser vertidos industriales. Además, a partir de los últimos gráficos (estudio dinámico) hemos podido observar que es una clase de aguas con gran número de partículas en suspensión donde la decantación surte máximo efecto.

- **Classer392:** Es una clase muy diferenciada del resto por tomar valores extremadamente pequeños en casi todas las variables (caudales con valores bajos y agua que entra, sale y está en el proceso poco sucia) menos el índice V30, la suciedad del licor mezcla y la edad celular que son elevados. Según el experto esto se corresponde a un periodo de tempestad. En Gerona, cuando llueve mucho, se cierran las compuertas que pasan por debajo del río para evitar que se rompan las conducciones. Ésto provoca que llegue muy poca agua ($Q-E$ baja) y muy diluida. Por contra, el cabezal de planta cierra completamente la purga ($QP-G$ es 0) y el nivel de la biomasa en los reactores (MLSS, MLVSS) y la edad celular (MCRT-B) aumentan mucho. Ésto podrá hacer frente al choque de carga que llegará después de la tormenta. A partir de los últimos gráficos (estudio dinámico) únicamente hemos podido observar que esta clase posee un comportamiento muy diferenciado del resto de clases, tomando valores muy diferenciados del resto.

18.2 Interpretaciones generadas por cada propuesta

A continuación se presenta el resumen de los resultados de las interpretaciones obtenidas tanto con la propuesta Best global concept and Close-World Assumption de construcción del concepto y Best local-global concept and Close-World Assumption:

18.2.1 Best global concept and Close-World Assumption:

- Classer392: “ $Q-E$ es bajo y $QB-B$ es bajo y $QR-G$ es bajo y $QP-G$ es bajo y $MCRT-B$ es alto”
- Classer389: “ $Q-E$ es alto y $QB-B$ es alto y $QR-G$ es alto y $QP-G$ es alto y $MCRT-B$ es bajo y $DQO-D$ no alto”
- Classer383: “ $Q-E$ es alto y $QB-B$ es alto y $QR-G$ es alto y $QP-G$ es alto y $MCRT-B$ es bajo y $DQO-D$ alto y $DBO-E$ es no bajo”
- Classer390: “ $Q-E$ es alto y $QB-B$ es alto y $QR-G$ es alto y $QP-G$ es alto y $MCRT-B$ es bajo y $DQO-D$ alto y $DBO-E$ es bajo”

18.2.2 Best local-global concept and Close-World Assumption:

- Classer392: “ $Q-E$ es bajo y $QB-B$ es bajo y $QR-G$ es bajo y $QP-G$ es bajo y $MCRT-B$ es alto”
- Classer389: “($Q-E$ es alto y $QB-B$ es alto y $QR-G$ es alto y $QP-G$ es alto y $MCRT-B$ es bajo) y ($SSV-D$ es bajo o $DQO-D$ no alto)”
- Classer383: “($Q-E$ es alto y $QB-B$ es alto y $QR-G$ es alto y $QP-G$ es alto y $MCRT-B$ es bajo) y ($SSV-D$ no es bajo o $DQO-D$ alto) y $DBO-E$ no es bajo”
- Classer390: “($Q-E$ es alto y $QB-B$ es alto y $QR-G$ es alto y $QP-G$ es alto y $MCRT-B$ es bajo) y ($SSV-D$ no es bajo o $DQO-D$ alto) y $DBO-E$ es bajo”

18.3 Análisis de los resultados

18.3.1 Análisis cualitativo

A modo de resumen se ha construido la tabla 18.1, con los conceptos generados para cada clase y por cada propuesta.

Método	clase	Q-E	QB-B	QR-G	QP-G	DBO-E	SSV-D	DQO-D	MCRT-B
BG & CWA	Classer392	bajo	bajo	bajo	bajo	-	-	-	alto
	Classer389	alto	alto	alto	alto	-	-	no alto	bajo
	Classer383	alto	alto	alto	alto	no bajo	-	alto	bajo
	Classer390	alto	alto	alto	alto	bajo	-	alto	bajo
BL + G &CWA	Classer392	bajo	bajo	bajo	bajo	-	-	-	alto
	Classer389	alto	alto	alto	alto	-	bajo	no alto	bajo
	Classer383	alto	alto	alto	alto	no bajo	no bajo	alto	bajo
	Classer390	alto	alto	alto	alto	bajo	no bajo	alto	bajo

Tabla 18.1: Resumen de las interpretaciones obtenidas por las diferentes propuestas.

Observando la tabla 18.1, se puede concluir que la propuesta Best global concept and Close-World Assumption, produce conceptos menos complejos a diferencia de la propuesta Best local-global concept and Close-World Assumption, incorporando una variable nueva a la interpretación (SSV – D) que en Best global concept and Close-World Assumption no estaba.

Si se observa la interpretación de las clases proporcionada por el experto la que más se acerca es Best local-global concept and Close-World Assumption, básicamente por la aportación de una nueva variable que enriquece la interpretación.

La clase “classer390” es la que contiene los días seleccionados por la reglas de clasificación P y Q, con lo cual se puede incluir en la interpretación generada por la propuesta Best global concept and Close-World Assumption las variables SS-S y DBO-S.

Según la interpretación generada por nuestra metodología las clases más difíciles de distinguir son las que se abren en la tercera iteración (Classer383 y Classer390), es aquí donde se hace evidente la posibilidad de introducir las variables de medidas de diferencia como las que se presentan en el class panel graph de la Tabla 18.4 y la Tabla 18.5.

Si observamos el class panel graph de la Tabla 18.4, se ve claramente que en el caso de la Classer383 llega agua muy sucia, y sale de calidad normal (ver también Figura 16.7 y Figura 16.8), es decir a pesar que el agua no sale totalmente limpia, si se está depurando bien, pues los valores de diferencia entre (SSV-E)-(SSV-D) y (SS-E)-(SS-D) son mas altos para esta clase de la que se distingue (classer390), y esta información si la han tenido en cuenta los expertos para generar la interpretación, es por ello que es mas rica a nivel semántico.

Esta hipótesis se reafirma al observar las siguientes coberturas relativas mas altas, que es el criterio que utiliza nuestra metodología para permitir el ingreso de variables a la interpretación, por ejemplo, en el caso de la partición en 3 clases, en donde no hay variables totalmente caracterizadoras de podrían incorporar nuevas variables que también aparecen en la interpretación generada por el experto (SS-D Y SSV-D), ver Tabla 18.2.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,classer391}^{DQO-D}$	146	45,625%
$r_{1,classer389}^{SS-D}$	24	34,28%
$r_{1,classer389}^{SSV-D}$	25	35,71%

Tabla 18.2: Cobertura relativa de DQO-D, SS-D y SSV-D en $\mathcal{S}(\mathcal{P}_3^*)$.

Otro ejemplo es el que se da en el caso de la partición en 4 clases, ver Tabla 18.3, en donde también puede entrar a la interpretación la variable DQO-E con el segundo valor mas alto de cobertura relativa.

Con lo anterior, se reafirma la buena calidad a nivel semántico de las interpretaciones generadas por nuestra metodología. De aquí el siguiente paso será valorar si se mejoran los indices de calidad y se enriquece la interpretación al incorporar mas variables, considerando las siguientes coberturas relativas mayores y no sólo “la cobertura relativa mayor”

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,classer390}^{DQO-E}$	113	39,64%
$r_{1,classer390}^{DBO-E}$	126	44,21%

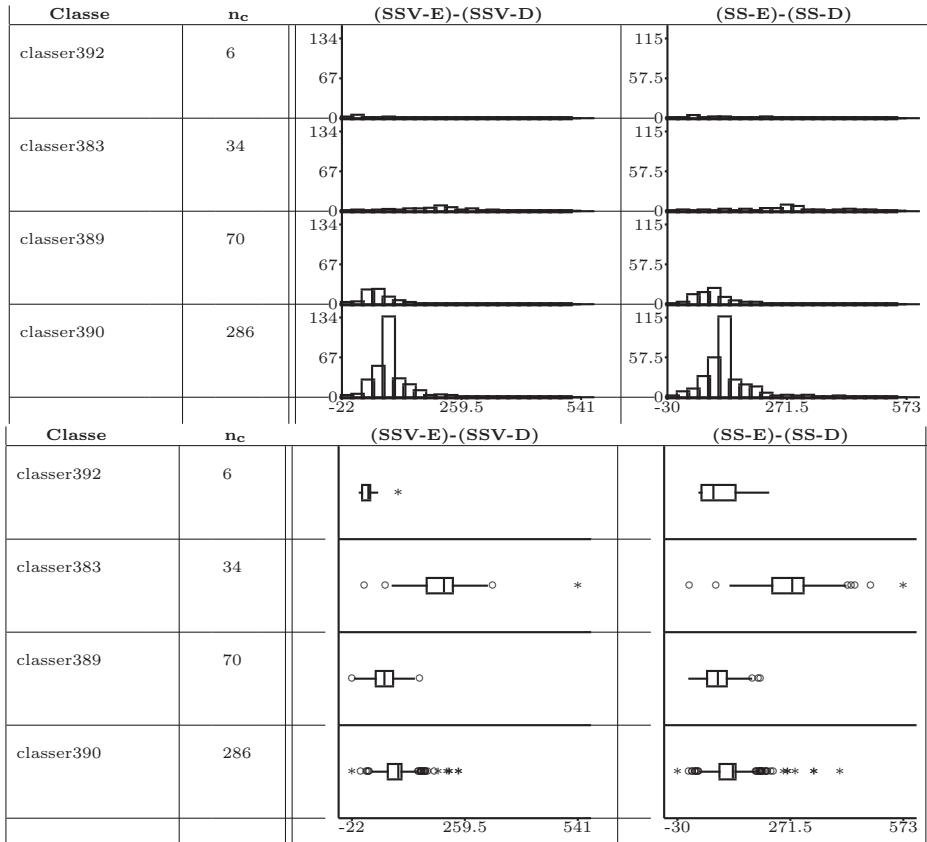
Tabla 18.3: Cobertura relativa de DQO-E y DBO-E en $\mathcal{S}(\mathcal{P}_4^*)$.

Tabla 18.4: Class panel Graph para las variables de diferencia.

	CLASSE	classer390	classer389	classer383	classer392
VARIABLE	N = 396	$n_c = 286$	$n_c = 70$	$n_c = 34$	$n_c = 6$
(SS-E)-(SS-D)	\bar{X}	114.6578	81.1076	278.2873	88.7306
	S	51.3135	39.4703	114.9987	71.5584
	min	-30	-1	2	26
	max	404	191	573	215
	N*	0	0	0	0
(SSV-E)-(SSV-D)	\bar{X}	90.1329	60.7783	202.5659	25.6033
	S	35.7877	29.7616	89.5304	34.8238
	min	-22	-22	9	-5
	max	244	147	541	93.6196
	N*	0	0	0	0

Tabla 18.5: Descriptiva estadística por clase.

18.3.2 Análisis cuantitativo

Como ya se ha dicho anteriormente, una vez determinados los conceptos asociados a cada clase, se ha procedido a evaluar cual(es) era(n) satisfecho(s) por cada objeto de \mathcal{I} a partir de evaluar la correcta correspondencia entre la interpretación generada por el experto y la conceptualización inducida. En la tabla 18.6 se muestra un resumen de los resultados de calcular el soporte $Sup(r)$, ver ecuación (9.1), la cobertura relativa $CovR(r)$, ver ecuación (9.5) y la confianza $p(r)$, ver ecuación (9.2), de cada una de las reglas compuestas inducidas para cada clase de la partición final y para las 2 propuestas para la integración del conocimiento aquí estudiadas.

En cuanto a las **Inconsistencias** (lo que significa que hay objetos que satisfacen el antecedente de más de una regla) en el caso de la propuesta Best global concept and Close-World Assumption es evidente que no habrá inconsistencias al evaluar cada regla por separado ya que los intervalos que conforman cada regla compuesta, son disjuntos entre una clase y otro debido a la forma en que se construyen los conceptos.

Met.	Ruler	Conc.	$\#\{i \in A_C^\xi\}$	$\#\{i \in A_C^\xi \cap i \in C\}$	n_c	$p(r)$	$Sup(r)$	$CovR(r)$
BG	$r_{Classer392}$	Classer392	6	6	6	100,0%	1,5%	100,0%
&	$r_{Classer389}$	Classer389	227	70	70	30,8%	57,5%	100,0%
CWA	$r_{Classer383}$	Classer383	55	10	34	18,2%	13,9%	29,4%
	$r_{Classer390}$	Classer390	61	61	285	100,0%	15,4%	21,4%
	$\bar{p}(\mathbb{R})$					62,3%		62,7%
	Suma		349	147	395		88,4%	
	$CovGlobal(\mathbb{R})$							37,2%
BL+	$r_{Classer392}$	Classer392	6	6	6	100,0%	1,5%	100,0%
G	$r_{Classer389}$	Classer389	227	70	70	30,8%	57,5%	100,0%
&	$r_{Classer383}$	Classer383	108	27	34	25,0%	27,3%	79,4%
CWA	$r_{Classer390}$	Classer390	156	130	285	83,3%	39,5%	45,6%
	$\bar{p}(\mathbb{R})$					59,8%		81,3%
	Suma		497	233	395		125,8%	
	$CovGlobal(\mathbb{R})$							59%

Tabla 18.6: Resumen resultados.

En el caso de la propuesta Best local-global concept and Close-World Assumption habrá

inconsistencias, es por esto que el total del número de objetos que satisfacen el antecedente es mayor al total de objetos de la base de datos. Esto ocurre debido a que los intervalos que conforman cada reglas compuesta no son disjuntos entre cada clase, es decir cada una de las reglas compuestas asociadas a cada clase presentan intersecciones, esto es debido a la forma en que se construyen los conceptos.

En cuanto al **Soporte**, si se observa la Tabla 18.6, se tiene que se obtiene dividiendo $\#\{i \in A_C^\xi\}$ entre el total de objetos de la base de datos (395).

En la tabla 18.6 se puede observar que el número de objetos asignados por las reglas compuestas asociadas a cada clase se puede obtener con $\#\{i \in I_s^{k,\xi}\}$ y el valor total viene cuantificado por el soporte, ya que la suma total de los soportes por clase corresponde al porcentaje de objetos asignados, cuando este es menor que el 100% significa que hay objetos sin asignar y cuando es mayor hay inconsistencias, es decir, hay objetos que han satisfacen el antecedente de mas de una regla compuesta.

Para la propuesta Best global concept and Close-World Assumption en la Tabla 18.6 se puede observar que se tiene un promedio de soporte menor que 100%, con lo cual hay objetos sin asignar. A diferencia del caso anterior en la propuesta Best local-global concept and Close-World Assumption la mayor parte de los objetos presentan inconsistencias, ya que tiene un soporte mayor que el 100%.

En cuanto a la **Confianza**, si se observa la Tabla 18.6, la confianza se obtiene dividiendo las celdas de la columna $\#\{i \in A_C^\xi \cap i \in C\}$ por las de la columna $\#\{i \in A_C^\xi\}$.

En la propuesta Best global concept and Close-World Assumption hay 2 clases en donde $\#\{i \in I_s^{k,\xi}\} > \#\{i \in I_s^{k,\xi} \cap i \in C\}$ y 3 en el caso de la propuesta Best local-global concept and Close-World Assumption por tanto se puede concluir que hay objetos mal asignados, es decir objetos que no satisfacen el consecuente de la regla, pero que satisfacen el antecedente de esta (esta característica es aplicable a todas las propuestas). Este valor no se considera en el soporte ni en la cobertura relativa, pero queda de manifiesto en la confianza ya que cuando se obtienen confianzas del 100% significa que todos los objetos que satisfacen el antecedente de la regla, también satisfacen simultáneamente el antecedente y el consecuente de la regla y por lo tanto pertenecen a la clase que la regla asigna. Con lo cual el se puede concluir, en este caso, que el número de objetos correctamente asignados por clase viene dado por $\#\{i \in I_s^{k,\xi} \cap i \in C\}$ y el número de objetos mal asignados por clases, en este caso, se puede calcular haciendo $\#\{i \in I_s^{k,\xi}\} - \#\{i \in I_s^{k,\xi} \cap i \in C\}$.

En cuanto a la **Cobertura relativa**, si se observa la Tabla 18.6, la cobertura relativa se obtiene dividiendo cada celda de la columna $\#\{i \in A_C^\xi \cap i \in C\}$ entre la correspondiente celda de la columna $\#\{C\}$.

La relación entre cobertura relativa y confianza es inversamente proporcional, a medida que se pierde confianza se gana cobertura relativa.

En estas tablas, en lugar de representar la cobertura relativa media de cada base de conocimiento, como se hace con la confianza, representamos la media de la cobertura relativa ponderada por el tamaño de cada clase, que coincide con el porcentaje global de objetos de la base de datos que se asignan correctamente, dando una idea mas ajustada de la calidad predictiva de la base de conocimiento inducida. Si observamos la Tabla 18.6 se puede ver que la propuesta Best local-global concept and Close-World Assumption tiene un porcentaje de cobertura global mejor que la propuesta Best global concept and Close-World Assumption, con lo cuál además de enriquecer la interpretación a nivel conceptual, la base de conocimiento inducida para la partición objetivo es de mejor calidad en cuanto al porcentaje de objetos bien asignados.

18.4 Conclusiones de la aplicación

- En cuanto a la interpretación de clases.
 1. Si se observa la interpretación de las clases proporcionada por el experto la que más se acerca es *Best local-global concept and Close-World Assumption*, básicamente por la incorporación de más conceptos dada la aparición de más variables que a diferencia de la propuesta original no existían.
 2. La propuesta *Best local-global concept and Close-World Assumption* está a medio camino hacia *Best global concept and Close-World Assumption*, aunque producen conceptos algo mas elaborados.
- En cuanto a las medidas de calidad de las bases de conocimiento inducidas según las distintas propuestas.
 1. Si se considera el índice de calidad $CovGlobal$ que representa el porcentaje de objetos de \mathcal{I} correctamente asignados por la base de conocimientos final. La propuesta *Best local-global concept and Close-World Assumption* es la que tiene la mejor cobertura global ($CovGlobal = 59\%$) lo que confirma la decisión que es la más acertada, tanto cualitativamente como en el valor de este índice de calidad.
 2. La suma total de los soportes por clase corresponde al porcentaje de objetos asignados, cuando este es menor que el 100% significa que hay objetos sin asignar y cuando es mayor hay inconsistencias.
 3. Cuando no hay inconsistencias ($Sup(r) \leq 100\%$) el número de objetos mal asignados por clase se puede calcular haciendo $\#\{i \in A_C^\xi\} - \#\{i \in A_C^\xi \cap i \in C\}$
 4. Si $\#\{i \in A_C^\xi\} > \#\{i \in A_C^\xi \cap i \in C\}$ se puede concluir que hay objetos mal asignados. Este valor queda de manifiesto en la confianza ya que cuando se obtienen confianzas del 100% significa que todos los objetos que satisfacen el antecedente de la regla, también satisfacen simultáneamente el antecedente y el consecuente de la regla y por lo tanto pertenecen a la clase a la que la regla asigna el objeto.
 5. La relación entre Soporte ($Sup(r)$) y Confianza ($p(r)$) es inversamente proporcional, a medida que se pierde confianza se gana soporte.
 6. La relación entre Cobertura relativa ($CovR(r)$) y Confianza ($p(r)$) es también inversamente proporcional, a medida que se pierde confianza se gana cobertura relativa.

18.5 Resumen

En general si se observa la interpretación de las clases proporcionada por el experto, la que más se acerca es *Best local-global concept and Close-World Assumption*, básicamente por la forma en que se comporta la temperatura y por la incorporación de más conceptos dada la aparición de más variables que en deferencia de la propuesta original no aparecían y por el valor de la cobertura global ($CovGlobal = 59\%$). Si el objetivo primordial es favorecer la riqueza conceptual de la interpretación o grado de interpretabilidad (y/o utilidad) de las clases formadas, la propuesta *Best local-global concept and Close-World Assumption* es la mejor, y también lo es, ya que la propuesta original sólo tiene una cobertura global ($CovGlobal = 37\%$), si se considera la capacidad predictiva (permitir para un nuevo objeto (día), predecir la clase (situación típica de la planta) que le corresponde y generar las caracterización e interpretación

conceptual correspondiente a esa clase), el hecho que existan inconsistencias genera conflictos, lo que ha se ha estudiado previamente en (Pérez-Bonilla and Gibert 2007a) (para más detalles ver el reporte de investigación (Pérez-Bonilla and Gibert 2008a)), donde se proponen diversos métodos de resolución de conflictos.

Capítulo 19

Caso de Estudio 2: Planta eslovena

19.1 Descripción general

La planta de tratamiento de aguas residuales Domzale-Kamnik, ver Figura 19.1 y 19.2 es la planta eslovena más grande actualmente en funcionamiento (200,000 PE), trata aguas residuales de origen municipal e industrial (50,000 personas + aguas residuales de provenientes de industrias y de origen meteorológico) de 4 municipios (Domzale, Kamnik, Menges and Trzin) (Hvala 2004), ver Figure 19.3



Figura 19.1: Vista aérea de la planta (1999).

La planta es de tipo convencional, diseñada para la eliminación de carbono orgánico. Es un sistema mecánico convencional que consta de 2 fases biológicas con digestión anaeróbica (proceso de descomposición de materia orgánica) y utilización de biogás en motores de biogás, ver Tabla 19.1. Las aguas, una vez depuradas, se vierten en el río Kamniška Bistrica.

La planta recibe aguas residuales con una alta concentración de componentes orgánicas e inorgánicas de diferentes industrias y municipios. Del total del caudal de entrada a la planta, el 30% son aguas residuales por carga hidráulica (37% origen doméstico, 28% industrial, 35% meteorológico) y el 70% restante por carga bioquímica (32% origen doméstico, 68% origen industrial) (Hvala 2004).

Las aguas residuales entrantes son difícil de biodegradar debido a la alta proporción de origen industrial, entre ellas farmacéuticas, galvanizadoras, mataderos, alimentación, muebles, textil, producción de tintes y otras, o sea que las condiciones en las que opera la planta

son muy complicadas debido a la diversidad de industrias que generan estas aguas residuales. Además la variación diario de carga orgánica y de nitrógeno varías hasta en un 100%



Figura 19.2: Otra vista área de la planta (2006).



Figura 19.3: Localización geográfica de la planta(WWTP).

Planta	Capacidad diseñada (PE)	Configuración del proceso
Domzale-Kamnik (Slovenia)	200,000 PE	Sistema convencional de 2 fases, mecánica y biológica para eliminación de carbono. Depuración primaria, digestión anaeróbica y uso de biogas

Tabla 19.1: Tamaño de la planta y configuración del proceso.

19.2 Descripción detallada del sistema

19.2.1 Sistema de alcantarillas

Uno de las alcantarillas principales lleva las aguas residuales sin tratar a la planta (EDAR) por flujo de gravedad. La alcantarilla transporta una mezcla de aguas residuales y de lluvia. La alcantarilla tiene varios desagües para cuando se producen tormentas lo que genera un exceso de agua que se vierte directamente al río Kamniska Bistrica. No existen depósitos ni en la planta ni en las alcantarillas para almacenar el exceso de aguas en caso de tormentas o lluvias extremas para su posterior tratamiento (Hvala 2004).

El único tratamiento de las aguas torrenciales en la planta de tratamiento de aguas residuales (EDAR) es el siguiente:

- Tratamiento mecánico de aguas torrenciales. La planta tiene una capacidad para aguas absorbidas de 48000 m³/d, cuando se sobrepasa esta capacidad, el exceso se vierte a una nueva corriente receptora en una tubería con una gruesa pantalla que filtra, detrás de la cual se encuentra un dique de desbordamiento de aguas pluviales.

Medidas y posibilidades de control en la alcantarilla:

- Actualmente hay un único lugar donde el Flujo y el nivel de Ph se miden continuamente. Sin embargo las medidas no tienen un seguimiento online en la EDAR.
- No están disponibles otros puntos de control (no es posible tomar muestras del agua residual en otros lugares de la EDAR) en los tubos de desagüe.

Debido a las medidas limitadas en los tubos de desagüe y a las falta de lugares donde estas muestras puedan ser recogidas. El control de los tubos de desagüe no se considera en el SMAC.¹.

La Figura 19.4 muestra un esquema simplificado del sistema de tubos de desagüe. Los puntos en rojo (uno en la alcantarilla y otro en la entrada de la EDAR) indican lugares donde el flujo y el Ph se miden continuamente. En estos lugares se ha instalado

¹SMAC - SMART CONTROL OF WASTEWATER SYSTEMS. Es un sistema Inteligente de control de los procesos de depuración de aguas residuales integrado por los sistemas de la red de alcantarillado y por las plantas de tratamiento de aguas residuales que están sujetos a grandes fluctuaciones del flujo y de las concentraciones. Hoy en día el control de los sistemas de aguas residuales se realiza en sub-unidades por medio de optimización local. Los resultados de un control afectan a otros, pero esto, hasta ahora, no se tiene en cuenta. El mantenimiento del sistema de alcantarillado y el control de la planta de tratamiento normalmente no está coordinado. El proyecto SMAC tiene por objeto ampliar el control funciones a inteligentes y globales, abarca todo el sistema de control con un seguimiento en línea del estado general de todo el sistema. SMAC permitirá integrar en el control la recogida y el tratamiento de las aguas residuales, incorporando variables como la precipitación pluvial y también, a largo plazo, hacer que la planificación de los procesos de eliminación de nutrientes sea más rentable y con una mejor sostenibilidad. Al conectarse con SMAC los usuarios tendrán acceso a las descripciones de la planta, los resúmenes de todos los informes internos, boletines informativos, etc. (<http://www.smac.dk/>)

shows a simplified scheme of the sewer system. Red points indicate places (one in the sewer and one at the wastewater treatment plant inlet) where flow and pH are continuously measured. En estos lugares, se encuentran instalados equipos estacionarios de muestreo. Los puntos negros indican posibles lugares de muestreo sin ningún tipo de equipo (Marjeta Strazar and Burica 2006).

19.2.2 Características de la EDAR

Influent

La capacidad de la planta 200.000 PE (population equivalents = habitantes equivalentes y 1 PE = 60 gBOD5/d) con un flujo diario of 24.000 m³/día. La Tabla 19.2 muestra el diseño de la EDAR y la información real del flujo de entrada.

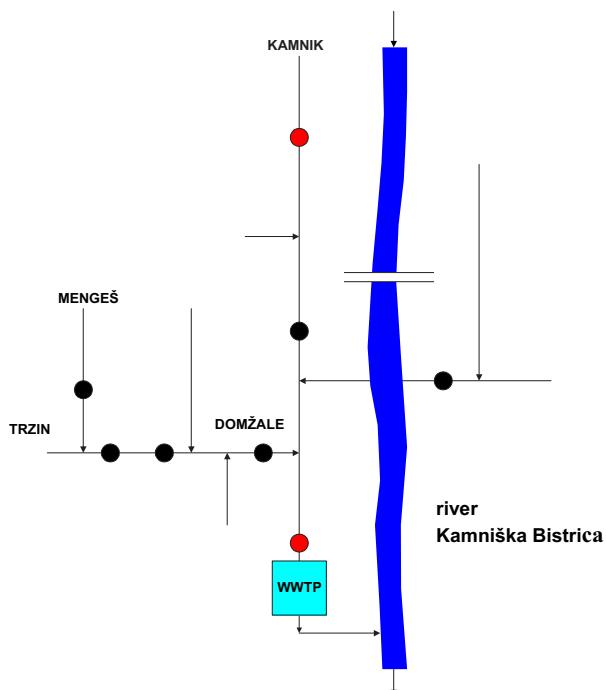


Figura 19.4: Esquema simplificado del sistema de alcantarillas.

Parameter	Datos teóricos	Datos reales
PE (BOD5)	200,000	130,000
Flujo medio con tiempo seco	24,000 m ³ /h	19,010 m ³ /h
Flujo máximo con tormentas	48,000 m ³ /h	48,000 m ³ /h
BOD5 carga	12,000 kg/día	15,019 kg/día (including technological recycles)
N carga total	-	1,741 kg/day (including technological recycles)
P carga total	-	306 kg/day (including technological recycles)

Tabla 19.2: Caudal de entrada: datos teóricos y datos reales.

Unidades de la EDAR

La planta de tratamiento de aguas residuales (EDAR) (Marjeta Strazan and Burica 2006) consta de una estación de bombeo (de las aguas residuales sin tratar que llegan a la planta a través de las alcantarillas), una estructura de detección de grava gaseoso y una cámara de grasa, tanques que son decantadores del flujo longitudinal primario y un proceso convencional de dos etapas de lodos activados, (Vrecko and Hvala 2006) (muy cargados en la primera etapa biológica y menos cargado en la segunda etapa biológica), sin eliminación de nutrientes específicos, ver Figura 19.5.

El lodo es bombeado desde un *"pre-espesante"* a cuatro digestores funcionando a temperaturas mesófilas. En los digestores el lodo es estabilizado biológicamente y se genera gas metano que se utiliza en motores de biogás; lo cual cubre aproximadamente el 25% de la jornada de consumo de electricidad diaria de planta. El lodo digerido se centrifuga, deshidrata y deposita en un vertedero interno. Más tarde, los lodos mezclados con serrín y tierra se utiliza como relleno de suelo y para fines no agrícolas.

Los volúmenes de cada una de las unidades de la planta se muestran en la Tabla 19.3

UNIDAD		VOLUMEN TOTAL
Tratamiento mecánico	2 cámaras de grava aereado	500 m ³
	decantadores primarios (2 tanques)	2000 m ³
Primera etapa biológica altamente cargada	2 tanques aeróbicos	2000 m ³
	2 tanques de sedimentación (cada uno dividido en 2)	2400 m ³
Segunda etapa biológica menos cargada	4 tanques aeróbicos	4000 m ³
	4 tanques de sedimentación (cada uno dividido en 2)	6800 m ³
Tratamiento de lodos	4 digestores anaeróbicos contenedor de gas	7200 m ³ 800 m ³

Tabla 19.3: Unidades de la planta y volúmenes.

Calidad de las aguas tratadas (salida de la planta)

La calidad exigida para las aguas tratadas se muestra en la Tabla 19.4. Los valores establecidos durante 24 horas se basan en muestras que son proporcionales al tiempo o al flujo. Como actualmente la EDAR Domzale-Kamnik no se localiza en una región clasificada como sensible al tratamiento terciario debe sólo alcanzar los requisitos para el tratamiento secundario.

La planta siempre ha cumplido con los requisitos del efluente en cuanto a la cantidad de carbono que lo compone. La reducción del COD es alrededor del 90%, DBO₅ 95% y los sólidos en suspensión de un 98%. La eliminación de otros nutrientes se realiza con menor éxito. Como la planta, actualmente, no está diseñada para eliminación de nitrógeno, no es capaz de reducir el nitrógeno total por más del 50% (ver Tabla 19.5).

Costos de operación

Los costos operacionales de la EDAR Domzale-Kamnik se pueden ver en la Tabla 19.6. En Eslovenia, actualmente, no se pagan multas por la calidad de las aguas residuales de las industrias.

Supervisory Control and Data Acquisition -SCADA (Sistema de supervisión de control y adquisición de datos).

La operación de la planta se supervisa mediante un Sistema de control y adquisición de datos (SCADA²), ver Figura 19.6). Un computadora central se utiliza para controlar el funcionamiento de la planta en todos los lugares importantes. El laboratorio de la planta trabaja con el apoyo de LIMS (Sistema de Gestión de la Información en laboratorios). El actual sistema consta de varios Power Line Communications (Comunicaciones mediante línea de energía) PLC-S conectado a la del sistema SCADA. Sobre la base de una conexión en línea, el equipo de supervisión es conectado a la computadora personal con MS Access y SQL Server (Figura 19.6). Los datos del equipo (ordenador) de supervisión se actualizan cada 20 segundos. Cada 15 minutos, el proceso de datos es almacenado en la base de datos de SQL y está disponible para el técnico.

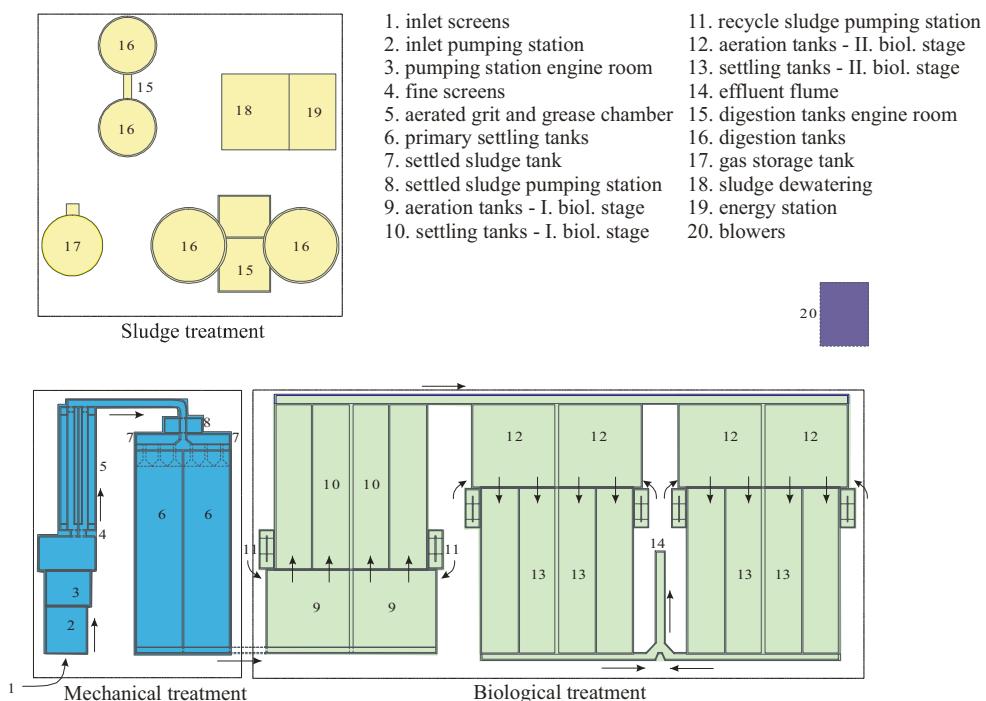


Figura 19.5: Diseño con las principales unidades de la planta.

Plant Capacity > 100,000 Pe	Secondary Treatment	Tertiary Treatment
COD	100 mg/L	100 mg/L
BOD5	20 mg/L	20 mg/L
Suspended solids	35 mg/L	35 mg/L
NH4-N*	10 mg/L	-
Ntotal	-	10 mg/L
Ptotal	-	2 mg/L

Tabla 19.4: Normas del efluente.

²El término se refiere a un sistema de medición (y control) distribuidos a gran escala. SCADA se utiliza para supervisar o para controlar químicos, físicos o procesos de transporte, en el sistema de abastecimiento municipal de agua, para controlar la generación y la distribución de energía eléctrica y para gaseoductos y oleoductos, y otros procesos distribuidos.

Parameter	unit	Influent			Effluent			Discharge limit
		avg	max	min	avg	max	min	
COD	mg O ₂ /l	589	962	370	72	105	50	100
BOD	mg O ₂ /l	340	480	209	17	39	7	20
SS	mg/l	239	520	104	16	37	3	35
NH4-N	mg/l	19,6	29,6	12,3	9,3	17,5	1,1	10
N-total	mg/l	42,6	62,4	27,8	24,7	34,5	13,7	-
P-total	mg/l	6,5	11,0	3,7	3,1	5,0	1,4	-

Tabla 19.5: Afluentes y efluentes de aguas residuales característicos de la planta.

Costos operacionales	EDAR Domzale-Kamnik
BOD ₅ , removed	0,77 /kg
COD, removed	0,47 /kg
KWh	0,062 /kWh
wastewater TN, removed	0,5 /m ³
TP, removed	12,66 /kg
Waste sludge	70,23 /kg
Polymers	43,943 /ton
	31,8 /ton

Tabla 19.6: Costos de operación de la EDAR:Domzale-Kamnik wastewater.

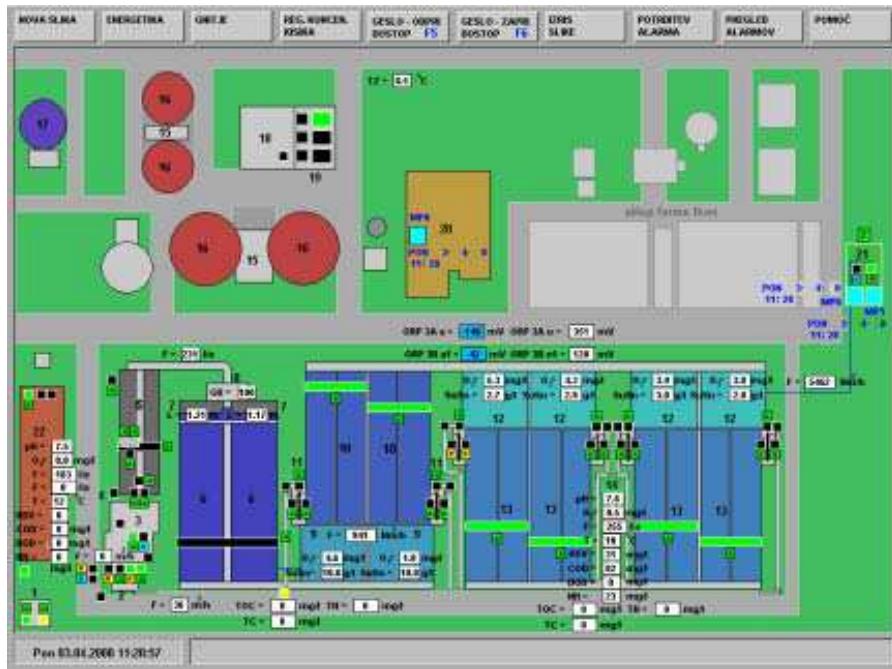


Figura 19.6: Diseño en la planta del sistema SCADA.

19.2.3 Estructuras de control actualmente disponibles

Control de oxígeno

El oxígeno necesario para la aireación es suministrada por difusores de burbuja fina a una profundidad de 4m. Los turbo-compresores necesarios para la producción de aire comprimido tienen una capacidad variable progresivamente, de 2 x 8.000 m³/h.

Los compresores mantienen la presión constante automáticamente en un colector común, a partir del cual el aire se distribuye a los seis tanques (ver Figura 19.7). El controlador de presión en el tubería principal es un controlador PI.

Un total de 6 sondas de oxígeno continuamente miden la concentración de oxígeno (DO) disuelto en los tanques. Como las mediciones de DO no siempre son fiables, el control de DO se implementa como un control de presión de DO tipo cascada. En total, 6 de los controladores de PI en cascada controlan las concentraciones de DO en los tanques.

En cada controlador de presión, el servidor esclavo es un controlador de presión PI, que ajusta la apertura de la válvula de aire a fin de que la deseada diferencia de presión (P) se mantenga en la tubería correspondiente a cada tanque.

Para cada controlador de presión, la diferencia de presión entre los puntos RV1 a RV6 se determina por el controlador master PI y el oxígeno depende de la concentración deseada y real de DO en el correspondiente tanque.

Si la medición de DO falta, el sistema mantiene la presión, que había en el tanque antes de los datos que faltan. El punto de referencia de DO en la primera etapa biológica es 2,0-2,2 mg/L y en el segunda etapa biológica 2,5-2,7 mg/L.

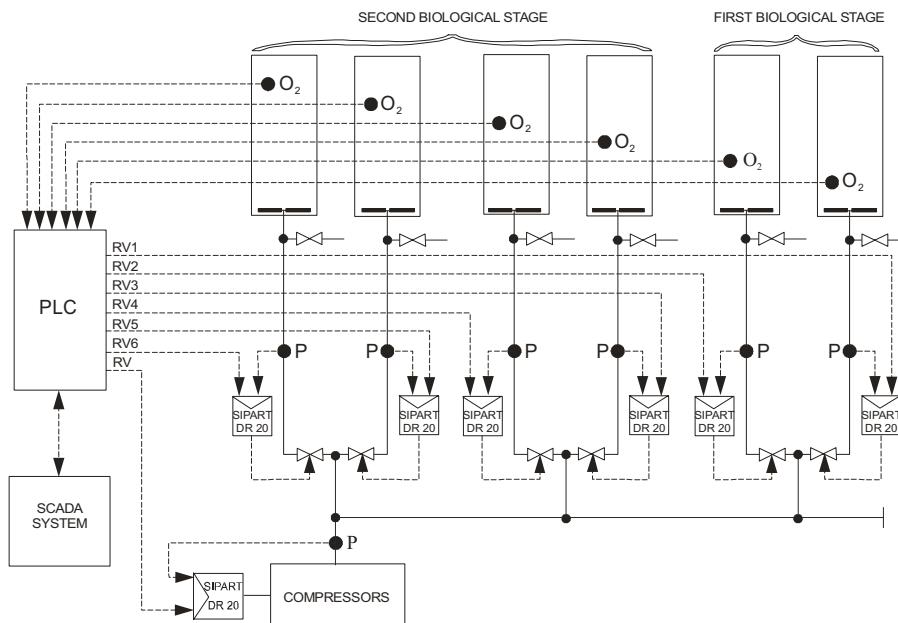


Figura 19.7: Control de Oxígeno.

Control del retorno de lodos

El control automático no se aplica para el control retorno de lodos. El retorno de lodo se bombea desde los decantadores a los tanques aeróbicos en un flujo constante de 100 L/s, lo que equivale entre 100-300% del flujo de entrada.

Control de lodos residuales

No hay control automático para el control de lodos residuales. El exceso de lodos se devuelve desde los encantadores secundarios a la estación de bombeo de la entrada. El exceso de lodos se elimina según sea necesario. La cantidad de lodos residuales se determinará según la medición de los parámetros del proceso (SVI, MLS, MLVSS e investigación microscópica) y sobre la base del conocimiento experto del operador.

19.3 Planta piloto considerada en SMAC

La EDAR Domzale-Kamnik eslovena, está planificando el mejoramiento del proceso de nitrificación-desnitrificación. Para estos fines se han instalado 2 plantas pilotos en 2 depósitos que ya existían (2 x 1.125m³) para comparar el funcionamiento del proceso convencional de los lodos activos (Vrecko and Hvala 2006) y el proceso alternativo denominado MBBR (Moving Bed Biofilm Reactor).

Las plantas pilotos han estado en operación por mas de 2 años. El funcionamiento de ambos procesos ha sido analizado por modelos matemáticos y de simulación. La tecnología elegida para el mejoramiento de la planta, de acuerdo a los estudios realizados, es MBBR (Moving Bed Biofilm Reactor). En SMAC, la planta piloto con MBBR (Moving Bed Biofilm Reactor) será usada como sistema experimental para pruebas simultaneas del proceso tecnológico y del control del mejoramiento de la planta, ver Figure 19.8.

19.3.1 Variables del proceso manipuladas

El plan de mejoramiento prevé un tipo de cascada con desnitrificación posterior.

El caudal de entrada de aguas residuales y se divide y se dirige hacia diferentes unidades. Como puede verse en la Figura19.10, esta configuración no prevé ningún tipo de control de los que normalmente se utilizan en la planta de tratamiento de aguas residuales. Hay flujo de lodo residual y otro flujo de lodo residual que proviene de la recirculación interna. Además, para esta configuración no es necesario añadir carbono, ya que una parte de las aguas residuales de la entrada va a los tanques de desnitrificación.

En la configuración de la planta piloto es posible manipular el flujo aire que va a los tanques aeróbico. Es por eso que el control de aireación (Control de oxígeno disuelto en cada tanque aeróbico, en nuestro caso estas variables son O₂-1aerobic and O₂-2aerobic) se estudiará y evaluará en el proyecto SMAC.

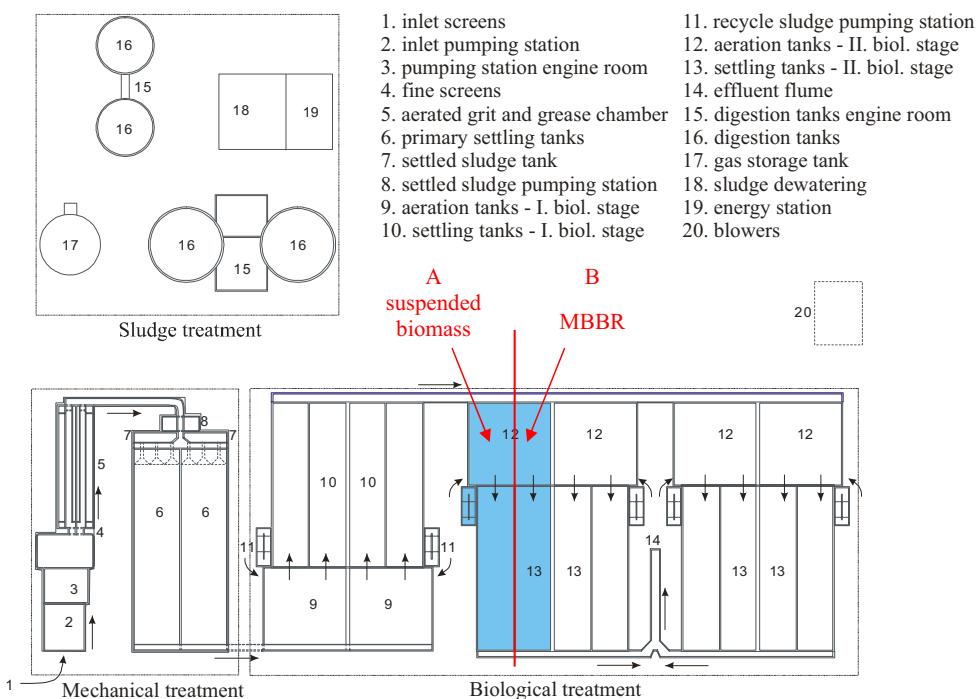


Figura 19.8: El diseño actual de la plantas piloto dentro de la EDAR eslovena.

19.3.2 Configuración de la planta piloto

El sistema de control de aireación se pondrá a prueba en el MBBR (Moving Bed Biofilm reactor) de la planta piloto.

Un plan detallado de la planta piloto se muestra en la Figura 19.10.

La planta piloto consta de dos tanques anóxicos y dos tanques aeróbicos, que están tienen plataformas de plástico sobre las cuales la biomasa se desarrolla.

Durante el período aeróbico, las bacterias autotróficas³ oxidan amoníaco a nitratos utilizando la reacción de nitrificación. Durante el período anóxico, las de bacterias heterótrofas consumen el carbono orgánico y convierten los nitratos procedentes del caudal de entrada y aquellos generados las bacterias autotróficas en gas nitrógeno. Los ciclos entre las condiciones aeróbicas y anóxicas permiten a los sistemas de reactores eliminar el amoníaco, el carbono orgánico y los nitratos de manera eficiente.

1. Tanque anóxico: Durante el período anóxico, se produce el consumo de bacterias heterótrofas (carbono orgánico) y se convierten los nitratos procedentes de los afluente y los generados por las bacterias autotróficas en nitrógeno gas (desnitrificación).

La digestión anaerobia es un proceso bacteriano que se lleva a cabo en la falta de oxígeno. El proceso puede ser termófilo la digestión, en el que el lodo se fermenta en tanques a una temperatura de 55 °C, o mesófilas, a una temperatura de alrededor de 36 °C. Aunque permitiendo que el tiempo de retención más cortos (y por lo tanto tanques más pequeños), termófilo la digestión es más caro en términos de energía consumo para la calefacción de los lodos. Una de las principales características de anaerobios la digestión es la producción de biogás, que puede ser utilizado en generadores para la producción de electricidad y / o en calderas para fines de calefacción.

2. Tanque aeróbico: Durante el período aeróbico, las bacterias autotróficas oxidan amoníaco a nitratos utilizando la reacción de nitrificación.

La digestión aeróbica es un proceso bacteriano que ocurren en presencia de oxígeno. Bajo condiciones aeróbias, las bacterias consumen rápidamente materia orgánica y la convierten en dióxido de carbono. El funcionamiento se caracterizan por tener costos mucho mayores que para el proceso anaeróbico porque para la digestión, los costos de energía necesaria para añadir oxígeno a la proceso son mas elevados.

La eliminación de nitrógeno se efectúa a través de la oxidación biológica de nitrógeno de amoníaco (nitrificación) a nitrato, seguido por la desnitrificación, reducción de nitratos a nitrógeno gas. El gas nitrógeno se libera a la atmósfera y, por tanto, queda eliminado del agua.

La nitrificación es la oxidación biológica de amonio con oxígeno en nitrito, seguido por la oxidación de esos nitritos en nitratos. La oxidación del amonio en nitrito, y la subsecuente oxidación a nitrato es hecha por dos especies de bacterias nitrificantes. La primera etapa la hacen bacterias (entre otras) del género microbiológico Nitrosomonas y Nitrosococcus. La

³Bacterias encargadas de la descomposición de los desechos orgánicos. Se pueden agrupar en dos grupos, las heterotróficas y las autotróficas. Las bacterias autotróficas son algo más exigentes para reproducirse que las heterotróficas y su crecimiento es mucho más lento, necesitando incluso semanas para un desarrollo en número suficiente. Su alimentación se basa en reacciones químicas y no requiere de la presencia de desechos orgánicos. Su necesidad básica es el carbono y pueden obtenerlo si fuera necesario del dióxido de carbono del agua. Además de este proceso emplean oxígeno para metabolizar compuestos nitrogenados. Entre las principales bacterias autotróficas podemos citar las Nitrosomas y las Nitrobacter, pero existen otras especies como las Nitrocystis que desarrollan similares procesos metabólicos. Las especies de éstas varían en función del tipo y las condiciones del agua.

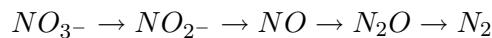
segunda etapa (oxidación de nitrito a nitrato) la hacen, mayormente, bacterias del género Nitrobacter, y en ambas etapas produciendo energía que destinan a la síntesis de ATP. Esos microorganismos nitrificantes son quimioautótrofos, y usan dióxido de carbono como su fuente de carbono para crecer.

La nitrificación también juega un importante rol en la remoción del nitrógeno orgánico de aguas servidas, donde la remoción convencional es por esa nitrificación bacteriana, seguida de desnitrificación. El costo de este proceso reside mayormente en la aereación (dar oxígeno en el reactor) y la adición de una fuente externa de carbono (e.g. metanol) para la desnitrificación.

En muchos ambientes, ambos organismos se hallan juntos, rindiendo nitrato como el producto final. Sin embargo, es posible diseñar sistemas donde se forme selectivamente nitrito (el proceso Sharon).

En conjunto con la amonificación, la nitrificación forma parte del proceso de mineralización, que hace referencia a la descomposición completa de materia orgánica, con la liberación de compuestos nitrogenados disponibles para los vegetales (formas minerales, no orgánicas).

La desnitrificación es un proceso que realizan ciertas bacterias durante la respiración usando el nitrato como acceptor de electrones en condiciones anóxicas (ausencia de oxígeno). El proceso de reducción de nitratos hasta nitrógeno gas ocurre en etapas seriales, catalizadas por sistemas enzimáticos diferentes, apareciendo como productos intermedios nitritos, óxido nítrico y óxido nitroso:



La desnitrificación requiere un sustrato oxidable ya sea orgánico o inorgánico que actúe como fuente de energía, por lo que la desnitrificación puede llevarse a cabo tanto por bacterias heterótrofas como autótrofas. En la desnitrificación heterótrofa, un sustrato orgánico, como metanol, etanol, ácido acético, glucosa, etc. actúa como fuente de energía (donador de electrones) y fuente de carbono. En la desnitrificación autótrofa, la fuente de energía es inorgánica, como hidrógeno o compuestos reducidos de azufre: sulfídrico (H_2S) o tiosulfato ($S_2O_3^{2-}$), la fuente de carbono, también inorgánica, es el CO_2 .

El mayor problema de la desnitrificación biológica es la contaminación potencial del agua tratada con: bacterias, fuente de carbono residual (desnitrificación heterótrofa) y la posibilidad de formación de nitritos, lo cual hace necesario un post-tratamiento. A día de hoy, los procesos desarrollados para la desnitrificación biológica son diversos usando distintos sustratos y diferentes configuraciones de reactores. Pero hay que destacar que prácticamente la totalidad de los sistemas de desnitrificación desarrollados se basan en la desnitrificación heterótrofa habiendo un gran vacío en el conocimiento y desarrollo de la desnitrificación autótrofa.

A veces, la conversión de amoníaco tóxico de nitrato por sí solo se conoce como tratamiento terciario.

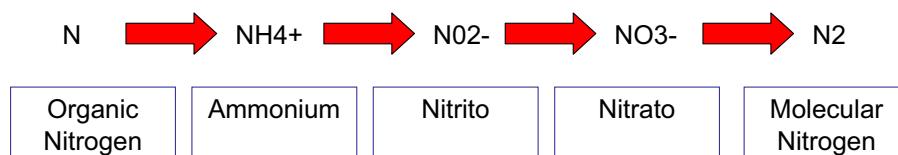


Figura 19.9: Proceso.

Las aguas residuales entran a la planta piloto después de tratamiento mecánico de la EDAR, se bombea a la planta piloto.

El caudal entrada se mantiene constante para fijar el tiempo de retención hidráulico.

El caudal entrada puede ser ajustado periódicamente para observar el desempeño de la planta piloto en diferentes tiempos de retención hidráulicos.

El sistema de aireación tiene las siguientes limitaciones:

- Sólo el flujo de aire total que va hacia a los 2 tanques aeróbicos puede ser manipulado *on-line*. Esto se logra mediante la manipulación de la válvula *V2* en la tubería principal a la MBBR (Moving Bed Biofilms Reactor) de planta piloto. La distribución de aire a cada uno de los tanques está determinada por las válvulas en las tuberías individual. La apertura de estas dos válvulas se ajusta manualmente y rara vez se cambia (puede ser una vez al mes o cada varios meses). Las aberturas habituales de las válvulas son de 80% y 40% para el primer y el segundo tanque de aeróbico, respectivamente.
- La apertura mínima de *V2* es de un 30%.
- La apertura mínima de las válvulas manuales en cada una de las tuberías se determina por el mínimo caudal de aire que se necesita para lograr la mezcla en los tanques de aeróbicos.
- La máxima apertura de la válvula manual al segundo tanque aeróbico es determinada por la máxima concentración de DO en el tanque de manera que no perturba el proceso de desnitrificación en los tanques afónicos.

El sistema que controla la aireación actualmente aplicado sobre el MBBR (Moving Bed Biofilm reactor) de la planta piloto no es muy preciso y se diferencia de la situación que se explica en el inciso anterior. Por lo tanto, para el efecto de la SMAC, el mismo controlador cascada para presión de DO se implementará también en la MBBR (Moving Bed Biofilms Reactor) de la planta piloto .

La planta piloto está equipado con los siguientes sensores en línea: concentración total de nitrógeno(TN), concentración total de carbono orgánico (TOC), un sensor que mide la inhibición y un sensor para la concentración de NH₄-N. Para los efectos de la SMAC, los sensores de los nutrientes adicionales(NH₄-N, N-NO₃) se pondrán en la planta piloto. Los sensores adicionales permitirán un control más avanzado de aireación (DO control, DO punto de referencia de control) basándose en los datos en línea de los nutrientes (feed-forward, basada en el modelo predictivo de control en línea de datos de NH₄-N y NO₃-N).

19.3.3 Experimentación con el “DO set-point control”

El experimento se realiza en 2 fases:

1. En primer lugar, los experimentos pueden realizarse sin algoritmos de control. Estos experimentos se llevarán a cabo para determinar la relación entre el punto de referencia de DO y la tasa de nitrificación para el proceso en el MBBR (Moving Bed Biofilms Reactor). Debido a la evolución de la temperatura, el tiempo de retención hidráulico se ajustará periódicamente para lograr la concentración TN en el caudal de entrada por debajo de 10 mg/L y, así mismo, la de concentración NH₄-N por debajo de 5 mg/L. Los datos de estos experimentos también se utilizarán para el diseño de modelos (a efectos de control) y la algoritmos de control, para mas información véase (Vrecko and Hvala 2006).
2. En la segunda fase, los experimentos se llevarán a cabo con algoritmos de control SMAC, que se implementarán en la planta piloto. Para evaluar la eficiencia y los costos económicos de la propuesta de control, se comparará el rendimiento de la planta experimentando con los diferentes algoritmos de control cuando usa un punto de referencia fijo de DO.

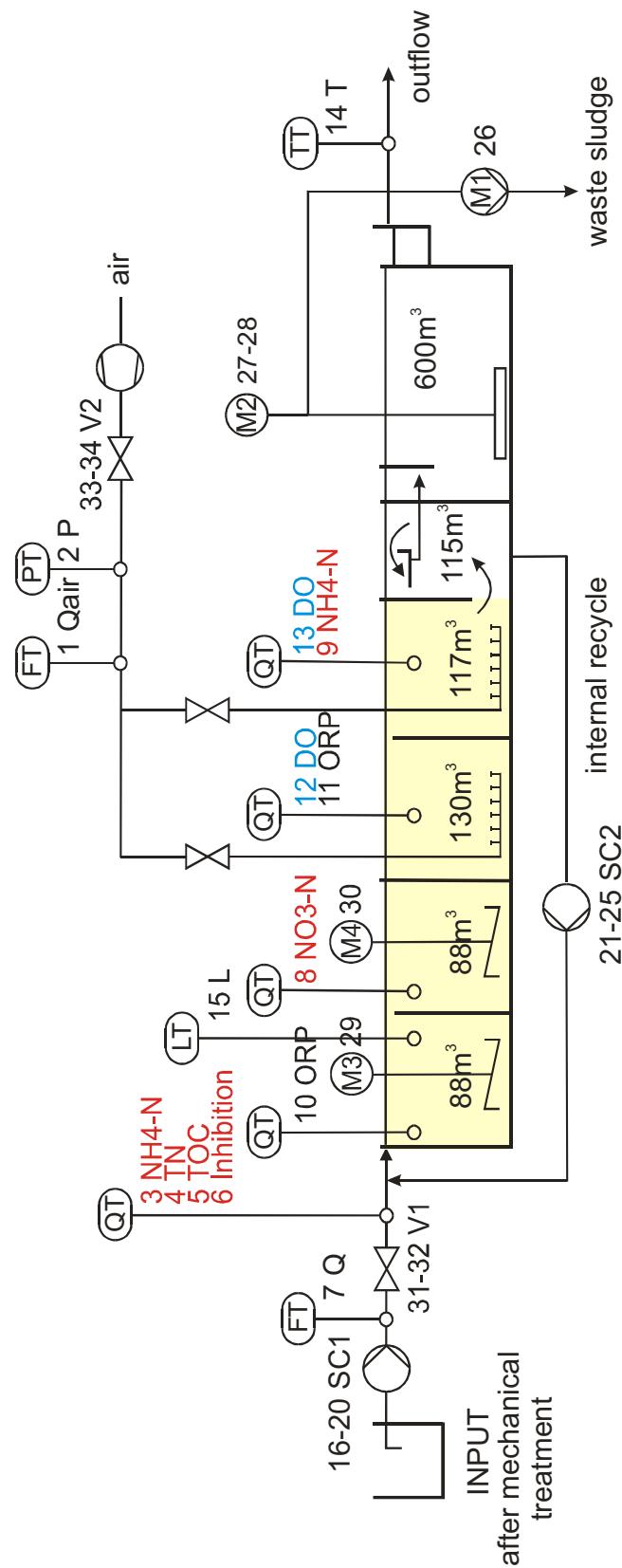


Figura 19.10: MBBR (*Moving Bed Biofilm Reactor*) planta piloto con sensores, actuadores y variables.

19.3.4 Inhibición del caudal de entrada a la planta piloto

El rendimiento de la planta está muy influenciado por la presencia de compuestos tóxicos que a menudo deterioran el rendimiento de esta (calidad del efluente). En el SMAC, la inhibición será supervisada y analizada. Potencialmente, se implementará, en el MBBR (Moving Bed Biofilms Reactor), el control del caudal de entrada a la planta piloto para prevenir a largo plazo la mala calidad del agua que sale. En caso de un aumento excesivo de la inhibición, la entrada de la bomba (SC1) se apaga.

19.3.5 Datos actualmente disponibles en linea para la planta piloto

Los siguientes datos dinámicos en línea están disponibles para el seguimiento de la operación de la MBBR (Moving Bed Biofilms Reactor) en planta piloto : Influent TN (ver Fig. 19.11), Influent TOC (see Fig. 19.12), Influent Inhibition (Toxicity (ver Fig. 19.15 y 19.16)), Effluent Temperature, Effluent NH4-N (ver Fig. 19.13 y 19.14)

19.3.6 Objetivos del SMAC

La planta de tratamiento de aguas residuales Domzale-Kamnik nitrificación-desnitrificación. Por lo tanto, el objetivo principal es encontrar los parámetros de funcionamiento, que garanticen la calidad deseada del efluente. Los siguientes objetivos referentes a la implementación del sistema de control SMAC en el sitio de prueba P13 deberán ser alcanzados en el proyecto SMAC:

- Implementación de sensores adicionales de nutrientes (NH4-N, NO3-N) en la planta piloto.
- Implementación de control de DO(control rápido por capa) sobre la planta piloto.
- Desarrollo de un medio ambiente experimental para pruebas de control del SMAC.
- La experimentación con punto de referencia de DO en la piloto planta.
- Implementación de un punto de control de referencia de DO en la planta piloto.
- Análisis de la inhibición y afluentes. Posiblemente el diseño de un control de inhibición.

19.3.7 Legislación

La planta piloto donde hemos recogido los datos de momento son utilizados para el ensayo de la nueva tecnología MBBR (Moving Bed Biofilms Reactor) para la eliminación de nitrógeno, véase (Kocjan 2004). Si los resultados son buenos entonces la tecnología MBBR (Moving Bed Biofilms Reactor) será utilizada para el mejoramiento de la planta (Vrecko and Hvala 2006).

En cuanto a las limitaciones de los valores de efluentes:

- Por el momento sólo se definen límites de concentración para amoníaco y carbono orgánico en el efluente.
- La concentración de amoníaco en el efluente debe ser inferior a 10 mg/l y la concentración total carbono orgánico debe ser inferior a 100 mg/l. (véase (Stare, Hvala, and Vrecko 2006)).

En Eslovenia la concentración de amoníaco (en nuestro caso NH₄-2aerobic) debe ser inferior a 10 mg/l. En algunas zonas más sensibles la concentración de nitrógeno total debe ser inferior a 10 mg/l. Sin embargo, esta es una legislación más estricta y no se considera en nuestro caso. En algunos zonas mas sensibles en Eslovenia las normas son más estrictas, para las las concentraciones de amoníaco y nitrógeno. En estas áreas la concentración total de nitrógeno en el efluente debe ser inferior a 10 mg/l y la eficiencia de eliminación de nitrógeno total, véase (Kocjan 2004), debe ser superior al 80%. Sin embargo, nuestra planta no se encuentra en una zona. Por lo tanto, me permito sugerir a utilizar las dos primeras limitaciones, que no son tan estrictas. Si desea disponer de más limitaciones, también puede solicitar una extra, por ejemplo, que el efluente de concentración de nitrógeno total debe ser inferior a 18 mg/l.



Figura 19.11: Máquina de medición en la línea de las variables TOC y TN.



Figura 19.12: Máquina de medición en la línea de las variables TOC y TN (zoom).



Figura 19.13: Máquina de medición en la línea de la variable NH4-N (NH4-2aerobic)(punto superior).



Figura 19.14: Máquina de medición en la línea de la variable NH4-N (NH4-2aerobic)(punto inferior).

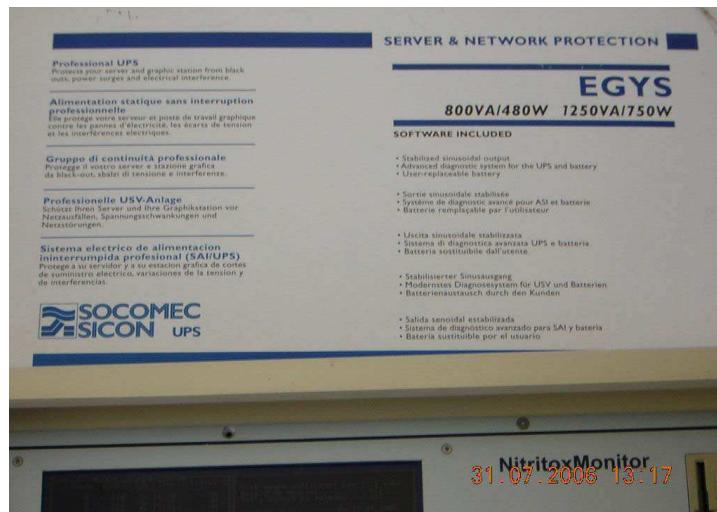


Figura 19.15: Máquina de medición en la línea de la variable *influent inhibition* (Nitritox).



Figura 19.16: Máquina de medición en la línea de la variable *influent inhibition* (Nitritox).

19.3.8 Presentación de los datos

Los datos obedecen a la colaboración existente con el Department of Systems and Control Jozef Stefan Institute de Ljubljana Esovenia y en particular con el doctor Darko Vrecko, quien realiza investigación con esta Planta depuradora de aguas residuales.

La base de datos original consta de 8.760 observaciones una por hora (24 observaciones por día) a partir de junio, 1º de 2005 y mayo, 31 de 2006 medidos en la planta piloto. Hay dos tanques aeróbicos (3º y 4º tanque) oxigenados y dos tanques anóxicos (1º y 2º tanque) que no están oxigenados. El último tanque (5º tanque) es un zona muerta (depósito) sin que transporte de plástico y por tanto, sin biomasa (más información en la subsección §19.3). Cada observación incluye la medición de 16 variables consideradas pertinentes por los expertos.

Para esta tesis, se ha calculado la media diaria y, por lo tanto tenemos una base de datos con 365 observaciones de las 16 variables.

19.3.9 Descripción de las variables

Las 16 variables numéricas que se han medido son las siguientes:

1. Influent
 - (a) NH4-influent (mg/l): Ammonia concentration at the influent of the pilot plant (mg of ammonia per liter of wastewater). It corresponds to the variable 3 in the figure 19.10.
 - (b) Q-influent (m³/h): Wastewater influent flow rate. It corresponds to the variable 7 in the figure 19.10, for more detail see subsection §19.3.2.
 - (c) FR1-DOTOK-20s (Hz): It is a frequency of the influent flow rate meter. This frequency is multiplied with a factor to obtain the influent flow rate. It is therefore completely correlated with the influent flow rate (Q-influent) and can therefore be omitted from the study.
 - (d) TN-influent (mg/l): Concentration of the total nitrogen at the influent of the pilot plant. It corresponds to the variable 4 in the figure 19.10, for more detail see subsection §19.3.2.

- (e) TOC-influent (mg/l): Total organic carbon concentration at the influent of the pilot plant. It corresponds to the variable 5 in the figure 19.10.
- (f) Nitritox-influent: Measurement of the inhibition at the influent of the pilot plant. It corresponds to the variable 6 in the figure 19.10.

2. First anoxic tank

3. Second anoxic tank

- (a) h-wastewater (m): Height of the wastewater in the tank. I think it is measured in the second anoxic tank. This variable is not shown in the figure 19.10.

4. First aerobic tank

- (a) O₂-1aerobic (mg/l): Dissolved oxygen concentration in the 1st aerobic tank (3rd tank). It corresponds to the variable 12 in the figure 19.10.
- (b) Valve-air: It is an openness of the air valve in percentage (between 0 - 100 %). It corresponds to the valve V2 in the figure 19.10. The minimal opening of V2 is 30%, for more detail see subsection §19.3.2.
- (c) Q-air (m³/h): Total air flow that is dosed in both aerobic tanks. It corresponds to the variable 1 in the figure 19.10. Only the total air flow to the both aerobic tanks can be on-line manipulated. This is achieved by manipulating valve V2 in the main pipe to the MBBR (Moving Bed Biofilm Reactor) pilot plant, for more detail see subsection §19.3.2.

5. Second aerobic tank

- (a) NH₄-2aerobic (mg/l): Ammonia concentration in the second aerobic tank. It corresponds to the variable 9 in the figure 19.10.
- (b) O₂-2aerobic (mg/l): Dissolved oxygen concentration in the 2nd aerobic tank (4th tank). It corresponds to the variable 13 in the figure 19.10.

6. Effluent

- (a) TN-effluent (mg/l): Concentration of the total nitrogen at the effluent (outflow) of the pilot plant. It is not shown in the figure 19.10, for more detail see subsection §19.3.2.
- (b) Temp-wastewater (C): Temperature of the wastewater. It corresponds to the variable 14 in the figure.
- (c) TOC-effluent (mg/l): Total organic carbon concentration at the effluent of the pilot plant. It is not shown in the figure 19.10.

7. Other

- (a) Freq-rec (Hz) : It is a frequency of the internal recycle flow rate meter. It gives you the information about the internal recycle flow rate.

Capítulo 20

Análisis descriptivo de los datos de la planta eslovena

20.1 Análisis univariante y bivariante

20.1.1 Introducción

Una vez definidos los parámetros a medir procedemos a una descripción exhaustiva de cada una de las variables disponibles.

El análisis descriptivo nos permitirá, en primera aproximación, hacernos una idea de la composición de la muestra. Este análisis se divide en dos grandes bloques: un primer análisis descriptivo de cada una de las variables (análisis univariante) y un segundo análisis de ciertos parámetros, en algunos casos, con gráficos bivariantes de puntos (Plots), que nos permiten intuir la relación entre dos variables y el sentido de ésta, es decir, si es positivo o negativo (análisis bivariante). Este último análisis es interesante de hacer debido que hay variables que son medidas en cada uno de los dos puntos clave que podemos encontrar en el proceso de depuración (en la entrada y en la salida de la Planta piloto), ver Figura 19.10.

20.1.2 Análisis univariante

El primer análisis presenta los estadísticos de descripción clásicos (número de valores, número de valores no missing, media, mediana, media truncada, desviación estándar, mínimo, máximo, 1º cuartil y 3º cuartil), un Boxplot y un Histograma, que muestra la distribución de la variable. Finalmente presentamos un resumen de los hallazgos más importantes derivados de este análisis. En este Capítulo sólo se presenta una variable a modo ilustrativo, las 15 variables restantes están en el Anexo F en la sección §F.1.

Variable Temp-ww

La variabilidad es bastante alta como se aprecia en el coeficiente de variación y el el time series plot de la Figura 20.1.

La temperatura controlada en la salida de la planta piloto y medida en grados Celsius, tiene un valor de 16.2514($^{\circ}$ C) en media, con una desviación de 3.7807 ($^{\circ}$ C) y unos valores que van desde 8.217($^{\circ}$ C) a 22.583($^{\circ}$ C).

No existe ningún dato missing para esta variable.

La distribución que se muestra en la Figura 20.2 presenta 2 claros picos correspondientes a los 2 distintos niveles de temperatura que se pueden encontrar, uno de ellos en torno a los 12 $^{\circ}$ C - 13 $^{\circ}$ C y el otro en torno a los 19 $^{\circ}$ C - 20 $^{\circ}$ C.

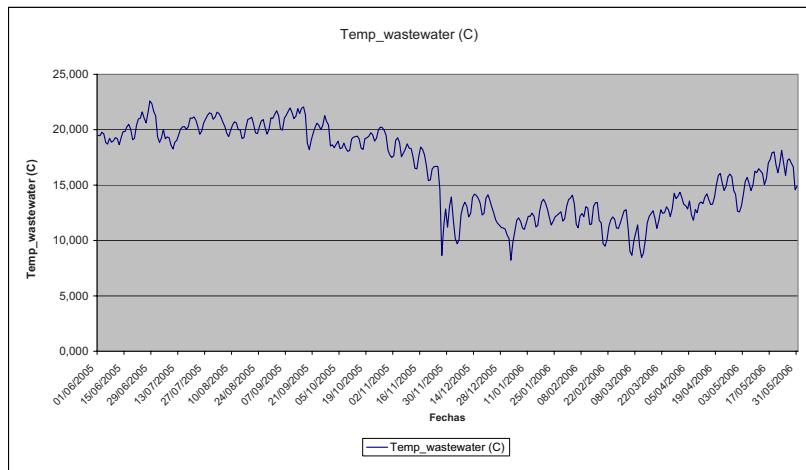


Figura 20.1: Serie temporal para la variable Temp-wastewater.

Se observa un comportamiento de los datos que se podría ajustar a 2 distribuciones, y esto podemos considerarlo en términos de que la temperatura en una variable que se manifiesta de forma estacionaria, es decir una distribución para valores "primavera-verano" y otra para valores "otoño-invierno"

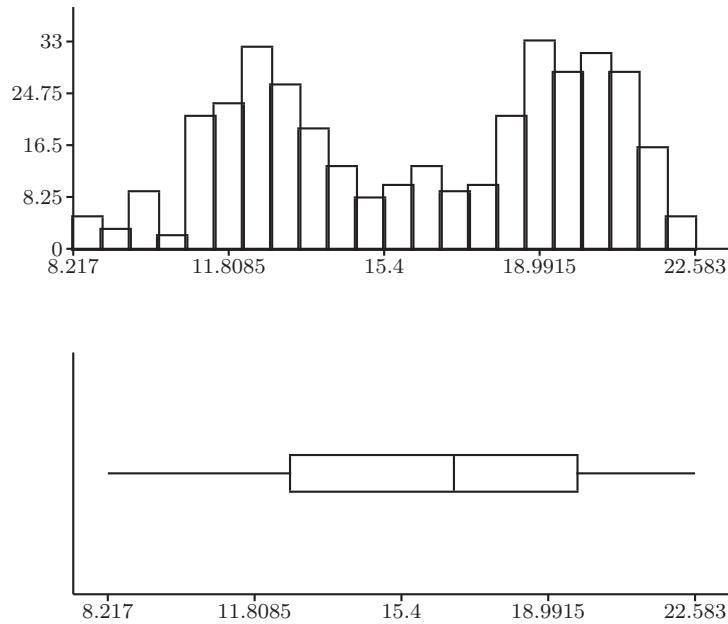


Figura 20.2: Histograma y Boxplot de la variable Temp-ww.

Este hecho nos podría hacer pensar que esta variable cambia según la estación del año en la que nos encontramos pero, aunque no encontramos los cambios de media próximos a los cambios de estación (sino en la parte central de estos), si que observamos valores más elevados para verano y más bajos para invierno.

Tabla de frecuencias	
Modalidades	Freq. absol.
8.217 - 8.87	5
8.87 - 9.523	3
9.523 - 10.176	9
10.176 - 10.829	2
10.829 - 11.482	21
11.482 - 12.135	23
12.135 - 12.788	32
12.788 - 13.441	26
13.441 - 14.094	19
14.094 - 14.747	13
14.747 - 15.4	8
15.4 - 16.053	10
16.053 - 16.706	13
16.706 - 17.359	9
17.359 - 18.012	10
18.012 - 18.665	21
18.665 - 19.318	33
19.318 - 19.971	28
19.971 - 20.624	31
20.624 - 21.277	28
21.277 - 21.93	16
21.93 - 22.583	5
Missings	0

20.1.3 Análisis bivariante

El análisis bivariante nos muestra la evolución de un conjunto de variables que se miden en los puntos claves del proceso que tiene lugar en la planta piloto. Ello contribuye a poner de manifiesto el efecto global del proceso de depuración de aguas residuales en la parte correspondiente a la planta piloto de la Domzale-Kamnik Wastewater Treatment Plant (WWTP). Como en el caso del análisis univariante sólo presentamos un caso a modo de ejemplo, en el Anexo F en la sección §F.2 se pueden ver otras relaciones entre pares de variables que se han estudiado.

TOC-influent and TOC-effluent.

Correlació per les Variables TOC-influent i TOC-effluent		
$r = 0.5359$ (coVar = 203.592)		
Informació sobre dades mancants		
TOC-influent \ TOC-effluent	útil	mancant
útil	365	0
mancant	0	0

Debido al poco parecido en las observaciones de estas dos últimas variables hemos realizado este gráfico que corrobora nuestra suposición; estas dos variables están relacionadas (correlación de 0.54), aunque no en gran medida, y existe una relación directa entre ellas (a mayor valor de una variable, mayor el valor de la otra) a excepción de muy pocos valores;

lo cual era de esperar por ser mediciones sobre datos en la misma fase de depuración. Ver Figura 20.3

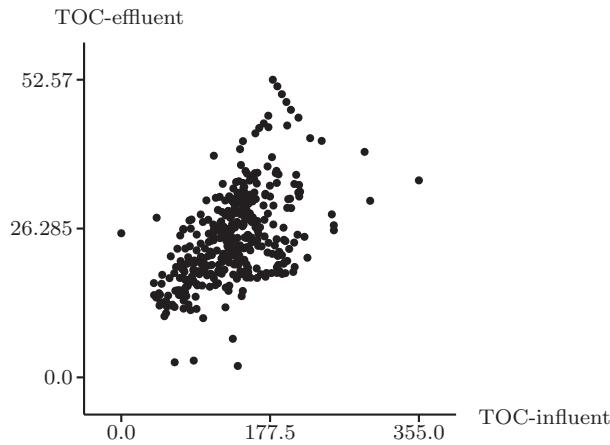


Figura 20.3: Diagrama bivariante para las variables TOC-influent and TOC-effluent.

20.2 Time series plot

Los siguientes gráficos muestran la distribución de todas las variables consideradas en el MBBR (Moving Bed Biofilms Reactor), véase (Hvala 2004), las cuales dan una indicación de la carga de entrada a la planta piloto.

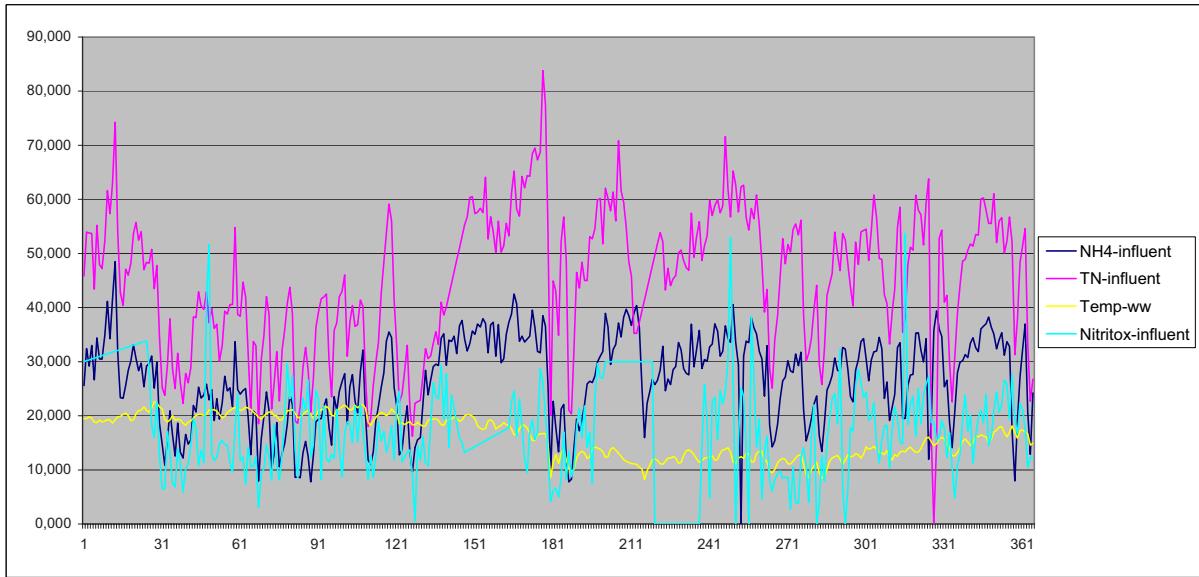


Figura 20.4: Serie temporal para 4 variables.

Los datos pertenecen al período comprendido entre el 1º de junio de 2005 y 31 de mayo de 2006. En el gráfico puede verse que, en el período comprendido entre septiembre-diciembre de 2005, en la entrada, el NH₄-2aerobic (Figura F.3) aumenta la concentración (Figura 20.4) y presenta una disminución al final del año debido a bajada de temperaturas. En todo el período, el tiempo de retención hidráulico en los tanques aeróbicos y anóxicos es de 6 horas en promedio.

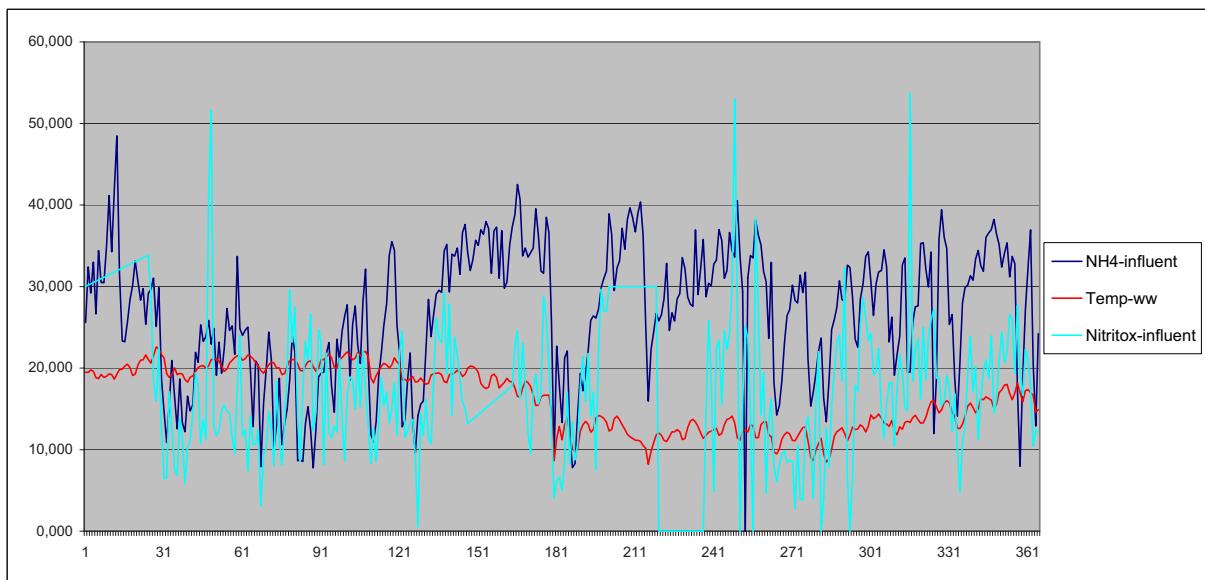


Figura 20.5: Serie temporal para 3 variables.

Los datos comprendidos entre enero y junio de 2006 muestran un lento aumento de la temperatura (Figura 20.4). En este período el tiempo promedio de retención hidráulica se estableció en 8 horas, favoreciendo la nitrificación a temperaturas más bajas (especialmente en el comienzo del año). De la Figura 20.1, F.3, F.23, F.21 se puede ver que hay bastante períodos de tiempo largo en que el proceso de ejecución del MBBR es muy pobre. Esto puede ser parcialmente explicado por la inhibición. A partir de la inhibición se pueden identificar 5 eventos, cuando la inhibición es muy alta (cerca de 100%), cuando la inhibición va en aumento durante un tiempo más largo (6-13 horas), ver información detallada de las cifras Figura F.27, Figura 20.1 y Figura F.3, a largo plazo (varios días o semanas) provoca un aumento de la concentración de NH₄-2. Las cifras muestran que la recuperación de la planta tras el aumento de inhibición puede ser bastante largo, especialmente a bajas temperaturas.

La Figura 20.4 muestra que la pérdida de capacidad debida a la nitrificación y al aumento de la inhibición puede ser un gran problema en la planta de tratamiento de aguas residuales.

La actividad de la biomasa autotróficas se analizó por la tasa de respiración, la actividad de la biomasa se midió en uno de los tanques de aeróbicos con suspensión de biomasa. Un repentina disminución de la tasa de respiración al comienzo de la gráfica parece más probable, debido a una mayor inhibición, ello puede ser entendido también por que la recuperación de la biomasa después de este evento es muy lenta, probablemente también porque la reducción de la concentración de TN en los afluentes es mayor, ver Figura 20.4(segunda parte del diagrama), este periodo corresponde a vacaciones de verano cuando las descargas industriales son más bajas.

El NH₄-2aerobic en la entrada (Figura 20.5) también muestra el aumento de la concentración de NH₄-2aerobic. Estos picos se notan especialmente bien a temperaturas más bajas (Figura 20.5) en caso de que el proceso pierda nitrificación. Los picos pueden ser en gran medida explicados por el retorno de los lodos de tratamiento de la estación de bombeo en la entrada.

Capítulo 21

Clustering planta eslovena

21.1 Introducción

Se tiene una familia de ficheros originales, estos ficheros son los que contienen la información para realizar la clasificación utilizando Java-KLASS y es la base de datos original de todas las variables medidas en la planta.

La estructura contiene un número total de 16 variables de las cuales 16 variables son numéricas y 0 variables son categóricas y un número total de 365 observaciones.

Según los expertos estas 16 variables son las más informativas para evaluar el funcionamiento de la planta piloto y son aquellas con las que se trabajará en el proceso de clasificación (ver la descripción de las variables en sección §19.3.9). Éstas corresponden a: NH4-influent (mg/l), Q-influent (m³/h), FR1-DOTOK-20s (Hz), TN-influent (mg/l), TOC-influent (mg/l), Nitritox-influent, h-wastewater (m), O₂-1aerobic (mg/l), Valve-air (%), Q-air (m³/h), NH4-2aerobic (mg/l), O₂-2aerobic (mg/l), TN-effluent (mg/l), Temp-wastewater (C), TOC-effluent (mg/l), Freq-rec (Hz).

En éste capítulo se presenta los detalles de la mejor clasificación de acuerdo a la recomendación y el conocimiento proporcionado por los expertos. Para más información sobre otras clasificaciones realizadas con esta base de datos ver en (Pérez-Bonilla, Gibert, and Vrecko 2007b), (Pérez-Bonilla, Gibert, and Vrecko 2007a) y (Pérez-Bonilla, Gibert, and Vrecko 2008).

Los aspectos técnicos del clustering jerárquico se han presentado previamente en el capítulo §12, en la sección §12.2.

21.2 Base de conocimiento para la clasificación basada en reglas

La base de conocimiento proporcionada por los expertos y que recoge las limitaciones legales expuestas en la sección §19.3.7, se traduce en las siguientes reglas donde:

- $r_1 : ((\text{AND } (>= (\text{NH4-2aerobic}) 10.0) (> (\text{TN-effluent}) 18.0)) \rightarrow \text{Mmonia})$
- $r_2 : ((\text{AND } (< (\text{NH4-2aerobic}) 10.0) (> (\text{TN-effluent}) 18.0)) \rightarrow \text{Nitrogen})$

En la Figura 21.1 se muestran los *subárboles* inducidos por cada regla.

La planta piloto donde hemos recogido los datos de momento son utilizados para el ensayo de la nueva tecnología MBBR (Moving Bed Biofilms Reactor) para la eliminación de nitrógeno, véase (Kocjan 2004). Si los resultados son buenos entonces la tecnología MBBR (Moving Bed Biofilms Reactor) será utilizada para el mejoramiento de la planta ((Vrecko and Hvala 2006)).

En cuanto a las limitaciones de los valores de efluentes:

- Por el momento sólo se definen límites de concentración para amoníaco y carbono orgánico en el efluente.
- La concentración de amoníaco en el efluente debe ser inferior a 10 mg/L y la concentración total carbono orgánico debe ser inferior a 100 mg/l. (véase (Stare, Hvala, and Vrecko 2006)).

En Eslovenia la concentración de amoníaco (en nuestro caso NH4-2aerobic) debe ser inferior a 10 mg/l. En algunas zonas más sensibles la concentración de nitrógeno total debe ser inferior a 10 mg/l. Sin embargo, esta es una legislación más estricta y no se considera en nuestro caso.

En algunos zonas mas sensibles en Eslovenia las normas son más estrictas, para las las concentraciones de amoníaco y nitrógeno. En estas áreas la concentración total de nitrógeno en el efluente debe ser inferior a 10 mg/l y la eficiencia de eliminación de nitrógeno total, véase (Kocijan 2004), debe ser superior al 80%. Sin embargo, nuestra planta no se encuentra en una zona. Por lo tanto, me permito sugerir a utilizar las dos primeras limitaciones, que no son tan estrictas. Si desea disponer de más limitaciones, también puede solicitar una extra, por ejemplo, que el efluente de concentración de nitrógeno total debe ser inferior a 18 mg/l.

21.3 Clustering

Así se construye la siguiente familia de ficheros de clasificación que una vez clasificados dan origen a la partición objetivo que se desea interpretar.

21.3.1 Familia de ficheros de clasificación

Se construye una familia de clasificación necesaria para operar con el software Java-KLASS, véase Capítulo §7 y anexo A, que incluye 365 observaciones (una observación por día), 16 variables y la base de conocimiento proporcionada por experto para la clasificación basada en reglas, ver §21.2, se puede ver la descripción de las reglas y mas detalles de ésta y otras clasificaciones realizadas con esta base de datos en el reporte (Pérez-Bonilla, Gibert, and Vrecko 2007a).

1. Nombre de los ficheros de clsificación:

- depdarko1.dat
- depdarko1.obj
- depdarko1.pro
- depdarko1.reg

2. Estructura de los ficheros de clasificación:

- Número de variables utilizadas para clasificar = 16.
- Número de variables numéricas en la clasificación = 16.
- Número de variables categóricas en la clasificación = 0.
- Número de objetos en la clasificación = 365.
- Descripción de las reglas utilizadas, proporcionadas por el experto (se encuentran el fichero depdarko1.reg) y se presentan en la sección §21.2.

La clasificación aquí presentada se obtiene utilizando la familia de clasificación antes descrita.

21.3.2 Clasificación basada en reglas

1. Parámetros de entrada para la Clasificación

- Métrica utilizada = Euclidea normalizada.
- Criterio de Clasificación = Ward.
- Ponderación de objetos = no.
- Tipo de Ponderación = Global.

2. Resultados:

- Nombre del Fichero de resultados .his: residual2.his
- Porcentaje de missing en los datos = 0.003%
- Árbol de clasificación (o dendrograma) $[\tau_{Lj3,R2}^{EnW,G}]$, ver Figura 21.3
- Gráfica de inercia interna entre clases, ver Figura 21.2
- Se recomienda cortar en (4 10 3 12 6 11 5) clases.
- Partición inducida por las reglas:
 - Seleccionados como clase Mmonia = 38 objetos, ver Figura 21.1 izquierda
 - Seleccionados como clase Nitrogen = 80 objetos, ver Figura 21.1 derecha
 - Seleccionados como clase Residual = 247 objetos

En la Figura 21.1 se muestran los *subárboles* inducidos por cada regla.

3. Cortes realizados:

De acuerdo al conocimiento proporcionado por los expertos y al criterio heurístico que tiene en cuenta la relación entre la variabilidad intra y entre clases que implementa Java-KLASS la mejor partición es la que se obtiene al cortar el árbol en 4 clases, ver Figura 21.3, y es con la que se trabajará para generar la interpretación final utilizando la metodología CCCS.

A partir de lo anterior se procede a cortar el árbol en 4 clases, ver Figura 21.3, y de esta manera obtener la partición de referencia (partición objetivo):

$$\mathcal{P}4_{Lj3,R2}^{EnW,G} = \{Cr358, Cr360, Cr353, Cr357\}$$

Como ya se ha explicado en capítulos anteriores, la propuesta metodológica aprovecha la estructura jerárquica del clustering y por lo tanto también se presenta, en este capítulo, los cortes en 2 y 3 clases necesarios para aplicar la metodología CCCS.

Los ficheros que incluyen los 3 cortes realizados tienen los siguientes nombres:

- (a) En 2 clases: ddreg22.cls (y .par) - $P2_{Lj3,R2}^{EnW,G}$. ver Figura 21.4
- (b) En 3 clases: ddreg23.cls (y .par) - $P3_{Lj3,R2}^{EnW,G}$. ver Figura 21.5
- (c) En 4 clases: ddreg24.cls (y .par) - $P4_{Lj3,R2}^{EnW,G}$. ver Figura 21.6 y Figura 21.7.

21.3.3 Secuencia de particiones

Los cortes sucesivos en 2, 3 y 4 clases a partir del dendrograma, ver Figura 21.3, son:

$$\begin{aligned} & Cr361 \{ Cr353 \\ & Cr363 \{ Cr357 \\ & Cr362 \{ Cr358 \\ & Cr360 \end{aligned}$$

Cada partición está compuesta de las siguientes clases:

$$\mathcal{P}_2^{EnW,G}_{Lj3,R2} = \{Cr361, Cr362\}$$

El análisis descriptivo por clases para esta partición se puede ver en la Figura 21.4.

$$\mathcal{P}_3^{EnW,G}_{Lj3,R2} = \{Cr362, Cr353, Cr357\}$$

El análisis descriptivo por clases para esta partición se puede ver en la Figura 21.5.

$$\mathcal{P}_4^{EnW,G}_{Lj3,R2} = \{Cr358, Cr360, Cr353, Cr357\}$$

El análisis descriptivo por clases para esta partición se puede ver en la Figura 21.5 y Figura 21.7

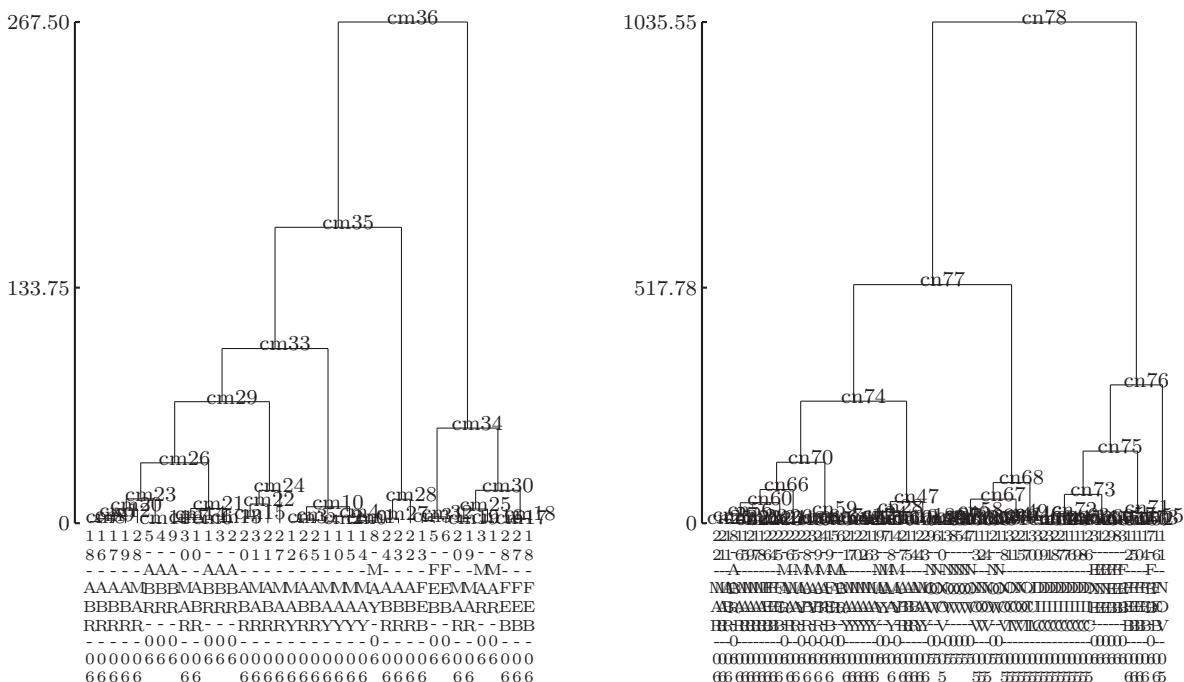


Figura 21.1: Árboles inducidos por las reglas de clasificación (Mmonia izq. y Nitrogen der.).

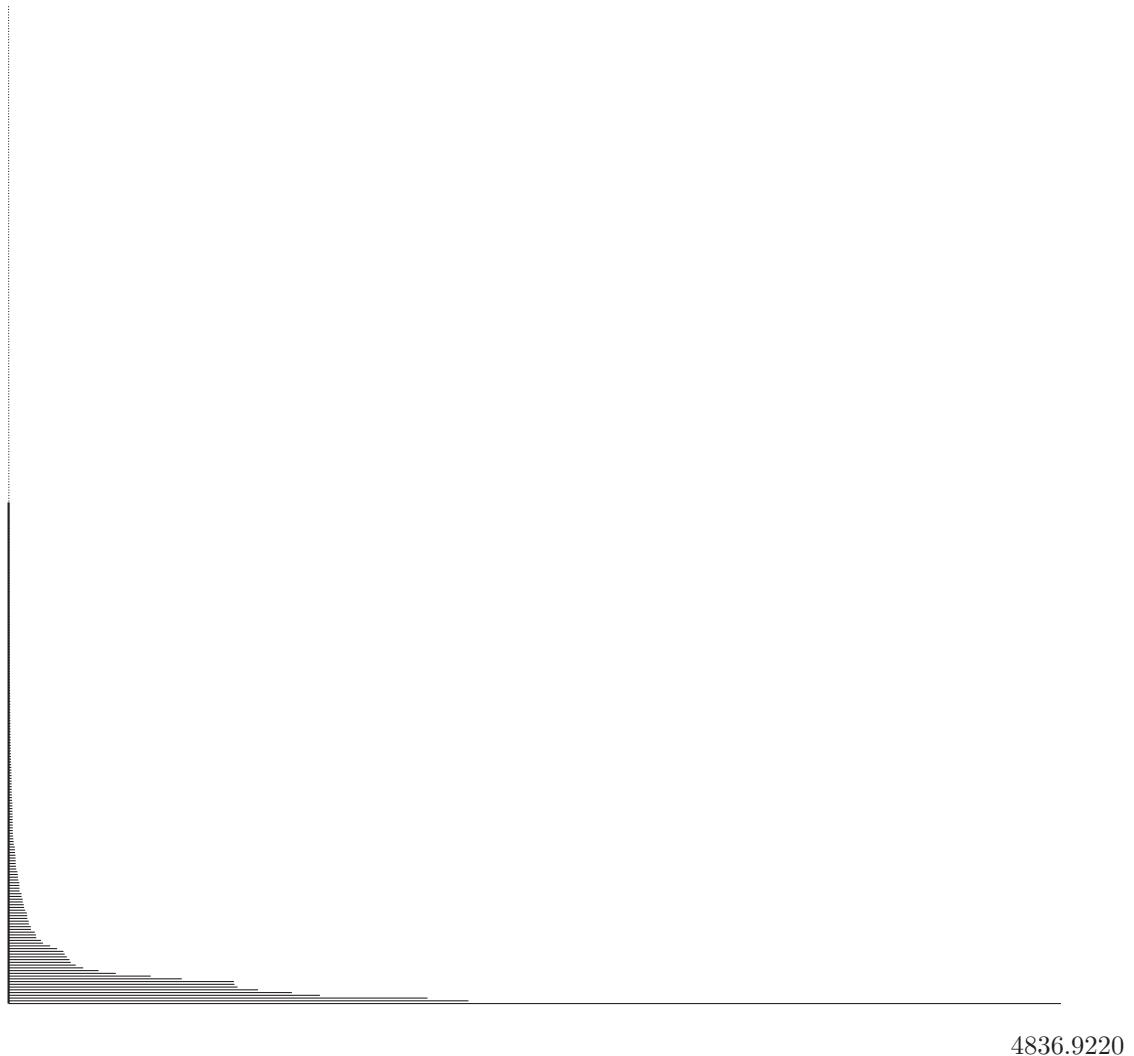
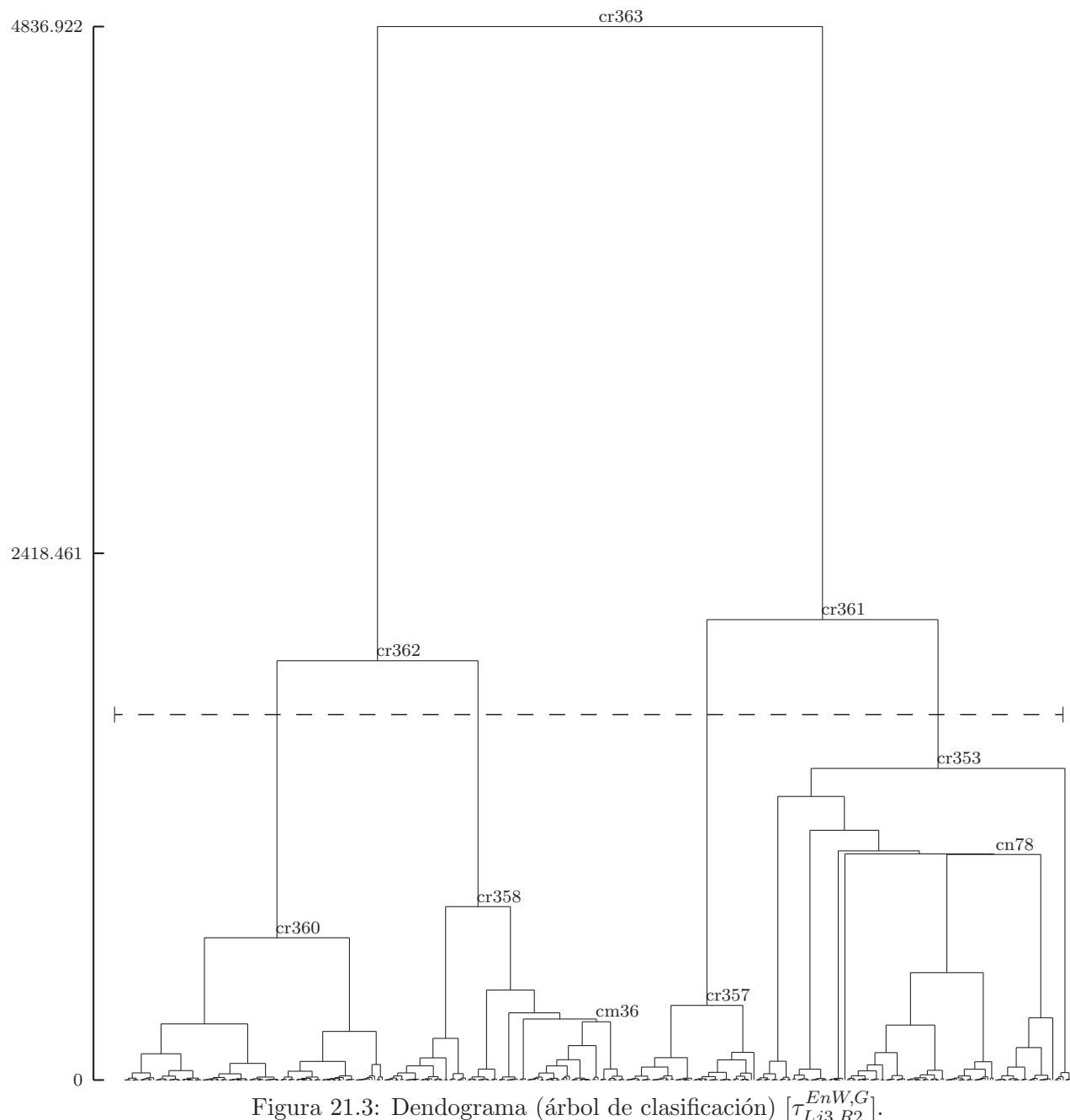
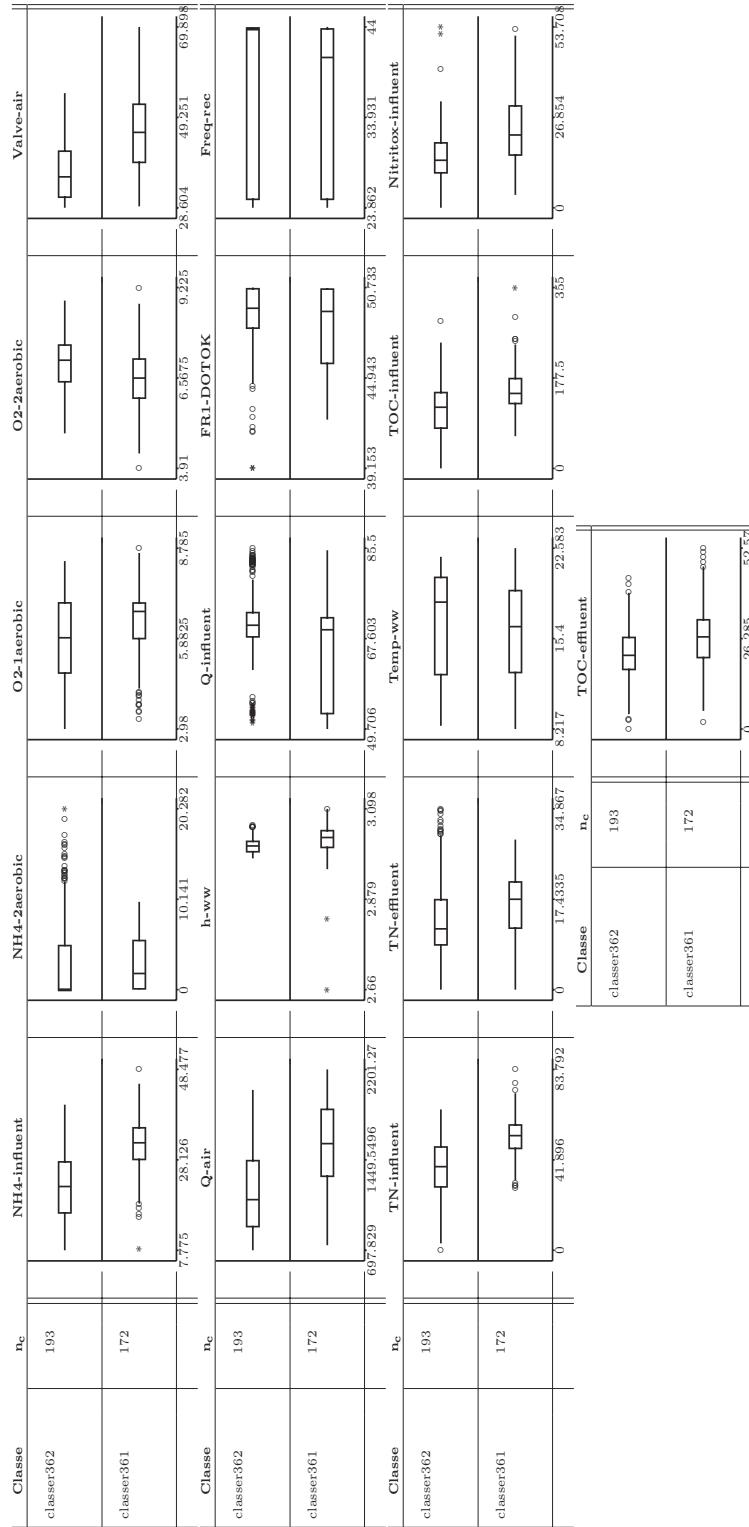
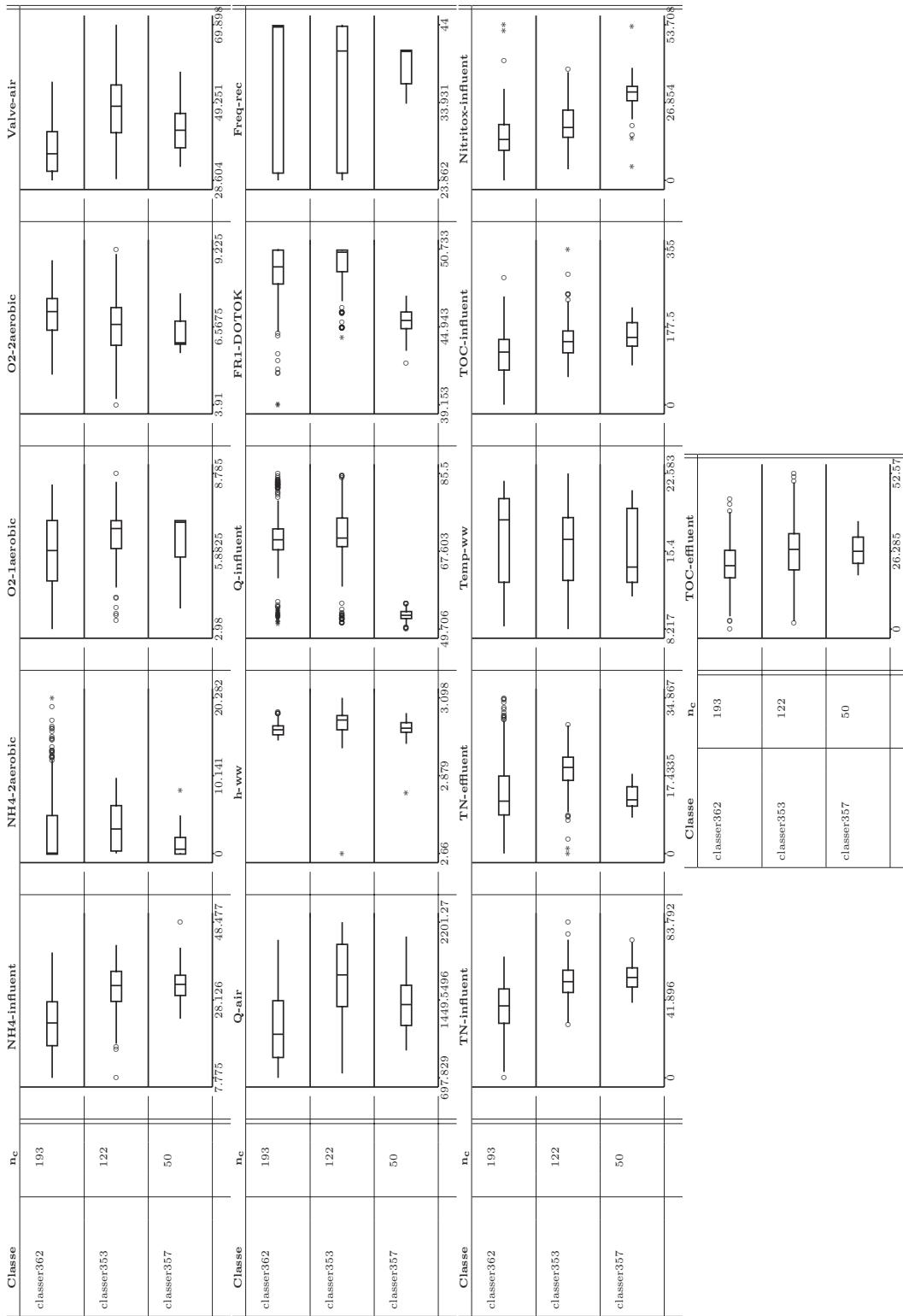


Figura 21.2: Gráfico de inercia interna de las clases $[\tau_{Lj3,R2}^{EnW,G}]$.



Figura 21.4: Análisis descriptivo por clases para $[P2_{Lj3,R2}^{EnW,G}]$.

Figura 21.5: Análisis descriptivo por clases para $[P3^{EnW,G}_{Lj3,R2}]$.

21.4 Interpretación validada por el experto

Los expertos han proporcionado la siguiente interpretación para la partición de referencia $\mathcal{P}4_{Lj3,R2}^{EnW,G} = \{Cr358, Cr360, Cr353, Cr357\}$:

- *Cr353*, represents the plant operation under the high load. In this case influent nitrogen concentrations are high and also influent flow rate is quite high as well. Even though the oxygen concentration in the aerobic tanks are high this can not decrease the effluent nitrogen concentrations. It means that when the plant is overloaded high effluent concentrations at the effluent of the plant can be expected.
- *Cr357*, represents the situation when the influent flow rate is low, that is, when the hydraulic retention time of the plant is high. In this case we get quite low effluent nitrogen concentrations if of course oxygen concentration in the aerobic tank is high enough. It means when the influent flow rate to the plant is low the effluent concentrations of the plant can be obtained at the low level if the oxygen concentration in the aerobic tanks is high.
- *Cr358*, explains the situation when the wastewater temperature is low. In this case nitrogen removal efficiency of the plant is rather low. This is so because microorganisms in the tanks don't work so intensively in cold conditions and therefore higher concentrations at the effluent of the plant can be expected.
- *Cr360*, shows the situation when the wastewater temperature is high. In warmer conditions the microorganisms in the plant work faster, so the effluent nitrogen concentrations can be low even when the oxygen concentrations in the aerobic tanks are quite low.

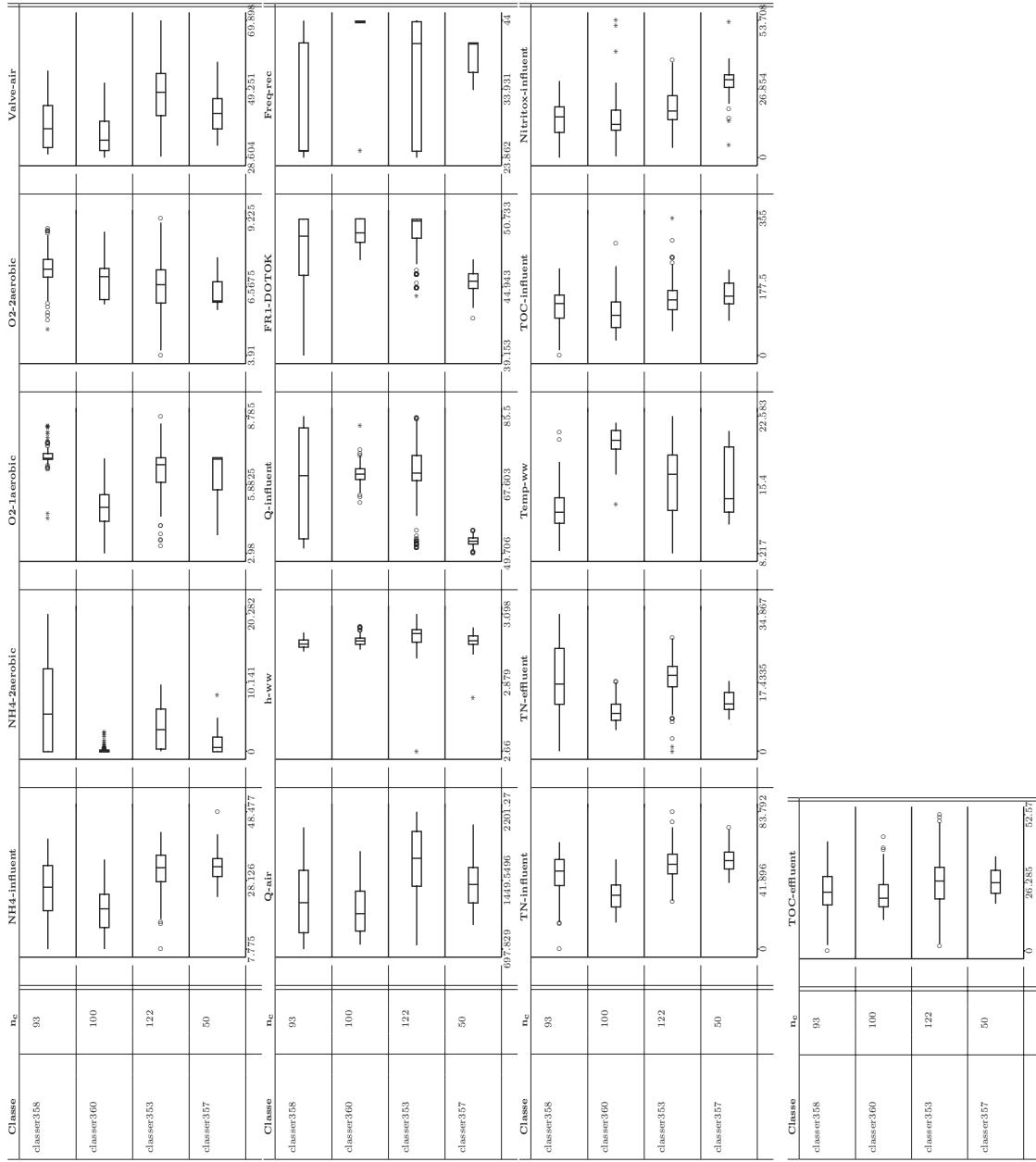


Figura 21.6: Análisis descriptivo por clases para $[P4_{Lj3,R2}^{EnW,G}]$.

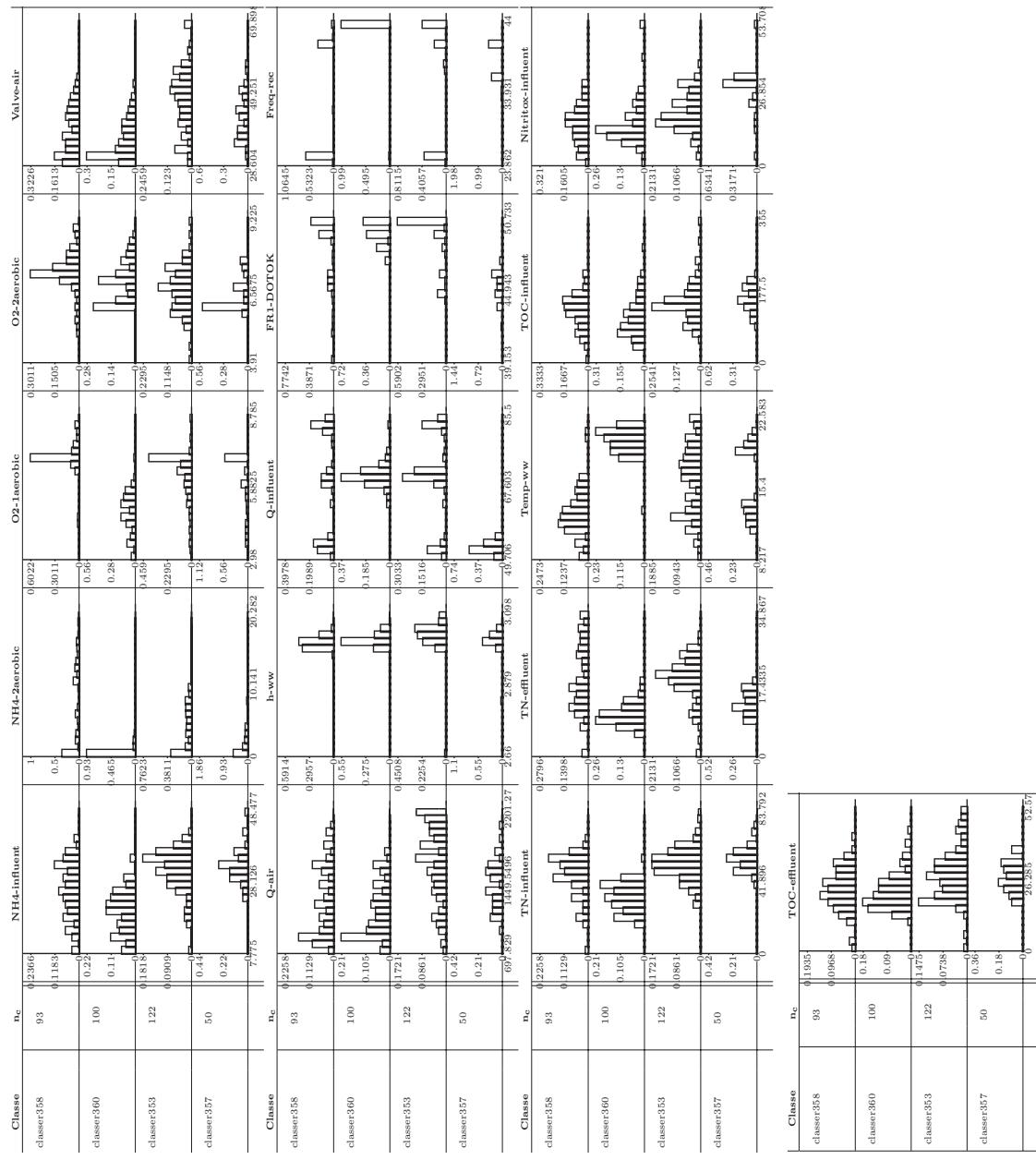


Figura 21.7: Análisis descriptivo por clases para $[P4_{Lj3,R2}^{EnW,G}]$.

Capítulo 22

Aplicación, planta eslovena

22.1 Interpretación de \mathcal{P}_4 utilizando Best global concept and Close-World Assumption

1. $\xi = 2$: Así, $\mathcal{P}2_{Lj3,R2}^{EnW,G} = \{Cr361, Cr362\}$. La raíz del árbol $Cr363$ tiene 2 hijos:

$$Cr363 \left\{ \begin{array}{l} Cr361 \\ Cr362 \end{array} \right.$$

2. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición $\mathcal{P}2_{Lj3,R2}^{EnW,G} = \{Cr361, Cr362\}$ se presenta en el Apéndice H.1
3. Con el BbIR se obtienen los sistemas de reglas $\mathcal{R}(X_k, \mathcal{P}_2^*)$ $k \in 1 : K$, inducidos para todas las variables numéricas que caracteriza ambas clases, y $\mathcal{R}(\mathcal{P}_2^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_2^*)$, se presenta en el Apéndice H.2
4. Como resultado de los pasos anteriores, se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_2^*)$ y se trabaja con $\mathcal{S}(\mathcal{P}_2^*) = \bigcup_{k=1}^K \mathcal{S}(X_k, \mathcal{P}_2^*)$ donde $\mathcal{S}(\mathcal{P}_2^*) \subseteq \mathcal{R}(\mathcal{P}_2^*)$.

$$\begin{aligned} \mathcal{S}(\mathcal{P}_2^*) = \{ & r_{3,Cr361}^{NH4-influent} : x_{NH4-influent,i} \in (40.541, 48.477] \xrightarrow{1.0} Cr361 , \\ & r_{3,Cr361}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (8.371, 8.785] \xrightarrow{1.0} Cr361 , \\ & r_{1,Cr361}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 4.94] \xrightarrow{1.0} Cr361 , \\ & r_{3,Cr361}^{Valve-air} : x_{Valve-air,i} \in (54.777, 69.898] \xrightarrow{1.0} Cr361 , \\ & r_{3,Cr361}^{Q-air} : x_{Q-air,i} \in (2030.77, 2201.27] \xrightarrow{1.0} Cr361 , \\ & r_{3,Cr361}^{h-ww} : x_{h-ww,i} \in (3.058, 3.098] \xrightarrow{1.0} Cr361 , \\ & r_{1,Cr361}^{Q-influent} : x_{Q-influent,i} \in [49.706, 50.99] \xrightarrow{1.0} Cr361 , \\ & r_{3,Cr361}^{Freq-rec} : x_{Freq-rec,i} \in (43.97, 44.0] \xrightarrow{1.0} Cr361 , \\ & r_{3,Cr361}^{TN-influent} : x_{TN-influent,i} \in (65.25, 83.792] \xrightarrow{1.0} Cr361 , \\ & r_{3,Cr361}^{Temp-ww} : x_{Temp-ww,i} \in (21.896, 22.583] \xrightarrow{1.0} Cr361 , \\ & r_{3,Cr361}^{TOC-influent} : x_{TOC-influent,i} \in (290.212, 355.0] \xrightarrow{1.0} Cr361 , \\ & r_{3,Cr361}^{TOC-effluent} : x_{TOC-effluent,i} \in (44.053, 52.57] \xrightarrow{1.0} Cr361 , \\ & r_{1,Cr362}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.972) \xrightarrow{1.0} Cr362 , \\ & r_{3,Cr362}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (9.846, 20.282] \xrightarrow{1.0} Cr362 , \\ & r_{1,Cr362}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 3.297] \xrightarrow{1.0} Cr362 , \\ & r_{1,Cr362}^{Valve-air} : x_{Valve-air,i} \in [28.604, 28.934) \xrightarrow{1.0} Cr362 , \\ & r_{1,Cr362}^{Q-air} : x_{Q-air,i} \in [697.829, 739.819) \xrightarrow{1.0} Cr362 , \\ & r_{3,Cr362}^{Q-influent} : x_{Q-influent,i} \in (85.092, 85.5] \xrightarrow{1.0} Cr362 , \end{aligned}$$

$$\begin{aligned}
r_{1,Cr362}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 42.276] &\xrightarrow{1.0} Cr362 , \\
r_{1,Cr362}^{TN-influent} : x_{TN-influent,i} \in [0.0, 28.792] &\xrightarrow{1.0} Cr362 , \\
r_{3,Cr362}^{TN-effluent} : x_{TN-effluent,i} \in [28.933, 34.867] &\xrightarrow{1.0} Cr362 , \\
r_{1,Cr362}^{TOC-influent} : x_{TOC-influent,i} \in [0.0, 63.22] &\xrightarrow{1.0} Cr362 , \\
r_{1,Cr362}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.0, 3.833] &\xrightarrow{1.0} Cr362 , \\
r_{1,Cr362}^{TOC-effluent} : x_{TOC-effluent,i} \in [0.0, 2.014] &\xrightarrow{1.0} Cr362 \quad \}
\end{aligned}$$

5. El Cuadro 22.1 muestra la cobertura relativa de las reglas de $\mathcal{S}(\mathcal{P}_2^*)$. La regla con mayor cobertura relativa y $p_{sc} = 1$ de $\mathcal{S}(\mathcal{P}_2^*)$ es $r_{1,Cr362}^{TN-influent}$ con una $CovR(r) = 22,80\%$:

$$r_{1,Cr362}^{TN-influent} : x_{TN-influent,i} \in [0.0, 28.792] \xrightarrow{1.0} Cr362$$

Así (ver ecuación (10.1)):

$$A^2 = A^{2,TN-influent} = "x_{TN-influent,i} \in [0.0, 28.792]"$$

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr361}^{NH4-influent}$	5	2,91%
$r_{3,Cr361}^{O2-1aerobic}$	1	0,58%
$r_{1,Cr361}^{O2-2aerobic}$	2	1,16%
$r_{3,Cr361}^{Valve-air}$	28	16,28%
$r_{3,Cr361}^{Q-air}$	27	15,70%
$r_{3,Cr361}^{h-ww}$	16	9,30%
$r_{1,Cr361}^{Q-influent}$	6	3,49%
$r_{3,Cr361}^{Freq-rec}$	3	1,74%
$r_{3,Cr361}^{TN-influent}$	9	5,23%
$r_{3,Cr361}^{Temp-ww}$	5	2,91%
$r_{3,Cr361}^{TOC-influent}$	20	11,63%
$r_{3,Cr361}^{TOC-effluent}$	10	5,81%
$r_{1,Cr362}^{NH4-influent}$	3	1,55%
$r_{3,Cr362}^{NH4-2aerobic}$	38	19,69%
$r_{1,Cr362}^{O2-1aerobic}$	4	2,07%
$r_{1,Cr362}^{Valve-air}$	5	2,59%
$r_{1,Cr362}^{Q-air}$	2	1,04%
$r_{3,Cr362}^{Q-influent}$	1	0,52%
$r_{1,Cr362}^{FR1-DOTOK}$	5	2,59%
$r_{1,Cr362}^{TN-influent}$	44	22,80%
$r_{3,Cr362}^{TN-effluent}$	14	7,25%
$r_{1,Cr362}^{TOC-influent}$	20	10,36%
$r_{1,Cr362}^{Nitritox-influent}$	4	2,07%
$r_{1,Cr362}^{TOC-effluent}$	1	0,52%

Tabla 22.1: Cobertura relativa de $\mathcal{S}(\mathcal{P}_2^*)$.

En este caso no hay empate, ver ecuación (10.2) y ver ecuación (10.3), y por tanto:

$$\bullet A_{Cr362}^2 = A^2 = A_{Cr362}^{2,TN-influent} = "x_{TN-influent,i} \in [0.0, 28.792]"$$

- $A_{Cr361}^2 = \neg A^2 = A_{Cr361}^{2,TN-influent} = "x_{TN-influent,i} \in [28.792, 83.792]"$

6. Asociando una regla a cada clase

$$\begin{aligned} \mathbb{R}(\mathcal{P}_2) = \{ & r_{Cr362} : x_{TN-influent,i} \in [0.0, 28.792] \xrightarrow{1.0} Cr362, \\ & r_{Cr361} : x_{TN-influent,i} \in [28.792, 83.792] \xrightarrow{0.54} Cr361 \\ \} \end{aligned}$$

Donde $p_{sCr361} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{172}{321} = 0.54$

7. $\xi = 3$: Así, $\mathcal{P}_3^{EnW,G}_{Lj3,R2} = \{Cr362, Cr353, Cr357\}$. La clase $Cr361$ es la que se divide en 2 hijos y la clase $Cr362$ es la que ya estaba en la partición anterior:

$$Cr361 \left\{ \begin{array}{l} Cr353 \\ Cr357 \end{array} \right.$$

Así $C_i^3 = Cr353$; $C_j^3 = Cr357$; $C_t^2 = Cr361$.

8. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_3^* = \{Cr353, Cr357\}$, donde $\mathcal{P}_3^* \subseteq \mathcal{P}_3^{EnW,G}_{Lj3,R2}$, se presenta en el Apéndice H.3.
9. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_3^*)$ $k \in 1 : K$.
10. Se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_3^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_3^*)$ que se se presenta en el Apéndice H.4 y se trabaja con $\mathcal{S}(\mathcal{P}_3^*) = \bigcup_{k=1}^K \mathcal{S}(X_k, \mathcal{P}_3^*)$ donde $\mathcal{S}(\mathcal{P}_3^*) \subseteq \mathcal{R}(\mathcal{P}_3^*)$ el cual es:

$$\begin{aligned} \mathcal{S}(\mathcal{P}_3^*) = \{ & r_{3,Cr357}^{NH4-influent} : x_{NH4-influent,i} \in (42.511, 48.477] \xrightarrow{1.0} Cr357, \\ & r_{1,Cr357}^{Q-influent} : x_{Q-influent,i} \in [49.706, 51.123] \xrightarrow{1.0} Cr357, \\ & r_{1,Cr357}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [42.276, 44.167] \xrightarrow{1.0} Cr357, \\ & r_{3,Cr357}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (38.333, 53.0] \xrightarrow{1.0} Cr357, \\ & r_{1,Cr353}^{NH4-influent} : x_{NH4-influent,i} \in [7.972, 23.23] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (8.262, 9.846] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (7.018, 8.785] \xrightarrow{1.0} Cr353, \\ & r_{1,Cr353}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 5.675] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{Valve-air} : x_{Valve-air,i} \in (57.442, 69.898] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{Q-air} : x_{Q-air,i} \in (2062.554, 2201.27] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{h-ww} : x_{h-ww,i} \in (3.055, 3.098] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{Q-influent} : x_{Q-influent,i} \in (55.666, 85.092] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (47.265, 50.7] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{Freq-rec} : x_{Freq-rec,i} \in (40.633, 44.0] \xrightarrow{1.0} Cr353, \\ & r_{1,Cr353}^{TN-influent} : x_{TN-influent,i} \in [28.792, 40.417] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{TN-effluent} : x_{TN-effluent,i} \in (17.837, 28.933] \xrightarrow{1.0} Cr353, \\ & r_{1,Cr353}^{Temp-ww} : x_{Temp-ww,i} \in [8.217, 11.235] \xrightarrow{1.0} Cr353, \\ & r_{1,Cr353}^{TOC-influent} : x_{TOC-influent,i} \in [63.22, 89.833] \xrightarrow{1.0} Cr353, \\ & r_{1,Cr353}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [3.833, 4.875] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{TOC-effluent} : x_{TOC-effluent,i} \in (36.476, 52.57] \xrightarrow{1.0} Cr353 \} \end{aligned}$$

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr357}^{NH4-influent}$	1	2,00%
$r_{1,Cr357}^{Q-influent}$	6	12,00%
$r_{1,Cr357}^{FR1-DOTOK}$	8	16,00%
$r_{3,Cr357}^{Nitritox-influent}$	1	2,00%
$r_{1,Cr353}^{NH4-influent}$	11	9,02%
$r_{3,Cr353}^{NH4-2aerobic}$	9	7,38%
$r_{3,Cr353}^{O2-1aerobic}$	9	7,38%
$r_{1,Cr353}^{O2-2aerobic}$	12	9,84%
$r_{3,Cr353}^{Valve-air}$	11	9,02%
$r_{3,Cr353}^{Q-air}$	22	18,03%
$r_{3,Cr353}^{h-ww}$	20	16,39%
$r_{3,Cr353}^{Q-influent}$	100	81,97%
$r_{3,Cr353}^{FR1-DOTOK}$	98	80,33%
$r_{3,Cr353}^{Freq-rec}$	49	40,16%
$r_{1,Cr353}^{TN-influent}$	16	13,11%
$r_{3,Cr353}^{TN-effluent}$	81	66,39%
$r_{1,Cr353}^{Temp-ww}$	11	9,02%
$r_{1,Cr353}^{TOC-influent}$	13	10,66%
$r_{1,Cr353}^{Nitritox-influent}$	2	1,64%
$r_{3,Cr353}^{TOC-effluent}$	17	13,93%

Tabla 22.2: Cobertura relativa de $\mathcal{S}(\mathcal{P}_3^*)$.

11. El Cuadro 22.2 muestra la cobertura relativa de las reglas de $\mathcal{S}(\mathcal{P}_3^*)$. La regla con mayor cobertura relativa y $p_{sc} = 1$ de $\mathcal{S}(\mathcal{P}_3^*)$ es $r_{3,Cr353}^{Q-influent}$, con una cobertura relativa $CovR = 81,97\%$:

$$r_{3,Cr353}^{Q-influent} : x_{Q-influent,i} \in (55.666, 85.092] \xrightarrow{1.0} Cr353$$

En este caso no hay empate, ver ecuación (10.2) y (10.3), así:

- $A_{Cr353}^{*3} = A_{Cr353}^{3,Q-influent} = "x_{Q-influent,i} \in (55.666, 85.092]"$
- $A_{Cr357}^{*3} = A_{Cr357}^{3,Q-influent} = \neg A_{Cr353}^{3,Q-influent} = "x_{Q-influent,i} \in [49.706, 55.666]"$

12. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

- $A_{Cr362}^3 = A_{Cr362}^2$
- $A_{Cr362}^3 = "x_{TN-influent,i} \in [0.0, 28.792]"$
- $A_{Cr353}^3 = A_{Cr361}^2 \wedge A_{Cr353}^{*3}$
- $A_{Cr353}^3 = "x_{TN-influent,i} \in [28.79, 83.79]" \wedge "x_{Q-influent,i} \in (55.666, 85.092]"$
- $A_{Cr357}^3 = A_{Cr361}^2 \wedge A_{Cr357}^{*3}$
- $A_{Cr357}^3 = "x_{TN-influent,i} \in [28.79, 83.79]" \wedge "x_{Q-influent,i} \in [49.706, 55.666]"$

Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$p_{sCr353} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{99}{220} = 0.4587$$

$$p_{sCr357} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{46}{98} = 0.47$$

$$\mathbb{R}(\mathcal{P}_3) = \{ \quad r_{Cr362} : \quad x_{TN-influent,i} \in [0.0, 28.792) \xrightarrow{1.0} Cr362,$$

$$\begin{aligned} r_{Cr353} : \quad x_{TN-influent,i} &\in [28.792, 83.792] \wedge \\ x_{Q-influent,i} &\in (55.666, 85.092] \xrightarrow{0.46} Cr353, \end{aligned}$$

$$\begin{aligned} r_{Cr357} : \quad x_{TN-influent,i} &\in [28.792, 83.792] \wedge \\ x_{Q-influent,i} &\in [49.706, 55.666] \xrightarrow{0.47} Cr357 \quad \} \end{aligned}$$

13. $\xi = 4$: Así, $\mathcal{P}_4^{EnW,G}_{Lj3,R2} = \{Cr358, Cr360, Cr353, Cr357\}$. La clase $Cr362$ de $\mathcal{P}_3^{EnW,G}_{Lj3,R2}$ tiene 2 hijos:

$$Cr362 \left\{ \begin{array}{l} Cr358 \\ Cr360 \end{array} \right.$$

Así $C_i^4 = Cr358$; $C_j^4 = Cr360$; $C_t^3 = Cr362$

14. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_4^* = \{Cr358, Cr360\}$, donde $\mathcal{P}_4^* \subseteq \mathcal{P}_4^{EnW,G}_{Lj3,R2}$, se presenta en el Apéndice H.5.
15. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_4^*)$, $k \in 1 : K$.
16. Se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_4^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_4^*)$ que se presenta en el Apéndice H.6 y se trabaja con el sistema de reglas seguro $\mathcal{S}(\mathcal{P}_4^*) = \bigcup_{k=1}^K \mathcal{S}(X_k, \mathcal{P}_4^*)$ donde $\mathcal{S}(\mathcal{P}_4^*) \subseteq \mathcal{R}(\mathcal{P}_4^*)$ el cual es:

$$\begin{aligned} \mathcal{S}(\mathcal{P}_4^*) = \{ & \quad r_{1,Cr360}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.79) \xrightarrow{1.0} Cr360, \\ & r_{1,Cr360}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 4.479) \xrightarrow{1.0} Cr360, \\ & r_{1,Cr360}^{Valve-air} : x_{Valve-air,i} \in [28.604, 29.57) \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{h-ww} : x_{h-ww,i} \in (3.039, 3.058] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (50.692, 50.733] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{Temp-ww} : x_{Temp-ww,i} \in (20.928, 21.896] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{TOC-influent} : x_{TOC-influent,i} \in (225.293, 290.212] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (30.0, 53.708] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{TOC-effluent} : x_{TOC-effluent,i} \in (42.251, 44.053] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr358}^{NH4-influent} : x_{NH4-influent,i} \in (34.353, 40.541] \xrightarrow{1.0} Cr358, \\ & r_{3,Cr358}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (2.856, 20.282] \xrightarrow{1.0} Cr358, \\ & r_{3,Cr358}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (6.998, 8.371] \xrightarrow{1.0} Cr358, \\ & r_{1,Cr358}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [4.94, 5.889) \xrightarrow{1.0} Cr358, \\ & r_{3,Cr358}^{Valve-air} : x_{Valve-air,i} \in (51.168, 54.777] \xrightarrow{1.0} Cr358, \\ & r_{3,Cr358}^{Q-air} : x_{Q-air,i} \in (1770.852, 2030.77] \xrightarrow{1.0} Cr358, \\ & r_{1,Cr358}^{h-ww} : x_{h-ww,i} \in [2.978, 2.984) \xrightarrow{1.0} Cr358, \end{aligned}$$

$$\begin{aligned}
r_{1,Cr358}^{Q-influent} : x_{Q-influent,i} \in [50.99, 63.038] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 47.2] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Freq-rec} : x_{Freq-rec,i} \in [23.863, 24.899] &\xrightarrow{1.0} Cr358 , \\
r_{3,Cr358}^{TN-influent} : x_{TN-influent,i} \in (54.792, 65.25] &\xrightarrow{1.0} Cr358 , \\
r_{3,Cr358}^{TN-effluent} : x_{TN-effluent,i} \in (17.788, 34.867] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Temp-ww} : x_{Temp-ww,i} \in [8.472, 13.327] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{TOC-influent} : x_{TOC-influent,i} \in [0.0, 38.888] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.0, 0.542] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{TOC-effluent} : x_{TOC-effluent,i} \in [0.0, 11.879] &\xrightarrow{1.0} Cr358 \quad \}
\end{aligned}$$

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr360}^{O2-1aerobic}$	28	28,00%
$r_{1,Cr360}^{Valve-air}$	5	5,00%
$r_{3,Cr360}^{h-ww}$	9	9,00%
$r_{3,Cr360}^{FR1-DOTOK}$	3	3,00%
$r_{3,Cr360}^{Temp-ww}$	29	29,00%
$r_{3,Cr360}^{TOC-influent}$	1	1,00%
$r_{3,Cr360}^{Nitritox-influent}$	3	3,00%
$r_{3,Cr360}^{TOC-effluent}$	1	1,00%
$r_{3,Cr358}^{NH4-2aerobic}$	54	58,06%
$r_{3,Cr358}^{O2-1aerobic}$	54	58,06%
$r_{1,Cr358}^{O2-2aerobic}$	5	5,38%
$r_{3,Cr358}^{Valve-air}$	3	3,23%
$r_{3,Cr358}^{Q-air}$	9	9,68%
$r_{1,Cr358}^{h-ww}$	6	6,45%
$r_{1,Cr358}^{Q-influent}$	33	35,48%
$r_{1,Cr358}^{FR1-DOTOK}$	34	36,56%
$r_{1,Cr358}^{Freq-rec}$	13	13,98%
$r_{3,Cr358}^{TN-influent}$	19	20,43%
$r_{3,Cr358}^{TN-effluent}$	38	40,86%
$r_{1,Cr358}^{Temp-ww}$	62	66,67%
$r_{1,Cr358}^{TOC-influent}$	1	1,08%
$r_{1,Cr358}^{Nitritox-influent}$	1	1,08%
$r_{1,Cr358}^{TOC-effluent}$	6	6,45%

Tabla 22.3: Cobertura relativa de $\mathcal{S}(\mathcal{P}_4^*)$.

17. El Cuadro 22.3 muestra la cobertura relativa de las reglas de $\mathcal{S}(\mathcal{P}_4^*)$. La regla con mayor cobertura relativa de $\mathcal{S}(\mathcal{P}_4^*)$ es $r_{1,Cr358}^{Temp-ww}$, con una cobertura relativa $CovR = 66,67\%$:

$$r_{1,Cr358}^{Temp-ww} : x_{Temp-ww,i} \in [8.472, 13.327] \xrightarrow{1.0} Cr358$$

En este caso no hay empate, ver ecuación (10.2) y ver ecuación (10.3), y por tanto:

- $A_{Cr358}^{*4} = A_{Cr358}^{4,Temp-ww} = "x_{Temp-ww,i} \in [8.472, 13.327]"$
- $A_{Cr360}^{*4} = A_{Cr360}^{4,Temp-ww} = \neg A_{Cr358}^{4,Temp-ww} = "x_{Temp-ww,i} \in [13.327, 21.896]"$

18. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

- $A_{Cr353}^4 = A_{Cr353}^3 = A_{Cr361}^2 \wedge A_{Cr353}^{*3}$
 $A_{Cr353}^4 = "x_{TN-influent,i} \in [28.79, 83.79]" \wedge "x_{Q-influent,i} \in (55.666, 85.092]"$
- $A_{Cr357}^4 = A_{Cr357}^3 = A_{Cr361}^2 \wedge A_{Cr357}^{*3}$
 $A_{Cr357}^4 = "x_{TN-influent,i} \in [28.79, 83.79]" \wedge "x_{Q-influent,i} \in [49.706, 55.666]"$
- $A_{Cr358}^4 = A_{Cr362}^2 \wedge A_{Cr358}^{*4}$
 $A_{Cr358}^4 = "x_{TN-influent,i} \in [0.0, 28.792]" \wedge "x_{Temp-ww,i} \in [8.472, 13.327]"$
- $A_{Cr360}^4 = A_{Cr362}^2 \wedge A_{Cr360}^{*4}$
 $A_{Cr360}^4 = "x_{TN-influent,i} \in [0.0, 28.792]" \wedge "x_{Temp-ww,i} \in [13.327, 21.896]"$

Así pues asociando una regla compuesta a cada clase

$$\begin{aligned} \mathbb{R}(\mathcal{P}_4) = \{ & r_{Cr353} : x_{TN-influent,i} \in [28.792, 83.792] \wedge \\ & x_{Q-influent,i} \in (55.666, 85.092] \xrightarrow{0.45} Cr353, \\ & r_{Cr357} : x_{TN-influent,i} \in [28.792, 83.792] \wedge \\ & x_{Q-influent,i} \in [49.706, 55.666] \xrightarrow{0.47} Cr357, \\ & r_{Cr358} : x_{TN-influent,i} \in [0.0, 28.792] \wedge \\ & x_{Temp-ww,i} \in [8.472, 13.327] \xrightarrow{1.0} Cr358, \\ & r_{Cr360} : x_{TN-influent,i} \in [0.0, 28.792] \wedge \\ & x_{Temp-ww,i} \in [13.327, 21.896] \xrightarrow{0.86} Cr360 \\ & \} \end{aligned}$$

22.1.1 Interpretación final:

Que correspondería a la conceptualización:

- Cr353: “ $TN - influent$ no es bajo y $Q - influent$ es alto”
- Cr357: “ $TN - influent$ no es bajo y $Q - influent$ no es alto”
- Cr358: “ $TN - influent$ es bajo y $Temp - ww$ es bajo”
- Cr360: “ $TN - influent$ es bajo y $Temp - ww$ no es bajo”

22.1.2 Evaluación de la propuesta

Una vez determinados los conceptos asociados a cada clase se ha procedido a evaluar cual(es) era(n) satisfecho(s) por cada objeto de \mathcal{I} con tal de evaluar la correcta correspondencia entre la muestra y la conceptualización inducida. La Tabla 22.4 muestra los resultados de calcular el soporte $Sup(r)$, ver ecuación (9.1), la cobertura relativa $CovR(r)$, ver ecuación (9.5) y la confianza $p(r)$, ver ecuación (9.2), de cada una de las reglas compuestas inducidas para cada clase de la partición final.

Ruler	Consec.	$\#\{i \in A_C^\xi\}$	$\#\{i \in A_C^\xi \cap i \in C\}$	n_c	$p(r)$	$Sup(r)$	$CovR(r)$
$rCr353$	$Cr353$	220	99	122	45,00%	60,27%	81,15%
$rCr357$	$Cr357$	98	46	50	46,94%	26,85%	92,00%
$rCr358$	$Cr358$	6	6	93	100%	1,64%	6,45%
$rCr360$	$Cr360$	38	33	100	86,84%	10,41%	33,00%
<i>Media</i>					69,70%		53,15%
<i>Suma</i>		362	184	365		99,18%	
<i>CovGlobal(\mathbb{R})</i>							50,4%

Tabla 22.4: Evaluación: Best global concept and Close-World Assumption.

La evaluación de la regla se hace con respecto al total de objetos de la base de datos, es evidente que no habrá inconsistencias al evaluar cada regla por separado ya que los intervalos que conforman cada reglas compuesta, son disjuntos entre una clase y otra debido a la forma en que se construyen los conceptos, ver ecuaciones (10.1), (10.3), (10.2), (10.4), (10.5), (10.6) y (10.7), (10.8).

Si se observa la Tabla 22.4, se tiene que el soporte se obtiene dividiendo cada celda de la tercera columna entre el total de objetos de la base de datos (365), la cobertura relativa se obtiene dividiendo cada celda de la cuarta columna entre la correspondiente celda de la quinta columna y la confianza dividiendo las celdas de la cuarta entre las correspondientes celdas de la tercera columna.

Como hay 3 clases en donde $\#\{i \in A_C^\xi\} > \#\{i \in A_C^\xi \cap i \in C\}$ se puede concluir que hay objetos mal asignados, es decir objetos que no satisfacen el consecuente de la regla, pero que satisfacen el antecedente de esta. Este valor no se considera en el soporte ni en la cobertura relativa, pero queda de manifiesto en la confianza ya que cuando se obtienen confianzas del 100% significa que todos los objetos que satisfacen el antecedente de la regla, también satisfacen simultáneamente el antecedente y el consecuente de la regla y por lo tanto pertenecen a la clase que la regla asigna. Con lo cual el se puede concluir, en este caso, que el número de objetos correctamente asignados por clase viene dado por $\#\{i \in A_C^\xi \cap i \in C\}$ y el número de objetos mal asignados por clases, en este caso, se puede calcular $\#\{i \in A_C^\xi\} - \#\{i \in A_C^\xi \cap i \in C\}$. El porcentaje total de objetos correctamente asignados se puede obtener dividiendo 184 entre 365, 50,4%. Las confianzas en promedio rondan el 70%, lo cual se puede considerar como bueno.

También es interesante observar que el número de objetos asignados por las reglas compuestas asociadas a cada clase se puede obtener a partir de $\#\{i \in A_C^\xi\}$ y el valor total viene cuantificado por el soporte, ya que si la suma total de los soportes por clase corresponde al porcentaje de objetos asignados, cuando este es menor que el 100% significa que hay objetos sin asignar. En la Tabla 23.2 se puede observar que al restar a 365 la suma de $\#\{i \in A_C^\xi\}$ sólo hay 3 objetos que no han sido asignados, con lo cual se tiene un promedio del soporte cercano al 100%, esta conclusión se puede realizar sólo en el caso que no ocurran inconsistencias y para soportes menores o iguales al 100%.

La relación entre cobertura relativa y confianza es equilibrada en el sentido que no se diferencian entre si en muchos puntos porcentuales, asignando con esta propuesta casi el total de objetos.

22.2 Interpretación de \mathcal{P}_4 utilizando Best local concept and no Close-World Assumption:

1. $\xi = 2$: Así, $\mathcal{P}2_{Lj3,R2}^{EnW,G} = \{Cr361, Cr362\}$. La raíz del árbol $Cr363$ tiene 2 hijos:

$$Cr363 \left\{ \begin{array}{l} Cr361 \\ Cr362. \end{array} \right.$$

2. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición $\mathcal{P}2_{Lj3,R2}^{EnW,G} = \{Cr361, Cr362\}$ se presenta en el Apéndice H.1
3. El sistema de reglas inducido para todas las variables numéricas, con el BbIR, que caracteriza ambas clases, $\mathcal{R}(\mathcal{P}_2^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_2^*)$, se presenta en el Apéndice H.2
4. Como resultado de los pasos anteriores, se consideran los siguientes sistemas de reglas, con reglas seguras $\mathcal{S}_{Cr361}(\mathcal{P}_2^*) \subseteq \mathcal{S}(\mathcal{P}_2^*)$, donde $\mathcal{S}(\mathcal{P}_2^*) \subseteq \mathcal{R}(\mathcal{P}_2^*)$:

$$\mathcal{S}_{Cr361}(\mathcal{P}_2^*) = \{ \begin{array}{ll} r_{3,Cr361}^{NH4-influent} : x_{NH4-influent,i} \in (40.541, 48.477] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (8.371, 8.785] \xrightarrow{1.0} Cr361 , \\ r_{1,Cr361}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 4.94] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{Valve-air} : x_{Valve-air,i} \in (54.777, 69.898] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{Q-air} : x_{Q-air,i} \in (2030.77, 2201.27] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{h-ww} : x_{h-ww,i} \in (3.058, 3.098] \xrightarrow{1.0} Cr361 , \\ r_{1,Cr361}^{Q-influent} : x_{Q-influent,i} \in [49.706, 50.99] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{Freq-rec} : x_{Freq-rec,i} \in (43.97, 44.0] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{TN-influent} : x_{TN-influent,i} \in (65.25, 83.792] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{Temp-ww} : x_{Temp-ww,i} \in (21.896, 22.583] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{TOC-influent} : x_{TOC-influent,i} \in (290.212, 355.0] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{TOC-effluent} : x_{TOC-effluent,i} \in (44.053, 52.57] \xrightarrow{1.0} Cr361 \end{array} \}$$

y $\mathcal{S}_{Cr362}(\mathcal{P}_2^*) \subseteq \mathcal{S}(\mathcal{P}_2^*)$.

$$\mathcal{S}_{Cr362}(\mathcal{P}_2^*) = \{ \begin{array}{ll} r_{1,Cr362}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.972] \xrightarrow{1.0} Cr362 , \\ r_{3,Cr362}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (9.846, 20.282] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 3.297] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{Valve-air} : x_{Valve-air,i} \in [28.604, 28.934] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{Q-air} : x_{Q-air,i} \in [697.829, 739.819] \xrightarrow{1.0} Cr362 , \\ r_{3,Cr362}^{Q-influent} : x_{Q-influent,i} \in (85.092, 85.5] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 42.276] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{TN-influent} : x_{TN-influent,i} \in [0.0, 28.792] \xrightarrow{1.0} Cr362 , \\ r_{3,Cr362}^{TN-effluent} : x_{TN-effluent,i} \in (28.933, 34.867] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{TOC-influent} : x_{TOC-influent,i} \in [0.0, 63.22] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.0, 3.833] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{TOC-effluent} : x_{TOC-effluent,i} \in [0.0, 2.014) \xrightarrow{1.0} Cr362 \end{array} \}$$

5. Los Cuadros 22.26 y 22.27 muestran la cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$ y $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr361}^{NH4-influent}$	5	2,91%
$r_{3,Cr361}^{O2-1aerobic}$	1	0,58%
$r_{1,Cr361}^{O2-2aerobic}$	2	1,16%
$r_{3,Cr361}^{Valve-air}$	28	16,28%
$r_{3,Cr361}^{Q-air}$	27	15,70%
$r_{3,Cr361}^{h-ww}$	16	9,30%
$r_{1,Cr361}^{Q-influent}$	6	3,49%
$r_{3,Cr361}^{Freq-rec}$	3	1,74%
$r_{3,Cr361}^{TN-influent}$	9	5,23%
$r_{3,Cr361}^{Temp-ww}$	5	2,91%
$r_{3,Cr361}^{TOC-influent}$	20	11,63%
$r_{3,Cr361}^{TOC-effluent}$	10	5,81%

Tabla 22.5: Cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr362}^{NH4-influent}$	3	1,55%
$r_{3,Cr362}^{NH4-2aerobic}$	38	19,69%
$r_{1,Cr362}^{O2-1aerobic}$	4	2,07%
$r_{1,Cr362}^{Valve-air}$	5	2,59%
$r_{1,Cr362}^{Q-air}$	2	1,04%
$r_{3,Cr362}^{Q-influent}$	1	0,52%
$r_{1,Cr362}^{FR1-DOTOK}$	5	2,59%
$r_{1,Cr362}^{TN-influent}$	44	22,80%
$r_{3,Cr362}^{TN-effluent}$	14	7,25%
$r_{1,Cr362}^{TOC-influent}$	20	10,36%
$r_{1,Cr362}^{Nitritox-influent}$	4	2,07%
$r_{1,Cr362}^{TOC-effluent}$	1	0,52%

Tabla 22.6: Cobertura relativa de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$.

La regla con mayor cobertura relativa y $p_{sc} = 1$ de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$ es $r_{1,Cr362}^{Valve-air}$ con una $CovR(r)=16,28\%$ y la de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$ es $r_{1,Cr362}^{TN-influent}$ con una $CovR(r)=22,80\%$:

$$r_{3,Cr361}^{Valve-air} : x_{Valve-air,i} \in (54.777, 69.898] \xrightarrow{1.0} Cr361$$

$$r_{1,Cr362}^{TN-influent} : x_{TN-influent,i} \in [0.0, 28.792) \xrightarrow{1.0} Cr362$$

En este caso no hay empate, ver ecuaciones (10.9) y (10.10), y por tanto:

- $A_{Cr361}^2 = A_{Cr361}^{2,Valve-air} = "x_{Valve-air,i} \in (54.777, 69.898]"$
- $A_{Cr362}^2 = A_{Cr362}^{2,TN-influent} = "x_{TN-influent,i} \in [0.0, 28.792)"$

6. Asociando una regla a cada clase

$$\begin{aligned}\mathbb{R}(\mathcal{P}_2) = \{r_{Cr361} : x_{Valve-air,i} \in (54.777, 69.898] \xrightarrow{1.0} Cr361, \\ r_{Cr362} : x_{TN-influent,i} \in [0.0, 28.792) \xrightarrow{1.0} Cr362 \\ \}\end{aligned}$$

7. $\xi = 3$: Así, $\mathcal{P}3_{Lj3,R2}^{EnW,G} = \{Cr362, Cr353, Cr357\}$. La clase $Cr361$ es la que se divide en 2 hijos y la clase $Cr362$ es la que ya estaba en la partición anterior:

$$Cr361 \left\{ \begin{array}{l} Cr353 \\ Cr357 \end{array} \right.$$

Así $C_i^3 = Cr353$; $C_j^3 = Cr357$; $C_t^2 = Cr361$.

8. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_3^* = \{Cr353, Cr357\}$, donde $\mathcal{P}_3^* \subseteq \mathcal{P}3_{Lj3,R2}^{EnW,G}$, se presenta en el Apéndice H.3.
9. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_3^*)$, $k \in 1 : K$.
10. Como resultado de los pasos anteriores, se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_3^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_3^*)$ que se se presenta en el Apéndice H.4 y se trabaja con $\mathcal{S}_{Cr357}(\mathcal{P}_3^*) \subseteq \mathcal{S}(\mathcal{P}_3^*)$, donde $\mathcal{S}(\mathcal{P}_3^*) \subseteq \mathcal{R}(\mathcal{P}_3^*)$:

$$\begin{aligned}\mathcal{S}_{Cr357}(\mathcal{P}_3^*) = \{ & r_{3,Cr357}^{NH4-influent} : x_{NH4-influent,i} \in (42.511, 48.477] \xrightarrow{1.0} Cr357, \\ & r_{1,Cr357}^{Q-influent} : x_{Q-influent,i} \in [49.706, 51.123) \xrightarrow{1.0} Cr357, \\ & r_{1,Cr357}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [42.276, 44.167) \xrightarrow{1.0} Cr357, \\ & r_{3,Cr357}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (38.333, 53.0] \xrightarrow{1.0} Cr357 \}\end{aligned}$$

y $\mathcal{S}_{Cr353}(\mathcal{P}_3^*) \subseteq \mathcal{S}(\mathcal{P}_3^*)$

$$\begin{aligned}\mathcal{S}_{Cr353}(\mathcal{P}_3^*) = \{ & r_{1,Cr353}^{NH4-influent} : x_{NH4-influent,i} \in [7.972, 23.23) \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (8.262, 9.846] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (7.018, 8.785] \xrightarrow{1.0} Cr353, \\ & r_{1,Cr353}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 5.675) \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{Valve-air} : x_{Valve-air,i} \in (57.442, 69.898] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{Q-air} : x_{Q-air,i} \in (2062.554, 2201.27] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{h-ww} : x_{h-ww,i} \in (3.055, 3.098] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{Q-influent} : x_{Q-influent,i} \in (55.666, 85.092] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (47.265, 50.7] \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{Freq-rec} : x_{Freq-rec,i} \in (40.633, 44.0] \xrightarrow{1.0} Cr353, \\ & r_{1,Cr353}^{TN-influent} : x_{TN-influent,i} \in [28.792, 40.417) \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{TN-effluent} : x_{TN-effluent,i} \in (17.837, 28.933] \xrightarrow{1.0} Cr353, \\ & r_{1,Cr353}^{Temp-ww} : x_{Temp-ww,i} \in [8.217, 11.235) \xrightarrow{1.0} Cr353, \\ & r_{1,Cr353}^{TOC-influent} : x_{TOC-influent,i} \in [63.22, 89.833) \xrightarrow{1.0} Cr353, \\ & r_{1,Cr353}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [3.833, 4.875) \xrightarrow{1.0} Cr353, \\ & r_{3,Cr353}^{TOC-effluent} : x_{TOC-effluent,i} \in (36.476, 52.57] \xrightarrow{1.0} Cr353 \}\end{aligned}$$

11. Los Cuadros 22.28 y 22.29 muestran la coberturas relativas de las reglas de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$ y $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)(1)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr357}^{NH4-influent}$	1	2,00%
$r_{1,Cr357}^{Q-influent}$	6	12,00%
$r_{1,Cr357}^{FR1-DOTOK}$	8	16,00%
$r_{3,Cr357}^{Nitritox-influent}$	1	2,00%

Tabla 22.7: Cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr353}^{NH4-influent}$	11	9,02%
$r_{3,Cr353}^{NH4-2aerobic}$	9	7,38%
$r_{3,Cr353}^{O2-1aerobic}$	9	7,38%
$r_{1,Cr353}^{O2-2aerobic}$	12	9,84%
$r_{3,Cr353}^{Valve-air}$	11	9,02%
$r_{3,Cr353}^{Q-air}$	22	18,03%
$r_{3,Cr353}^{h-ww}$	20	16,39%
$r_{3,Cr353}^{Q-influent}$	100	81,97%
$r_{3,Cr353}^{FR1-DOTOK}$	98	80,33%
$r_{3,Cr353}^{Freq-rec}$	49	40,16%
$r_{1,Cr353}^{TN-influent}$	16	13,11%
$r_{3,Cr353}^{TN-effluent}$	81	66,39%
$r_{1,Cr353}^{Temp-ww}$	11	9,02%
$r_{1,Cr353}^{TOC-influent}$	13	10,66%
$r_{1,Cr353}^{Nitritox-influent}$	2	1,64%
$r_{3,Cr353}^{TOC-effluent}$	17	13,93%

Tabla 22.8: Cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$.

La regla con mayor cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$ es $r_{3,Cr353}^{Q-influent}$, con una cobertura relativa $CovR = 81,97\%$ y la regla con mayor cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$ es $r_{1,Cr357}^{FR1-DOTOK}$, con una cobertura relativa $CovR = 16\%$.

$$r_{3,Cr353}^{Q-influent} : x_{Q-influent,i} \in (55.666, 85.092] \xrightarrow{1.0} Cr353$$

$$r_{1,Cr357}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [42.276, 44.167) \xrightarrow{1.0} Cr357$$

En este caso no hay empate, ver ecuaciones (10.9) y (10.10), y por tanto:

- $A_{Cr353}^{*3} = A_{Cr353}^{3,Q-influent} = "x_{Q-influent,i} \in (55.666, 85.092]"$
- $A_{Cr357}^{*3} = A_{Cr357}^{3,FR1-DOTOK} = "x_{FR1-DOTOK,i} \in [42.276, 44.167)"$

12. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

- $A_{Cr362}^3 = A_{Cr362}^2$
 $A_{Cr362}^3 = "x_{TN-influent,i} \in [0.0, 28.792]"$
- $A_{Cr353}^3 = A_{Cr361}^2 \wedge A_{Cr353}^{*3}$
 $A_{Cr353}^3 = "x_{Valve-air,i} \in (54.777, 69.898]" \wedge "x_{Q-influent,i} \in (55.666, 85.092]"$

- $A_{Cr357}^3 = A_{Cr361}^2 \wedge {}_{Cr357}^{*3}$
 $A_{Cr357}^3 = "x_{Valve-air,i} \in (54.777, 69.898]" \wedge "x_{FR1-DOTOK,i} \in [42.276, 44.167]"$

Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$p_{sCr353} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{27}{27} = 1$$

$$p_{sCr357} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{1}{1} = 1$$

$$\mathbb{R}(\mathcal{P}_3) = \{ \quad r_{Cr362} : \quad x_{TN-influent,i} \in [0.0, 28.792] \quad \xrightarrow{1.0} Cr362,$$

$$\begin{aligned} r_{Cr353} : \quad &x_{Valve-air,i} \in (54.777, 69.898] \wedge \\ &x_{Q-influent,i} \in (55.666, 85.092] \quad \xrightarrow{1.0} Cr353, \end{aligned}$$

$$\begin{aligned} r_{Cr357} : \quad &x_{Valve-air,i} \in (54.777, 69.898] \wedge \\ &x_{FR1-DOTOK,i} \in [42.276, 44.167) \quad \xrightarrow{1.0} Cr357 \quad \} \end{aligned}$$

13. $\xi = 4$: Así, $\mathcal{P}_4^{EnW,G}_{Lj3,R2} = \{Cr358, Cr360, Cr353, Cr357\}$. La clase $Cr362$ de $\mathcal{P}_3^{EnW,G}_{Lj3,R2}$ tiene 2 hijos:

$$Cr362 \left\{ \begin{array}{l} Cr358 \\ Cr360 \end{array} \right.$$

Así $C_i^4 = Cr358$; $C_j^4 = Cr360$; $C_t^3 = Cr362$

14. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_4^* = \{Cr358, Cr360\}$, donde $\mathcal{P}_4^* \subseteq \mathcal{P}_4^{EnW,G}_{Lj3,R2}$, se presenta en el Apéndice H.5.
15. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_4^*)$, $k \in 1 : K$.
16. Como resultado de los pasos anteriores, se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_4^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_4^*)$ que se presenta en el Apéndice H.6 y se trabaja con los sistemas de reglas seguros; $\mathcal{S}_{Cr360}(\mathcal{P}_4^*) \subseteq \mathcal{S}(\mathcal{P}_4^*)$, donde $\mathcal{S}(\mathcal{P}_4^*) \subseteq \mathcal{R}(\mathcal{P}_4^*)$ el cual es:

$$\begin{aligned} \mathcal{S}_{Cr360}(\mathcal{P}_4^*) = \{ \quad &r_{1,Cr360}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.79] \xrightarrow{1.0} Cr360, \\ &r_{1,Cr360}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 4.479] \xrightarrow{1.0} Cr360, \\ &r_{1,Cr360}^{Valve-air} : x_{Valve-air,i} \in [28.604, 29.57] \xrightarrow{1.0} Cr360, \\ &r_{3,Cr360}^{h-ww} : x_{h-ww,i} \in (3.039, 3.058] \xrightarrow{1.0} Cr360, \\ &r_{3,Cr360}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (50.692, 50.733] \xrightarrow{1.0} Cr360, \\ &r_{3,Cr360}^{Temp-ww} : x_{Temp-ww,i} \in (20.928, 21.896] \xrightarrow{1.0} Cr360, \\ &r_{3,Cr360}^{TOC-influent} : x_{TOC-influent,i} \in (225.293, 290.212] \xrightarrow{1.0} Cr360, \\ &r_{3,Cr360}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (30.0, 53.708] \xrightarrow{1.0} Cr360, \\ &r_{3,Cr360}^{TOC-effluent} : x_{TOC-effluent,i} \in (42.251, 44.053] \xrightarrow{1.0} Cr360 \quad \} \end{aligned}$$

y $\mathcal{S}_{Cr358}(\mathcal{P}_4^*) \subseteq \mathcal{S}(\mathcal{P}_4^*)$ es:

$$\mathcal{S}_{Cr358}(\mathcal{P}_4^*) = \{ \begin{array}{ll} r_{3,Cr358}^{NH4-influent} : x_{NH4-influent,i} \in (34.353, 40.541] \xrightarrow{1.0} Cr358 , \\ r_{3,Cr358}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (2.856, 20.282] \xrightarrow{1.0} Cr358 , \\ r_{3,Cr358}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (6.998, 8.371] \xrightarrow{1.0} Cr358 , \\ r_{1,Cr358}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [4.94, 5.889] \xrightarrow{1.0} Cr358 , \\ r_{3,Cr358}^{Valve-air} : x_{Valve-air,i} \in (51.168, 54.777] \xrightarrow{1.0} Cr358 , \\ r_{3,Cr358}^{Q-air} : x_{Q-air,i} \in (1770.852, 2030.77] \xrightarrow{1.0} Cr358 , \\ r_{1,Cr358}^{h-ww} : x_{h-ww,i} \in [2.978, 2.984] \xrightarrow{1.0} Cr358 , \\ r_{1,Cr358}^{Q-influent} : x_{Q-influent,i} \in [50.99, 63.038] \xrightarrow{1.0} Cr358 , \\ r_{1,Cr358}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 47.2] \xrightarrow{1.0} Cr358 , \\ r_{1,Cr358}^{Freq-rec} : x_{Freq-rec,i} \in [23.863, 24.899] \xrightarrow{1.0} Cr358 , \\ r_{3,Cr358}^{TN-influent} : x_{TN-influent,i} \in (54.792, 65.25] \xrightarrow{1.0} Cr358 , \\ r_{3,Cr358}^{TN-effluent} : x_{TN-effluent,i} \in (17.788, 34.867] \xrightarrow{1.0} Cr358 , \\ r_{1,Cr358}^{Temp-ww} : x_{Temp-ww,i} \in [8.472, 13.327] \xrightarrow{1.0} Cr358 , \\ r_{1,Cr358}^{TOC-influent} : x_{TOC-influent,i} \in [0.0, 38.888] \xrightarrow{1.0} Cr358 , \\ r_{1,Cr358}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.0, 0.542] \xrightarrow{1.0} Cr358 , \\ r_{1,Cr358}^{TOC-effluent} : x_{TOC-effluent,i} \in [0.0, 11.879] \xrightarrow{1.0} Cr358 \end{array} \}$$

17. Los Cuadros 22.30 y 22.31 muestran la cobertura relativa de las reglas de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$ y $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr360}^{O2-1aerobic}$	28	28,00%
$r_{1,Cr360}^{Valve-air}$	5	5,00%
$r_{3,Cr360}^{h-ww}$	9	9,00%
$r_{3,Cr360}^{FR1-DOTOK}$	3	3,00%
$r_{3,Cr360}^{Temp-ww}$	29	29,00%
$r_{3,Cr360}^{TOC-influent}$	1	1,00%
$r_{3,Cr360}^{Nitritox-influent}$	3	3,00%
$r_{3,Cr360}^{TOC-effluent}$	1	1,00%

Tabla 22.9: Cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$.

La regla con mayor cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$ es $r_{1,Cr358}^{Temp-ww}$, con una cobertura relativa $CovR = 66,67\%$ y la regla con mayor cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$ es $r_{3,Cr358}^{Temp-ww}$, con una cobertura relativa $CovR = 29\%$.

$$\begin{aligned} r_{1,Cr358}^{Temp-ww} : x_{Temp-ww,i} &\in [8.472, 13.327] \xrightarrow{1.0} Cr358 \\ r_{3,Cr360}^{Temp-ww} : x_{Temp-ww,i} &\in (20.928, 21.896] \xrightarrow{1.0} Cr360 \end{aligned}$$

En este caso no hay empate, ver ecuaciones (10.9) y (10.10), y por tanto:

- $A_{Cr358}^{*4} = A_{Cr358}^{4,Temp-ww} = "x_{Temp-ww,i} \in [8.472, 13.327]"$
- $A_{Cr360}^{*4} = A_{Cr360}^{4,Temp-ww} = "x_{Temp-ww,i} \in (20.928, 21.896]"$

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr358}^{NH4-2aerobic}$	54	58,06%
$r_{3,Cr358}^{O2-1aerobic}$	54	58,06%
$r_{1,Cr358}^{O2-2aerobic}$	5	5,38%
$r_{3,Cr358}^{Valve-air}$	3	3,23%
$r_{3,Cr358}^{Q-air}$	9	9,68%
$r_{1,Cr358}^{h-ww}$	6	6,45%
$r_{1,Cr358}^{Q-influent}$	33	35,48%
$r_{1,Cr358}^{FR1-DOTOK}$	34	36,56%
$r_{1,Cr358}^{Freq-rec}$	13	13,98%
$r_{3,Cr358}^{TN-influent}$	19	20,43%
$r_{3,Cr358}^{TN-effluent}$	38	40,86%
$r_{1,Cr358}^{Temp-ww}$	62	66,67%
$r_{1,Cr358}^{TOC-influent}$	1	1,08%
$r_{1,Cr358}^{Nitritox-influent}$	1	1,08%
$r_{1,Cr358}^{TOC-effluent}$	6	6,45%

Tabla 22.10: Cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$.

18. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

- $A_{Cr353}^4 = A_{Cr353}^3 = A_{Cr361}^2 \wedge A_{Cr353}^{*3}$
- $A_{Cr353}^4 = "x_{Valve-air,i} \in (54.777, 69.898]" \wedge "x_{Q-influent,i} \in (55.666, 85.092]"$
- $A_{Cr357}^4 = A_{Cr357}^3 = A_{Cr361}^2 \wedge A_{Cr357}^{*3}$
- $A_{Cr357}^4 = "x_{Valve-air,i} \in (54.777, 69.898]" \wedge "x_{FR1-DOTOK,i} \in [42.276, 44.167]"$
- $A_{Cr358}^4 = A_{Cr362}^2 \wedge A_{Cr358}^{*4}$
- $A_{Cr358}^4 = "x_{TN-influent,i} \in [0.0, 28.792]" \wedge "x_{Temp-ww,i} \in [8.472, 13.327]"$
- $A_{Cr360}^4 = A_{Cr362}^2 \wedge A_{Cr360}^{*4}$
- $A_{Cr360}^4 = "x_{TN-influent,i} \in [0.0, 28.792]" \wedge "x_{Temp-ww,i} \in (20.928, 21.896]"$

Así pues asociando una regla compuesta a cada clase

$$\begin{aligned} \mathbb{R}(\mathcal{P}_4) = \{ & \quad r_{Cr353} : \quad x_{Valve-air,i} \in (54.777, 69.898] \wedge \\ & \quad x_{Q-influent,i} \in (55.666, 85.092] \xrightarrow{1.0} Cr353, \\ & \quad r_{Cr357} : \quad x_{Valve-air,i} \in (54.777, 69.898] \wedge \\ & \quad x_{FR1-DOTOK,i} \in [42.276, 44.167) \xrightarrow{1.0} Cr357, \\ & \quad r_{Cr358} : \quad x_{TN-influent,i} \in [0.0, 28.792) \wedge \\ & \quad x_{Temp-ww,i} \in [8.472, 13.327) \xrightarrow{1.0} Cr358, \\ & \quad r_{Cr360} : \quad x_{TN-influent,i} \in [0.0, 28.792) \wedge \\ & \quad x_{Temp-ww,i} \in (20.928, 21.896] \xrightarrow{1.0} Cr360 \\ & \} \end{aligned}$$

22.2.1 Interpretación final:

Que correspondería a la conceptualización:

- Cr353: “*Valve – air* es alto y *Q – influent* es alto”
- Cr357: “*Valve – air* es alto y *FR1 – DOTOK* es bajo”
- Cr358: “*TN – influent* es bajo y *Temp – ww* es bajo”
- Cr360: “*TN – influent* es bajo y *Temp – ww* es alto”

22.2.2 Evaluación de la propuesta

Una vez determinados los conceptos asociados a cada clase se ha procedido a evaluar cual(es) era(n) satisfecho(s) por cada objeto de \mathcal{I} con tal de evaluar la correcta correspondencia entre la muestra y la conceptualización inducida. La Tabla 22.11 muestra los resultados de calcular el soporte $Sup(r)$, ver ecuación (9.1), la cobertura relativa $CovR(r)$, ver ecuación (9.5) y la confianza $p(r)$, ver ecuación (9.2) de cada una de las reglas compuestas inducidas para cada clase de la partición final.

La evaluación de la regla se hace con respecto al total de objetos de la base de datos, es evidente que no habrá inconsistencias al evaluar cada regla por separado ya que los intervalos que conforman cada regla compuesta, son disjuntos entre una clase y otro debido a la forma en que se construyen los conceptos, ver ecuaciones (10.9) y (10.10), caracterizando los extremos de las clases.

Si se observa la Tabla 22.11, se tiene que el soporte se obtiene dividiendo cada celda de la tercera columna entre el total de objetos de la base de datos (365), la cobertura relativa se obtiene dividiendo cada celda de la cuarta columna entre la correspondiente celda de la quinta columna y la confianza dividiendo las celdas de la cuarta entre las correspondientes celdas de la tercera columna y, en este caso, al elegir la la regla con mayor cobertura relativa por clase y con $p_{sc} = 1$, lo que se hace, entonces, es caracterizar los extremos de las clases.

Como en todas las clases se cumple que $\#\{i \in A_C^\xi\} = \#\{i \in A_C^\xi \cap i \in C\}$, lo cual era ya predecible dada la forma en que se construyen los conceptos, se puede concluir que no hay objetos mal asignados, es decir, todos los objetos satisfacen el antecedente y el consecuente de la reglas. Este valor no se considera en el soporte ni en la cobertura relativa, pero queda de manifiesto en la confianza ya que cuando se obtienen confianzas del 100% significa que todos

los objetos que satisfacen el antecedente de la regla, también satisfacen simultáneamente el antecedente y el consecuente de la regla y por lo tanto pertenecen a la clase que la regla asigna. Con lo cual se puede concluir que el número de objetos correctamente asignados por clase, en este caso, viene dado por $\#\{i \in A_C^\xi \cap i \in C\}$ o por $\#\{i \in A_C^\xi\}$ y que el porcentaje total de objetos correctamente asignados se puede obtener dividiendo 37 entre 365, 10,14%.

Ruler	Consec.	$\#\{i \in A_C^\xi\}$	$\#\{i \in A_C^\xi \cap i \in C\}$	n_c	$p(r)$	$Sup(r)$	$CovR(r)$
r_{Cr353}	$Cr353$	27	27	122	100%	7,40%	22,13%
r_{Cr357}	$Cr357$	1	1	50	100%	0,27%	2,00%
r_{Cr358}	$Cr358$	6	6	93	100%	1,64%	6,45%
r_{Cr360}	$Cr360$	3	3	100	100%	0,82%	3,00%
<i>Media</i>					100%		8,40%
<i>Suma</i>		37	37	365		10,14%	
<i>CovGlobal(\mathbb{R})</i>							10,1%

Tabla 22.11: Evaluación: Best local concept and not Close-World Assumption.

También es interesante observar que el número de objetos asignados por las reglas compuestas asociadas a cada clase se puede obtener a partir de $\#\{i \in A_C^\xi\}$ y el porcentaje correspondiente al total de objetos de la base de datos viene cuantificado por el soporte. Como la suma total de los soportes por clase corresponden al porcentaje de objetos asignados, cuando este es menor que el 100% significa que hay objetos sin asignar. En la Tabla 22.11 se puede observar que el soporte es muy bajo, lo cual implica que el porcentaje de objetos sin asignar es muy alto. Al restar a 365 la suma de $\#\{i \in A_C^\xi\}$ hay 328 objetos que no han sido asignados, esta conclusión se puede realizar sólo en el caso que no existan inconsistencias y para soportes menores o iguales al 100%.

Hay una gran diferencia porcentual entre la cobertura relativa y confianza y podemos observar que con confianzas del 100% aparecen soportes muy bajos. Si se considera la propuesta la propuesta Best global concept and Close-World Assumption, analizada en la sección anterior, se puede concluir que al ganar confianza se pierde soporte y cobertura relativa.

22.3 Interpretación de \mathcal{P}_4 utilizando Best local concept and Close-World Assumption:

1. $\xi = 2$: Así, $\mathcal{P}^{EnW,G}_{Lj3,R2} = \{Cr361, Cr362\}$. La raíz del árbol $Cr363$ tiene 2 hijos:

$$Cr363 \left\{ \begin{array}{l} Cr361 \\ Cr362. \end{array} \right.$$

- 2. La discretización realizada con el BbD se presenta en el Apéndice H.1
- 3. El sistema de reglas inducido para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(\mathcal{P}_2^*) = \bigcup_{k=1}^k \mathcal{R}(X_k, \mathcal{P}_2^*)$, se presenta en el Apéndice H.2
- 4. Como resultado de los pasos anteriores, se consideran los siguientes sistemas de reglas con reglas seguras $\mathcal{S}_{Cr361}(\mathcal{P}_2^*) \subseteq \mathcal{S}(\mathcal{P}_2^*)$:

$$\mathcal{S}_{Cr361}(\mathcal{P}_2^*) = \{ \begin{array}{ll} r_{3,Cr361}^{NH4-influent} : x_{NH4-influent,i} \in (40.541, 48.477] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (8.371, 8.785] & \xrightarrow{1.0} Cr361, \\ r_{1,Cr361}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 4.94) & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{Valve-air} : x_{Valve-air,i} \in (54.777, 69.898] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{Q-air} : x_{Q-air,i} \in (2030.77, 2201.27] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{h-ww} : x_{h-ww,i} \in (3.058, 3.098] & \xrightarrow{1.0} Cr361, \\ r_{1,Cr361}^{Q-influent} : x_{Q-influent,i} \in [49.706, 50.99) & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{Freq-rec} : x_{Freq-rec,i} \in (43.97, 44.0] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{TN-influent} : x_{TN-influent,i} \in (65.25, 83.792] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{Temp-ww} : x_{Temp-ww,i} \in (21.896, 22.583] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{TOC-influent} : x_{TOC-influent,i} \in (290.212, 355.0] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{TOC-effluent} : x_{TOC-effluent,i} \in (44.053, 52.57] & \xrightarrow{1.0} Cr361 \end{array} \}$$

y $\mathcal{S}_{Cr362}(\mathcal{P}_2^*) \subseteq \mathcal{S}(\mathcal{P}_2^*)$.

$$\mathcal{S}_{Cr362}(\mathcal{P}_2^*) = \{ \begin{array}{ll} r_{1,Cr362}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.972] & \xrightarrow{1.0} Cr362, \\ r_{3,Cr362}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (9.846, 20.282] & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 3.297) & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{Valve-air} : x_{Valve-air,i} \in [28.604, 28.934) & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{Q-air} : x_{Q-air,i} \in [697.829, 739.819) & \xrightarrow{1.0} Cr362, \\ r_{3,Cr362}^{Q-influent} : x_{Q-influent,i} \in (85.092, 85.5] & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 42.276) & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{TN-influent} : x_{TN-influent,i} \in [0.0, 28.792] & \xrightarrow{1.0} Cr362, \\ r_{3,Cr362}^{TN-effluent} : x_{TN-effluent,i} \in (28.933, 34.867] & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{TOC-influent} : x_{TOC-influent,i} \in [0.0, 63.22) & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.0, 3.833) & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{TOC-effluent} : x_{TOC-effluent,i} \in [0.0, 2.014) & \xrightarrow{1.0} Cr362 \end{array} \}$$

5. Los Cuadros 22.26 y 22.27 muestran la cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$ y $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr361}^{NH4-influent}$	5	2,91%
$r_{3,Cr361}^{O2-1aerobic}$	1	0,58%
$r_{1,Cr361}^{O2-2aerobic}$	2	1,16%
$r_{3,Cr361}^{Valve-air}$	28	16,28%
$r_{3,Cr361}^{Q-air}$	27	15,70%
$r_{3,Cr361}^{h-ww}$	16	9,30%
$r_{1,Cr361}^{Q-influent}$	6	3,49%
$r_{3,Cr361}^{Freq-rec}$	3	1,74%
$r_{3,Cr361}^{TN-influent}$	9	5,23%
$r_{3,Cr361}^{Temp-ww}$	5	2,91%
$r_{3,Cr361}^{TOC-influent}$	20	11,63%
$r_{3,Cr361}^{TOC-effluent}$	10	5,81%

Tabla 22.12: Cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr362}^{NH4-influent}$	3	1,55%
$r_{3,Cr362}^{NH4-2aerobic}$	38	19,69%
$r_{1,Cr362}^{O2-1aerobic}$	4	2,07%
$r_{1,Cr362}^{Valve-air}$	5	2,59%
$r_{1,Cr362}^{Q-air}$	2	1,04%
$r_{3,Cr362}^{Q-influent}$	1	0,52%
$r_{1,Cr362}^{FR1-DOTOK}$	5	2,59%
$r_{1,Cr362}^{TN-influent}$	44	22,80%
$r_{3,Cr362}^{TN-effluent}$	14	7,25%
$r_{1,Cr362}^{TOC-influent}$	20	10,36%
$r_{1,Cr362}^{Nitritox-influent}$	4	2,07%
$r_{1,Cr362}^{TOC-effluent}$	1	0,52%

Tabla 22.13: Cobertura relativa de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$.

La regla con mayor cobertura relativa y $p_{sc} = 1$ de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$ es $r_{1,Cr362}^{Valve-air}$ con una $CovR(r)=16,28\%$ y la de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$ es $r_{1,Cr362}^{TN-influent}$ con una $CovR(r)=22,80\%$:

$$\begin{aligned} r_{3,Cr361}^{Valve-air} : x_{Valve-air,i} &\in (54.777, 69.898] \xrightarrow{1.0} Cr361 \\ r_{1,Cr362}^{TN-influent} : x_{TN-influent,i} &\in [0.0, 28.792) \xrightarrow{1.0} Cr362 \end{aligned}$$

Así,

$$\begin{aligned} A_{Cr361}^{2,Valve-air} &= "x_{Valve-air,i} \in (54.777, 69.898]" \\ A_{Cr362}^{2,TN-influent} &= "x_{TN-influent,i} \in [0.0, 28.792)" \end{aligned}$$

En este caso no hay empate, ver ecuaciones (10.15) y (10.16), y por tanto:

- $A_{Cr361}^2 = A_{Cr361}^{2,Valve-air} \vee \neg A_{Cr362}^{2,TN-influent}$

$$A_{Cr361}^2 = "x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]"$$

$$\bullet \quad A_{Cr362}^2 = A_{Cr362}^{2,TN-influent} \vee \neg A_{Cr361}^{2,Valve-air}$$

$$A_{Cr362}^2 = "x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]"$$

6. Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$p_{sCr361} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{172}{321} = 0.54$$

$$p_{sCr362} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{181}{334} = 0.54$$

$$\begin{aligned} \mathbb{R}(\mathcal{P}_2) = \{ & \quad r_{Cr361} : \quad x_{Valve-air,i} \in (54.777, 69.898] \vee \\ & \quad x_{TN-influent,i} \in [28.792, 83.792] \quad \xrightarrow{0.54} Cr361, \\ & \quad r_{Cr362} : \quad x_{TN-influent,i} \in [0.0, 28.792] \vee \\ & \quad x_{Valve-air,i} \in [28.604, 54.777] \quad \xrightarrow{0.54} Cr362 \} \end{aligned}$$

7. $\xi = 3$: Así, $\mathcal{P}_3^{EnW,G}_{Lj3,R2} = \{Cr362, Cr353, Cr357\}$. La clase $Cr361$ es la que se divide en 2 hijos y la clase $Cr362$ es la que ya estaba en la partición anterior:

$$Cr361 \left\{ \begin{array}{l} Cr353 \\ Cr357 \end{array} \right.$$

Así $C_i^3 = Cr353$; $C_j^3 = Cr357$; $C_t^2 = Cr361$.

8. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_3^* = \{Cr353, Cr357\}$, donde $\mathcal{P}_3^* \subseteq \mathcal{P}_3^{EnW,G}_{Lj3,R2}$, se presenta en el Apéndice H.3.
9. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_3^*)$, $k \in 1 : K$.
10. Como resultado de los pasos anteriores, se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_3^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_3^*)$ que se se presenta en el Apéndice H.4 y se trabaja con $\mathcal{S}_{Cr357}(\mathcal{P}_3^*) \subseteq \mathcal{S}(\mathcal{P}_3^*)$ el cual es:

$$\begin{aligned} \mathcal{S}_{Cr357}(\mathcal{P}_3^*) = \{ & \quad r_{3,Cr357}^{NH4-influent} : x_{NH4-influent,i} \in (42.511, 48.477] \xrightarrow{1.0} Cr357, \\ & \quad r_{1,Cr357}^{Q-influent} : x_{Q-influent,i} \in [49.706, 51.123] \xrightarrow{1.0} Cr357, \\ & \quad r_{1,Cr357}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [42.276, 44.167] \xrightarrow{1.0} Cr357, \\ & \quad r_{3,Cr357}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (38.333, 53.0] \xrightarrow{1.0} Cr357 \} \end{aligned}$$

y $\mathcal{S}_{Cr353}(\mathcal{P}_3^*) \subseteq \mathcal{S}(\mathcal{P}_3^*)$

$$\mathcal{S}_{Cr353}(\mathcal{P}_3^*) = \{ \begin{array}{ll} r_{1,Cr353}^{NH4-influent} : x_{NH4-influent,i} \in [7.972, 23.23] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (8.262, 9.846] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (7.018, 8.785] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 5.675] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Value-air} : x_{Value-air,i} \in (57.442, 69.898] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Q-air} : x_{Q-air,i} \in (2062.554, 2201.27] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{h-ww} : x_{h-ww,i} \in (3.055, 3.098] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Q-influent} : x_{Q-influent,i} \in (55.666, 85.092] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (47.265, 50.7] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Freq-rec} : x_{Freq-rec,i} \in (40.633, 44.0] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{TN-influent} : x_{TN-influent,i} \in [28.792, 40.417] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{TN-effluent} : x_{TN-effluent,i} \in (17.837, 28.933] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{Temp-ww} : x_{Temp-ww,i} \in [8.217, 11.235] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{TOC-influent} : x_{TOC-influent,i} \in [63.22, 89.833] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [3.833, 4.875] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{TOC-effluent} : x_{TOC-effluent,i} \in (36.476, 52.57] \xrightarrow{1.0} Cr353 \end{array} \}$$

11. Los Cuadros 22.28 y 22.29 muestran la coberturas relativas de las reglas de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$ y $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr357}^{NH4-influent}$	1	2,00%
$r_{1,Cr357}^{Q-influent}$	6	12,00%
$r_{1,Cr357}^{FR1-DOTOK}$	8	16,00%
$r_{3,Cr357}^{Nitritox-influent}$	1	2,00%

Tabla 22.14: Cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr353}^{NH4-influent}$	11	9,02%
$r_{3,Cr353}^{NH4-2aerobic}$	9	7,38%
$r_{3,Cr353}^{O2-1aerobic}$	9	7,38%
$r_{1,Cr353}^{O2-2aerobic}$	12	9,84%
$r_{3,Cr353}^{Value-air}$	11	9,02%
$r_{3,Cr353}^{Q-air}$	22	18,03%
$r_{3,Cr353}^{h-ww}$	20	16,39%
$r_{3,Cr353}^{Q-influent}$	100	81,97%
$r_{3,Cr353}^{FR1-DOTOK}$	98	80,33%
$r_{3,Cr353}^{Freq-rec}$	49	40,16%
$r_{1,Cr353}^{TN-influent}$	16	13,11%
$r_{3,Cr353}^{TN-effluent}$	81	66,39%
$r_{1,Cr353}^{Temp-ww}$	11	9,02%
$r_{1,Cr353}^{TOC-influent}$	13	10,66%
$r_{1,Cr353}^{Nitritox-influent}$	2	1,64%
$r_{3,Cr353}^{TOC-effluent}$	17	13,93%

Tabla 22.15: Cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$.

La regla con mayor cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$ es $r_{3,Cr353}^{Q-influent}$, con una cobertura relativa $CovR = 81,97\%$ y la regla con mayor cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$ es $r_{1,Cr357}^{FR1-DOTOK}$, con una cobertura relativa $CovR = 16\%$.

$$\begin{aligned} r_{3,Cr353}^{Q-influent} : x_{Q-influent,i} &\in (55.666, 85.092] \xrightarrow{1.0} Cr353 \\ r_{1,Cr357}^{FR1-DOTOK} : x_{FR1-DOTOK,i} &\in [42.276, 44.167) \xrightarrow{1.0} Cr357 \end{aligned}$$

Así,

$$A_{Cr353}^{3,Q-influent} = "x_{Q-influent,i} \in (55.666, 85.092]"$$

$$A_{Cr357}^{3,FR1-DOTOK} = "x_{FR1-DOTOK,i} \in [42.276, 44.167)"$$

En este caso no hay empate, ver ecuaciones (10.15) y (10.16), y por tanto:

$$\begin{aligned} \bullet \quad A_{Cr353}^{*3} &= A_{Cr353}^{3,Q-influent} \vee \neg A_{Cr357}^{3,FR1-DOTOK} \\ A_{Cr353}^{*3} &= "x_{Q-influent,i} \in (55.666, 85.092]" \vee "x_{FR1-DOTOK,i} \in [44.167, 50.7]" \\ \bullet \quad A_{Cr357}^{*3} &= A_{Cr357}^{3,FR1-DOTOK} \vee \neg A_{Cr353}^{3,Q-influent} \\ A_{Cr357}^{*3} &= "x_{Q-influent,i} \in [49.706, 55.666]" \vee "x_{FR1-DOTOK,i} \in [42.276, 44.167]" \end{aligned}$$

12. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

$$\begin{aligned} \bullet \quad A_{Cr362}^3 &= A_{Cr362}^2 \\ A_{Cr362}^3 &= "x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]" \\ \bullet \quad A_{Cr353}^3 &= A_{Cr361}^2 \wedge A_{Cr353}^{*3} \\ A_{Cr353}^3 &= ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge \\ &("x_{Q-influent,i} \in (55.666, 85.092]" \vee "x_{FR1-DOTOK,i} \in [44.167, 50.7]") \\ \bullet \quad A_{Cr357}^3 &= A_{Cr361}^2 \wedge A_{Cr357}^{*3} \\ A_{Cr357}^3 &= ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge \\ &("x_{Q-influent,i} \in [49.706, 55.666]" \vee "x_{FR1-DOTOK,i} \in [42.276, 44.167]") \end{aligned}$$

Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$p_{sCr353} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{122}{309} = 0.40$$

$$p_{sCr357} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{48}{101} = 0.48$$

$$\begin{aligned} \mathbb{R}(\mathcal{P}_3) = \{ & r_{Cr362} : \quad x_{TN-influent,i} \in [0.0, 28.792] \vee \\ & x_{Valve-air,i} \in [28.604, 54.777] \xrightarrow{0.54} Cr362 \} \\ r_{Cr353} : & (x_{Valve-air,i} \in (54.777, 69.898] \vee \\ & x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\ & (x_{Q-influent,i} \in (55.666, 85.092] \vee \\ & x_{FR1-DOTOK,i} \in [44.167, 50.7]) \xrightarrow{0.4} Cr353, \\ r_{Cr357} : & (x_{Valve-air,i} \in (54.777, 69.898] \vee \\ & x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\ & (x_{Q-influent,i} \in [49.706, 55.666] \vee \\ & x_{FR1-DOTOK,i} \in [42.276, 44.167]) \xrightarrow{0.48} Cr357 \} \end{aligned}$$

13. $\xi = 4$: Así, $\mathcal{P}_4^{EnW,G}_{Lj3,R2} = \{Cr358, Cr360, Cr353, Cr357\}$. La clase $Cr362$ de $\mathcal{P}_3^{EnW,G}_{Lj3,R2}$ tiene 2 hijos:

$$Cr362 \left\{ \begin{array}{l} Cr358 \\ Cr360 \end{array} \right.$$

Así $C_i^4 = Cr358$; $C_j^4 = Cr360$; $C_t^3 = Cr362$

14. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_4^* = \{Cr358, Cr360\}$, donde $\mathcal{P}_4^* \subseteq \mathcal{P}_4^{EnW,G}_{Lj3,R2}$, se presenta en el Apéndice H.5.
15. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_4^*)$, $k \in 1 : K$.
16. Como resultado de los pasos anteriores, se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_4^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_4^*)$ que se presenta en el Apéndice H.6 y se trabaja con $\mathcal{S}_{Cr360}(\mathcal{P}_4^*) \subseteq \mathcal{S}(\mathcal{P}_4^*)$ el cual es:

$$\begin{aligned} \mathcal{S}_{Cr360}(\mathcal{P}_4^*) = \{ & r_{1,Cr360}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.79] \xrightarrow{1.0} Cr360, \\ & r_{1,Cr360}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 4.479] \xrightarrow{1.0} Cr360, \\ & r_{1,Cr360}^{Value-air} : x_{Valve-air,i} \in [28.604, 29.57] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{h-ww} : x_{h-ww,i} \in (3.039, 3.058] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (50.692, 50.733] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{Temp-ww} : x_{Temp-ww,i} \in (20.928, 21.896] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{TOC-influent} : x_{TOC-influent,i} \in (225.293, 290.212] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (30.0, 53.708] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{TOC-effluent} : x_{TOC-effluent,i} \in (42.251, 44.053] \xrightarrow{1.0} Cr360 \} \end{aligned}$$

y $\mathcal{S}_{Cr358}(\mathcal{P}_4^*) \subseteq \mathcal{S}(\mathcal{P}_4^*)$ es:

$$\begin{aligned} \mathcal{S}_{Cr358}(\mathcal{P}_4^*) = \{ & r_{3,Cr358}^{NH4-influent} : x_{NH4-influent,i} \in (34.353, 40.541] \xrightarrow{1.0} Cr358, \\ & r_{3,Cr358}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (2.856, 20.282] \xrightarrow{1.0} Cr358, \\ & r_{3,Cr358}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (6.998, 8.371] \xrightarrow{1.0} Cr358, \end{aligned}$$

$$\begin{aligned}
r_{1,Cr358}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [4.94, 5.889] &\xrightarrow{1.0} Cr358 , \\
r_{3,Cr358}^{Valve-air} : x_{Valve-air,i} \in (51.168, 54.777] &\xrightarrow{1.0} Cr358 , \\
r_{3,Cr358}^{Q-air} : x_{Q-air,i} \in (1770.852, 2030.77] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{h-ww} : x_{h-ww,i} \in [2.978, 2.984] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Q-influent} : x_{Q-influent,i} \in [50.99, 63.038] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 47.2) &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Freq-rec} : x_{Freq-rec,i} \in [23.863, 24.899) &\xrightarrow{1.0} Cr358 , \\
r_{3,Cr358}^{TN-influent} : x_{TN-influent,i} \in (54.792, 65.25] &\xrightarrow{1.0} Cr358 , \\
r_{3,Cr358}^{TN-effluent} : x_{TN-effluent,i} \in (17.788, 34.867] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Temp-ww} : x_{Temp-ww,i} \in [8.472, 13.327) &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{TOC-influent} : x_{TOC-influent,i} \in [0.0, 38.888) &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.0, 0.542) &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{TOC-effluent} : x_{TOC-effluent,i} \in [0.0, 11.879) &\xrightarrow{1.0} Cr358 \quad \}
\end{aligned}$$

17. Los Cuadros 22.30 y 22.31 muestran la cobertura relativa de las reglas de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$ y $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr360}^{O2-1aerobic}$	28	28,00%
$r_{1,Cr360}^{Valve-air}$	5	5,00%
$r_{3,Cr360}^{h-ww}$	9	9,00%
$r_{3,Cr360}^{FR1-DOTOK}$	3	3,00%
$r_{3,Cr360}^{Temp-ww}$	29	29,00%
$r_{3,Cr360}^{TOC-influent}$	1	1,00%
$r_{3,Cr360}^{Nitritox-influent}$	3	3,00%
$r_{3,Cr360}^{TOC-effluent}$	1	1,00%

Tabla 22.16: Cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr358}^{NH4-2aerobic}$	54	58,06%
$r_{3,Cr358}^{O2-1aerobic}$	54	58,06%
$r_{1,Cr358}^{O2-2aerobic}$	5	5,38%
$r_{3,Cr358}^{Valve-air}$	3	3,23%
$r_{3,Cr358}^{Q-air}$	9	9,68%
$r_{1,Cr358}^{h-ww}$	6	6,45%
$r_{1,Cr358}^{Q-influent}$	33	35,48%
$r_{1,Cr358}^{FR1-DOTOK}$	34	36,56%
$r_{1,Cr358}^{Freq-rec}$	13	13,98%
$r_{3,Cr358}^{TN-influent}$	19	20,43%
$r_{3,Cr358}^{TN-effluent}$	38	40,86%
$r_{1,Cr358}^{Temp-ww}$	62	66,67%
$r_{1,Cr358}^{TOC-influent}$	1	1,08%
$r_{1,Cr358}^{Nitritox-influent}$	1	1,08%
$r_{1,Cr358}^{TOC-effluent}$	6	6,45%

Tabla 22.17: Cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$.

La regla con mayor cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$ es $r_{1,Cr358}^{Temp-ww}$, con una cobertura relativa $CovR = 66,67\%$ y la regla con mayor cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$ es $r_{3,Cr358}^{Temp-ww}$, con una cobertura relativa $CovR = 29\%$.

$$\begin{aligned} r_{1,Cr358}^{Temp-ww} : x_{Temp-ww,i} \in [8.472, 13.327] &\xrightarrow{1.0} Cr358 \\ r_{3,Cr360}^{Temp-ww} : x_{Temp-ww,i} \in (20.928, 21.896] &\xrightarrow{1.0} Cr360 \end{aligned}$$

Así,

$$\begin{aligned} A_{Cr358}^{4,Temp-ww} &= "x_{Temp-ww,i} \in [8.472, 13.327]" \\ A_{Cr360}^{4,Temp-ww} &= "x_{Temp-ww,i} \in (20.928, 21.896]" \end{aligned}$$

En este caso no hay empate, ver ecuaciones (10.15) y (10.16), y por tanto:

- $A_{Cr358}^{*4} = A_{Cr358}^{4,Temp-ww} \vee \neg A_{Cr360}^{4,Temp-ww}$
- $$A_{Cr358}^{*4} = "x_{Temp-ww,i} \in [8.472, 13.327]" \vee "x_{Temp-ww,i} \in [8.472, 20.928]"$$

- $A_{Cr360}^{*4} = A_{Cr360}^{4,Temp-ww} \vee \neg A_{Cr358}^{4,Temp-ww}$
- $$A_{Cr360}^{*4} = "x_{Temp-ww,i} \in (20.928, 21.896]" \vee "x_{Temp-ww,i} \in [13.327, 21.896]"$$

18. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

- $A_{Cr353}^4 = A_{Cr353}^3 = A_{Cr361}^2 \wedge A_{Cr353}^{*3}$
- $$A_{Cr353}^4 = ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge ("x_{Q-influent,i} \in (55.666, 85.092]" \vee "x_{FR1-DOTOK,i} \in [44.167, 50.7]")$$
- $A_{Cr357}^4 = A_{Cr357}^3 = A_{Cr361}^2 \wedge A_{Cr357}^{*3}$
- $$A_{Cr357}^4 = ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge ("x_{Q-influent,i} \in [49.706, 55.666]" \vee "x_{FR1-DOTOK,i} \in [42.276, 44.167]")$$
- $A_{Cr358}^4 = A_{Cr362}^2 \wedge A_{Cr358}^{*4}$
- $$A_{Cr358}^4 = ("x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]") \wedge ("x_{Temp-ww,i} \in [8.472, 13.327]" \vee "x_{Temp-ww,i} \in [8.472, 20.928]")$$
- $A_{Cr360}^4 = A_{Cr362}^2 \wedge A_{Cr360}^{*4}$
- $$A_{Cr360}^4 = ("x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]") \wedge ("x_{Temp-ww,i} \in (20.928, 21.896]" \vee "x_{Temp-ww,i} \in [13.327, 21.896]")$$

Así pues asociando una regla compuesta a cada clase

$$\begin{aligned}
\mathbb{R}(\mathcal{P}_4) = \{ & \quad r_{Cr353} : (x_{Valve-air,i} \in (54.777, 69.898] \vee \\
& \quad x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\
& \quad (x_{Q-influent,i} \in (55.666, 85.092] \vee \\
& \quad x_{FR1-DOTOK,i} \in [44.167, 50.7]) \xrightarrow{0.4} Cr353, \\
r_{Cr357} : & (x_{Valve-air,i} \in (54.777, 69.898] \vee \\
& x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\
& (x_{Q-influent,i} \in [49.706, 55.666] \vee \\
& x_{FR1-DOTOK,i} \in [42.276, 44.167]) \xrightarrow{0.48} Cr357 \\
r_{Cr358} : & (x_{TN-influent,i} \in [0.0, 28.792]) \vee \\
& ("x_{Valve-air,i} \in [28.604, 54.777]) \wedge \\
& ("x_{Temp-ww,i} \in [8.472, 13.327]) \vee \\
& ("x_{Temp-ww,i} \in [8.472, 20.928]) \xrightarrow{0.3} Cr358, \\
r_{Cr360} : & ("x_{TN-influent,i} \in [0.0, 28.792]) \vee \\
& ("x_{Valve-air,i} \in [28.604, 54.777]) \wedge \\
& ("x_{Temp-ww,i} \in (20.928, 21.896]) \vee \\
& ("x_{Temp-ww,i} \in [13.327, 21.896]) \xrightarrow{0.45} Cr360
\end{aligned}$$

22.3.1 Interpretación final:

Que correspondería a la conceptualización:

- Cr353: “ $TN - influent$ no es bajo o $Valve - air$ es alto y $Q - influent$ es alto o $FR1 - DOTOK$ no es bajo”
- Cr357: “ $TN - influent$ no es bajo o $Valve - air$ es alto y $Q - influent$ no es alto o $FR1 - DOTOK$ es bajo”
- Cr358: “ $TN - influent$ es bajo o $Valve - air$ no es alto y $Temp - ww$ es bajo o $Temp - ww$ no es alto”
- Cr360: “ $TN - influent$ es bajo o $Valve - air$ no es alto y $Temp - ww$ no es bajo o $Temp - ww$ es alto”

22.3.2 Evaluación de la propuesta

Una vez determinados los conceptos asociados a cada clase se ha procedido a evaluar cual(es) era(n) satisfecho(s) por cada objeto de \mathcal{I} con tal de evaluar la correcta correspondencia entre la muestra y la conceptualización inducida. La Tabla 22.18 muestra los resultados de calcular el soporte $Sup(r)$, ver ecuación (9.1), la cobertura relativa $CovR(r)$, ver ecuación (9.5) y la confianza $p(r)$, ver ecuación (9.2) de cada una de las reglas compuestas inducidas para cada clase de la partición final.

La evaluación de la regla se hace con respecto al total de objetos de la base de datos, en este caso habrá inconsistencias al evaluar cada regla por separado ya que los intervalos que conforman cada regla compuesta no son disjuntos entre cada clase, es decir cada una de las reglas compuestas asociadas a cada clase presentan intersecciones, esto es debido a la forma en que se construyen los conceptos, ver ecuaciones (10.15) y (10.16).

Si se observa la Tabla 22.18, se tiene que el soporte se obtiene dividiendo cada celda de la tercera columna entre el total de objetos de la base de datos (365), la cobertura relativa se obtiene dividiendo cada celda de la cuarta columna entre la correspondiente celda de la quinta columna y la confianza dividiendo las celdas de la cuarta entre las correspondientes celdas de la tercera columna.

Como en las 4 clases ocurre que $\#\{i \in A_C^\xi\} > \#\{i \in A_C^\xi \cap i \in C\}$ se puede concluir que hay objetos mal asignados, es decir objetos que no satisfacen el consecuente de la regla, pero que satisfacen el antecedente de esta. Este valor no se considera en el soporte ni en la cobertura relativa, pero queda de manifiesto en la confianza ya que cuando se obtienen confianzas del 100% significa que todos los objetos que satisfacen el antecedente de la regla, también satisfacen simultáneamente el antecedente y el consecuente de la regla y por lo tanto pertenecen a la clase que la regla asigna. En este caso el número de objetos correctamente asignados por clase viene dado por $\#\{i \in A_C^\xi \cap i \in C\}$, pero el número de objetos mal asignados por clases no se puede calcular haciendo $\#\{i \in A_C^\xi\} - \#\{i \in A_C^\xi \cap i \in C\}$, como en el caso de la propuesta Best global concept and Close-World Assumption debido a que hay inconsistencias.

También es interesante observar que el número de objetos asignados por las reglas compuestas asociadas a cada clase se puede obtener de $\#\{i \in A_C^\xi\}$ y el valor porcentual viene cuantificado por el soporte, ya que si la suma total de los soportes por clase corresponde al porcentaje de objetos asignados, cuando este es menor que el 100% significa que hay objetos sin asignar y cuando es mayor, como en este caso, hay inconsistencias. En la Tabla 22.18 se puede inducir que la mayor parte de los objetos presentan inconsistencias.

La relación entre cobertura relativa y confianza es similar a la propuesta Best global concept and Close-World Assumption. aunque se puede considerar que la relación entre cobertura relativa y confianza es inversamente proporcional, a medida que se pierde confianzas se gana cobertura.

Ruler	Consec.	$\#\{i \in A_C^\xi\}$	$\#\{i \in A_C^\xi \cap i \in C\}$	n_c	$p(r)$	$Sup(r)$	$CovR(r)$
r_{Cr353}	$Cr353$	309	122	122	39,48%	84,66%	100%
r_{Cr357}	$Cr357$	101	48	50	47,52%	27,67%	96,00%
r_{Cr358}	$Cr358$	299	91	93	30,43%	81,92%	97,85%
r_{Cr360}	$Cr360$	220	99	100	45,00%	60,27%	99,00%
<i>Media</i>					40,61%		98,21%
<i>Suma</i>		929*	360	365		254,52%**	
<i>CovGlobal(\mathbb{R})</i>							98,6%

Tabla 22.18: Evaluación: Best local concept and Close-World Assumption.

(*) Hay inconsistencias, con lo cuál hay objetos que satisfacen el antecedente de más de una regla, es por esto que el número total de objetos que satisfacen el antecedente es mayor al número total de objetos de la base de datos.

(**) El soporte supera el 100% debido a las inconsistencias.

22.4 Interpretación de \mathcal{P}_4 utilizando Best local concept and partial Close-World Assumption:

1. $\xi = 2$: Así, $\mathcal{P}^{EnW,G}_{Lj3,R2} = \{Cr361, Cr362\}$. La raíz del árbol $Cr363$ tiene 2 hijos:

$$Cr363 \left\{ \begin{array}{l} Cr361 \\ Cr362. \end{array} \right.$$

2. La discretización realizada con el BbD se presenta en el Apéndice H.1
3. El sistema de reglas inducido para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(\mathcal{P}_2^*) = \bigcup_{k=1}^k \mathcal{R}(X_k, \mathcal{P}_2^*)$, se presenta en el Apéndice H.2
4. Como resultado de los pasos anteriores, se consideran los siguientes sistemas de reglas con reglas seguras $\mathcal{S}_{Cr361}(\mathcal{P}_2^*) \subseteq \mathcal{S}(\mathcal{P}_2^*)$:

$$\mathcal{S}_{Cr361}(\mathcal{P}_2^*) = \{ \begin{array}{ll} r_{3,Cr361}^{NH4-influent} : x_{NH4-influent,i} \in (40.541, 48.477] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (8.371, 8.785] & \xrightarrow{1.0} Cr361, \\ r_{1,Cr361}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 4.94) & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{Valve-air} : x_{Valve-air,i} \in (54.777, 69.898] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{Q-air} : x_{Q-air,i} \in (2030.77, 2201.27] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{h-ww} : x_{h-ww,i} \in (3.058, 3.098] & \xrightarrow{1.0} Cr361, \\ r_{1,Cr361}^{Q-influent} : x_{Q-influent,i} \in [49.706, 50.99) & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{Freq-rec} : x_{Freq-rec,i} \in (43.97, 44.0] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{TN-influent} : x_{TN-influent,i} \in (65.25, 83.792] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{Temp-ww} : x_{Temp-ww,i} \in (21.896, 22.583] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{TOC-influent} : x_{TOC-influent,i} \in (290.212, 355.0] & \xrightarrow{1.0} Cr361, \\ r_{3,Cr361}^{TOC-effluent} : x_{TOC-effluent,i} \in (44.053, 52.57] & \xrightarrow{1.0} Cr361 \end{array} \}$$

y $\mathcal{S}_{Cr362}(\mathcal{P}_2^*) \subseteq \mathcal{S}(\mathcal{P}_2^*)$.

$$\mathcal{S}_{Cr362}(\mathcal{P}_2^*) = \{ \begin{array}{ll} r_{1,Cr362}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.972] & \xrightarrow{1.0} Cr362, \\ r_{3,Cr362}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (9.846, 20.282] & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 3.297) & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{Valve-air} : x_{Valve-air,i} \in [28.604, 28.934) & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{Q-air} : x_{Q-air,i} \in [697.829, 739.819) & \xrightarrow{1.0} Cr362, \\ r_{3,Cr362}^{Q-influent} : x_{Q-influent,i} \in (85.092, 85.5] & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 42.276) & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{TN-influent} : x_{TN-influent,i} \in [0.0, 28.792] & \xrightarrow{1.0} Cr362, \\ r_{3,Cr362}^{TN-effluent} : x_{TN-effluent,i} \in (28.933, 34.867] & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{TOC-influent} : x_{TOC-influent,i} \in [0.0, 63.22) & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.0, 3.833) & \xrightarrow{1.0} Cr362, \\ r_{1,Cr362}^{TOC-effluent} : x_{TOC-effluent,i} \in [0.0, 2.014) & \xrightarrow{1.0} Cr362 \end{array} \}$$

5. Los Cuadros 22.26 y 22.27 muestran la cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$ y $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr361}^{NH4-influent}$	5	2,91%
$r_{3,Cr361}^{O2-1aerobic}$	1	0,58%
$r_{1,Cr361}^{O2-2aerobic}$	2	1,16%
$r_{3,Cr361}^{Valve-air}$	28	16,28%
$r_{3,Cr361}^{Q-air}$	27	15,70%
$r_{3,Cr361}^{h-ww}$	16	9,30%
$r_{1,Cr361}^{Q-influent}$	6	3,49%
$r_{3,Cr361}^{Freq-rec}$	3	1,74%
$r_{3,Cr361}^{TN-influent}$	9	5,23%
$r_{3,Cr361}^{Temp-ww}$	5	2,91%
$r_{3,Cr361}^{TOC-influent}$	20	11,63%
$r_{3,Cr361}^{TOC-effluent}$	10	5,81%

Tabla 22.19: Cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr362}^{NH4-influent}$	3	1,55%
$r_{3,Cr362}^{NH4-2aerobic}$	38	19,69%
$r_{1,Cr362}^{O2-1aerobic}$	4	2,07%
$r_{1,Cr362}^{Valve-air}$	5	2,59%
$r_{1,Cr362}^{Q-air}$	2	1,04%
$r_{3,Cr362}^{Q-influent}$	1	0,52%
$r_{1,Cr362}^{FRI-DOTOK}$	5	2,59%
$r_{1,Cr362}^{TN-influent}$	44	22,80%
$r_{3,Cr362}^{TN-effluent}$	14	7,25%
$r_{1,Cr362}^{TOC-influent}$	20	10,36%
$r_{1,Cr362}^{Nitritox-influent}$	4	2,07%
$r_{1,Cr362}^{TOC-effluent}$	1	0,52%

Tabla 22.20: Cobertura relativa de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$.

La regla con mayor cobertura relativa y $p_{sc} = 1$ de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$ es $r_{1,Cr362}^{Valve-air}$ con una $CovR(r)=16,28\%$ y la de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$ es $r_{1,Cr362}^{TN-influent}$ con una $CovR(r)=22,80\%$:

$$\begin{aligned} r_{3,Cr361}^{Valve-air} : x_{Valve-air,i} &\in (54.777, 69.898] \xrightarrow{1.0} Cr361 \\ r_{1,Cr362}^{TN-influent} : x_{TN-influent,i} &\in [0.0, 28.792) \xrightarrow{1.0} Cr362 \end{aligned}$$

Así,

$$A_{Cr361}^{2,Valve-air} = "x_{Valve-air,i} \in (54.777, 69.898]"$$

$$A_{Cr362}^{2,TN-influent} = "x_{TN-influent,i} \in [0.0, 28.792)"$$

En este caso no hay empate y tampoco coincide la variable, ver ecuaciones (10.23) y (10.24), por tanto:

- $A_{Cr361}^2 = A_{Cr361}^{2, Valve-air} \vee \neg A_{Cr362}^{2, TN-influent}$

$$A_{Cr361}^2 = "x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]"$$

- $A_{Cr362}^2 = A_{Cr362}^{2, TN-influent} \vee \neg A_{Cr361}^{2, Valve-air}$

$$A_{Cr362}^2 = "x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]"$$

6. Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$psCr361 = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{172}{321} = 0.54$$

$$psCr362 = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{181}{334} = 0.54$$

$$\mathbb{R}(\mathcal{P}_2) = \{ \quad r_{Cr361} : \quad x_{Valve-air,i} \in (54.777, 69.898] \vee \\ x_{TN-influent,i} \in [28.792, 83.792] \quad \xrightarrow{0.54} Cr361,$$

$$r_{Cr362} : \quad x_{TN-influent,i} \in [0.0, 28.792] \vee \\ x_{Valve-air,i} \in [28.604, 54.777] \quad \xrightarrow{0.54} Cr362 \}$$

7. $\xi = 3$: Así, $\mathcal{P}3_{Lj3,R2}^{EnW,G} = \{Cr362, Cr353, Cr357\}$. La clase $Cr361$ es la que se divide en 2 hijos y la clase $Cr362$ es la que ya estaba en la partición anterior:

$$Cr361 \left\{ \begin{array}{l} Cr353 \\ Cr357 \end{array} \right.$$

Así $C_i^3 = Cr353$; $C_j^3 = Cr357$; $C_t^2 = Cr361$.

8. La discretización realizada con el *BbD* de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_3^* = \{Cr353, Cr357\}$, donde $\mathcal{P}_3^* \subseteq \mathcal{P}3_{Lj3,R2}^{EnW,G}$, se presenta en el Apéndice H.3.
9. Con el *BbIR* se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_3^*)$, $k \in 1 : K$.
10. Como resultado de los pasos anteriores, se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_3^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_3^*)$ que se presenta en el Apéndice H.4 y se trabaja con $\mathcal{S}_{Cr357}(\mathcal{P}_3^*) \subseteq \mathcal{S}(\mathcal{P}_3^*)$ el cual es:

$$\mathcal{S}_{Cr357}(\mathcal{P}_3^*) = \{ \quad \begin{aligned} &r_{3,Cr357}^{NH4-influent} : x_{NH4-influent,i} \in (42.511, 48.477] \xrightarrow{1.0} Cr357, \\ &r_{1,Cr357}^{Q-influent} : x_{Q-influent,i} \in [49.706, 51.123] \xrightarrow{1.0} Cr357, \\ &r_{1,Cr357}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [42.276, 44.167] \xrightarrow{1.0} Cr357, \\ &r_{3,Cr357}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (38.333, 53.0] \xrightarrow{1.0} Cr357 \end{aligned} \}$$

y $\mathcal{S}_{Cr353}(\mathcal{P}_3^*) \subseteq \mathcal{S}(\mathcal{P}_3^*)$

$$\mathcal{S}_{Cr353}(\mathcal{P}_3^*) = \{ \begin{array}{ll} r_{1,Cr353}^{NH4-influent} : x_{NH4-influent,i} \in [7.972, 23.23] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (8.262, 9.846] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (7.018, 8.785] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 5.675] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Value-air} : x_{Value-air,i} \in (57.442, 69.898] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Q-air} : x_{Q-air,i} \in (2062.554, 2201.27] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{h-ww} : x_{h-ww,i} \in (3.055, 3.098] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Q-influent} : x_{Q-influent,i} \in (55.666, 85.092] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (47.265, 50.7] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Freq-rec} : x_{Freq-rec,i} \in (40.633, 44.0] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{TN-influent} : x_{TN-influent,i} \in [28.792, 40.417] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{TN-effluent} : x_{TN-effluent,i} \in (17.837, 28.933] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{Temp-ww} : x_{Temp-ww,i} \in [8.217, 11.235] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{TOC-influent} : x_{TOC-influent,i} \in [63.22, 89.833] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [3.833, 4.875] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{TOC-effluent} : x_{TOC-effluent,i} \in (36.476, 52.57] \xrightarrow{1.0} Cr353 \end{array} \}$$

11. Los Cuadros 22.28 y 22.29 muestran la coberturas relativas de las reglas de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$ y $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr357}^{NH4-influent}$	1	2,00%
$r_{1,Cr357}^{Q-influent}$	6	12,00%
$r_{1,Cr357}^{FR1-DOTOK}$	8	16,00%
$r_{3,Cr357}^{Nitritox-influent}$	1	2,00%

Tabla 22.21: Cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr353}^{NH4-influent}$	11	9,02%
$r_{3,Cr353}^{NH4-2aerobic}$	9	7,38%
$r_{3,Cr353}^{O2-1aerobic}$	9	7,38%
$r_{1,Cr353}^{O2-2aerobic}$	12	9,84%
$r_{3,Cr353}^{Value-air}$	11	9,02%
$r_{3,Cr353}^{Q-air}$	22	18,03%
$r_{3,Cr353}^{h-ww}$	20	16,39%
$r_{3,Cr353}^{Q-influent}$	100	81,97%
$r_{3,Cr353}^{FR1-DOTOK}$	98	80,33%
$r_{3,Cr353}^{Freq-rec}$	49	40,16%
$r_{1,Cr353}^{TN-influent}$	16	13,11%
$r_{3,Cr353}^{TN-effluent}$	81	66,39%
$r_{1,Cr353}^{Temp-ww}$	11	9,02%
$r_{1,Cr353}^{TOC-influent}$	13	10,66%
$r_{1,Cr353}^{Nitritox-influent}$	2	1,64%
$r_{3,Cr353}^{TOC-effluent}$	17	13,93%

Tabla 22.22: Cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$.

La regla con mayor cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$ es $r_{3,Cr353}^{Q-influent}$, con una cobertura relativa $CovR = 81,97\%$ y la regla con mayor cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$ es $r_{1,Cr357}^{FR1-DOTOK}$, con una cobertura relativa $CovR = 16\%$.

$$\begin{aligned} r_{3,Cr353}^{Q-influent} : x_{Q-influent,i} &\in (55.666, 85.092] \xrightarrow{1.0} Cr353 \\ r_{1,Cr357}^{FR1-DOTOK} : x_{FR1-DOTOK,i} &\in [42.276, 44.167) \xrightarrow{1.0} Cr357 \end{aligned}$$

Así,
 $A_{Cr353}^{3,Q-influent} = "x_{Q-influent,i} \in (55.666, 85.092]"$

$$A_{Cr357}^{3,FR1-DOTOK} = "x_{FR1-DOTOK,i} \in [42.276, 44.167)"$$

En este caso no hay empate y tampoco coincide la variable, ver ecuaciones (10.23) y (10.24), por tanto:

- $A_{Cr353}^{*3} = A_{Cr353}^{3,Q-influent} \vee \neg A_{Cr357}^{3,FR1-DOTOK}$

$$A_{Cr353}^{*3} = "x_{Q-influent,i} \in (55.666, 85.092]" \vee "x_{FR1-DOTOK,i} \in [44.167, 50.7]"$$

- $A_{Cr357}^{*3} = A_{Cr357}^{3,FR1-DOTOK} \vee \neg A_{Cr353}^{3,Q-influent}$

$$A_{Cr357}^{*3} = "x_{Q-influent,i} \in [49.706, 55.666]" \vee "x_{FR1-DOTOK,i} \in [42.276, 44.167]"$$

12. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

- $A_{Cr362}^3 = A_{Cr362}^2$

$$A_{Cr362}^3 = "x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]"$$

- $A_{Cr353}^3 = A_{Cr361}^2 \wedge A_{Cr353}^{*3}$

$$A_{Cr353}^3 = ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge ("x_{Q-influent,i} \in (55.666, 85.092]" \vee "x_{FR1-DOTOK,i} \in [44.167, 50.7]")$$

- $A_{Cr357}^3 = A_{Cr361}^2 \wedge A_{Cr357}^{*3}$

$$A_{Cr357}^3 = ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge ("x_{Q-influent,i} \in [49.706, 55.666]" \vee "x_{FR1-DOTOK,i} \in [42.276, 44.167]")$$

Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$p_{scCr353} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{122}{309} = 0.40$$

$$p_{scCr357} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{48}{101} = 0.48$$

$$\begin{aligned} \mathbb{R}(\mathcal{P}_3) = \{ & r_{Cr362} : \quad x_{TN-influent,i} \in [0.0, 28.792] \vee \\ & x_{Valve-air,i} \in [28.604, 54.777] \xrightarrow{0.54} Cr362 \} \\ r_{Cr353} : & (x_{Valve-air,i} \in (54.777, 69.898] \vee \\ & x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\ & (x_{Q-influent,i} \in (55.666, 85.092] \vee \\ & x_{FR1-DOTOK,i} \in [44.167, 50.7]) \xrightarrow{0.4} Cr353, \\ r_{Cr357} : & (x_{Valve-air,i} \in (54.777, 69.898] \vee \\ & x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\ & (x_{Q-influent,i} \in [49.706, 55.666] \vee \\ & x_{FR1-DOTOK,i} \in [42.276, 44.167]) \xrightarrow{0.48} Cr357 \} \end{aligned}$$

13. $\xi = 4$: Así, $\mathcal{P}_4^{EnW,G}_{Lj3,R2} = \{Cr358, Cr360, Cr353, Cr357\}$. La clase $Cr362$ de $\mathcal{P}_3^{EnW,G}_{Lj3,R2}$ tiene 2 hijos:

$$Cr362 \left\{ \begin{array}{l} Cr358 \\ Cr360 \end{array} \right.$$

Así $C_i^4 = Cr358$; $C_j^4 = Cr360$; $C_t^3 = Cr362$

14. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_4^* = \{Cr358, Cr360\}$, donde $\mathcal{P}_4^* \subseteq \mathcal{P}_4^{EnW,G}_{Lj3,R2}$, se presenta en el Apéndice H.5.
15. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_4^*)$, $k \in 1 : K$.
16. Como resultado de los pasos anteriores, se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_4^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_4^*)$ que se presenta en el Apéndice H.6 y se trabaja con $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$ el cual es:

$$\begin{aligned} \mathcal{S}_{Cr360}(\mathcal{P}_4^*) = \{ & r_{1,Cr360}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.79] \xrightarrow{1.0} Cr360, \\ & r_{1,Cr360}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 4.479] \xrightarrow{1.0} Cr360, \\ & r_{1,Cr360}^{Value-air} : x_{Valve-air,i} \in [28.604, 29.57] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{h-ww} : x_{h-ww,i} \in (3.039, 3.058] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (50.692, 50.733] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{Temp-ww} : x_{Temp-ww,i} \in (20.928, 21.896] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{TOC-influent} : x_{TOC-influent,i} \in (225.293, 290.212] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (30.0, 53.708] \xrightarrow{1.0} Cr360, \\ & r_{3,Cr360}^{TOC-effluent} : x_{TOC-effluent,i} \in (42.251, 44.053] \xrightarrow{1.0} Cr360 \} \end{aligned}$$

y $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$ es:

$$\begin{aligned} \mathcal{S}_{Cr358}(\mathcal{P}_4^*) = \{ & r_{3,Cr358}^{NH4-influent} : x_{NH4-influent,i} \in (34.353, 40.541] \xrightarrow{1.0} Cr358, \\ & r_{3,Cr358}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (2.856, 20.282] \xrightarrow{1.0} Cr358, \\ & r_{3,Cr358}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (6.998, 8.371] \xrightarrow{1.0} Cr358, \\ & r_{1,Cr358}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [4.94, 5.889] \xrightarrow{1.0} Cr358, \\ & r_{3,Cr358}^{Valve-air} : x_{Valve-air,i} \in (51.168, 54.777] \xrightarrow{1.0} Cr358, \end{aligned}$$

$$\begin{aligned}
r_{3,Cr358}^{Q-air} : x_{Q-air,i} \in (1770.852, 2030.77] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{h-ww} : x_{h-ww,i} \in [2.978, 2.984] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Q-influent} : x_{Q-influent,i} \in [50.99, 63.038] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 47.2] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Freq-rec} : x_{Freq-rec,i} \in [23.863, 24.899] &\xrightarrow{1.0} Cr358 , \\
r_{3,Cr358}^{TN-influent} : x_{TN-influent,i} \in (54.792, 65.25] &\xrightarrow{1.0} Cr358 , \\
r_{3,Cr358}^{TN-effluent} : x_{TN-effluent,i} \in (17.788, 34.867] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Temp-ww} : x_{Temp-ww,i} \in [8.472, 13.327] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{TOC-influent} : x_{TOC-influent,i} \in [0.0, 38.888] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.0, 0.542] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{TOC-effluent} : x_{TOC-effluent,i} \in [0.0, 11.879] &\xrightarrow{1.0} Cr358 \}
\end{aligned}$$

17. Los Cuadros 22.30 y 22.31 muestran la cobertura relativa de las reglas de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$ y $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr360}^{O2-1aerobic}$	28	28,00%
$r_{1,Cr360}^{Valve-air}$	5	5,00%
$r_{3,Cr360}^{h-ww}$	9	9,00%
$r_{3,Cr360}^{FR1-DOTOK}$	3	3,00%
$r_{3,Cr360}^{Temp-ww}$	29	29,00%
$r_{3,Cr360}^{TOC-influent}$	1	1,00%
$r_{3,Cr360}^{Nitritox-influent}$	3	3,00%
$r_{3,Cr360}^{TOC-effluent}$	1	1,00%

Tabla 22.23: Cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr358}^{NH4-2aerobic}$	54	58,06%
$r_{3,Cr358}^{O2-1aerobic}$	54	58,06%
$r_{1,Cr358}^{O2-2aerobic}$	5	5,38%
$r_{3,Cr358}^{Valve-air}$	3	3,23%
$r_{3,Cr358}^{Q-air}$	9	9,68%
$r_{1,Cr358}^{h-ww}$	6	6,45%
$r_{1,Cr358}^{Q-influent}$	33	35,48%
$r_{1,Cr358}^{FR1-DOTOK}$	34	36,56%
$r_{1,Cr358}^{Freq-rec}$	13	13,98%
$r_{3,Cr358}^{TN-influent}$	19	20,43%
$r_{3,Cr358}^{TN-effluent}$	38	40,86%
$r_{1,Cr358}^{Temp-ww}$	62	66,67%
$r_{1,Cr358}^{TOC-influent}$	1	1,08%
$r_{1,Cr358}^{Nitritox-influent}$	1	1,08%
$r_{1,Cr358}^{TOC-effluent}$	6	6,45%

Tabla 22.24: Cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$.

La regla con mayor cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$ es $r_{1,Cr358}^{Temp-ww}$, con una cobertura relativa $CovR = 66,67\%$ y la regla con mayor cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$ es

$r_{3,Cr358}^{Temp-ww}$, con una cobertura relativa $CovR = 29\%$.

$$\begin{aligned} r_{1,Cr358}^{Temp-ww} : x_{Temp-ww,i} \in [8.472, 13.327] &\xrightarrow{1.0} Cr358 \\ r_{3,Cr360}^{Temp-ww} : x_{Temp-ww,i} \in (20.928, 21.896] &\xrightarrow{1.0} Cr360 \end{aligned}$$

Así,
 $A_{Cr358}^{4,Temp-ww} = "x_{Temp-ww,i} \in [8.472, 13.327]"$
 $A_{Cr360}^{4,Temp-ww} = "x_{Temp-ww,i} \in (20.928, 21.896]"$

En este caso la variable con mayor cobertura relativa en cada sistema de reglas que interpreta la partición final es la misma (Temp-ww) y como no hay empate, ver ecuaciones (10.21) y (10.22), los conceptos que se asocian son los siguientes:

- $A_{Cr358}^{*4} = A_{Cr358}^{4,Temp-ww}$
 $A_{Cr358}^{*4} = "x_{Temp-ww,i} \in [8.472, 13.327]"$
- $A_{Cr360}^{*4} = A_{Cr360}^{4,Temp-ww}$
 $A_{Cr360}^{*4} = "x_{Temp-ww,i} \in (20.928, 21.896]"$

18. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

- $A_{Cr353}^4 = A_{Cr353}^3 = A_{Cr361}^2 \wedge A_{Cr353}^{*3}$
 $A_{Cr353}^3 = ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge ("x_{Q-influent,i} \in (55.666, 85.092]" \vee "x_{FR1-DOTOK,i} \in [44.167, 50.7]")$
- $A_{Cr357}^4 = A_{Cr357}^3 = A_{Cr361}^2 \wedge A_{Cr357}^{*3}$
 $A_{Cr357}^3 = ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge ("x_{Q-influent,i} \in [49.706, 55.666]" \vee "x_{FR1-DOTOK,i} \in [42.276, 44.167]")$
- $A_{Cr358}^4 = A_{Cr362}^2 \wedge A_{Cr358}^{*4}$
 $A_{Cr358}^4 = ("x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]") \wedge "x_{Temp-ww,i} \in [8.472, 13.327]"$
- $A_{Cr360}^4 = A_{Cr362}^2 \wedge A_{Cr360}^{*4}$
 $A_{Cr360}^4 = ("x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]") \wedge "x_{Temp-ww,i} \in (20.928, 21.896]"$

Así pues asociando una regla compuesta a cada clase

$$\begin{aligned}
\mathbb{R}(\mathcal{P}_4) = \{ & \quad r_{Cr353} : (x_{Valve-air,i} \in (54.777, 69.898] \vee \\
& x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\
& (x_{Q-influent,i} \in (55.666, 85.092] \vee \\
& x_{FR1-DOTOK,i} \in [44.167, 50.7]) \xrightarrow{0.4} Cr353, \\
r_{Cr357} : & (x_{Valve-air,i} \in (54.777, 69.898] \vee \\
& x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\
& (x_{Q-influent,i} \in [49.706, 55.666] \vee \\
& x_{FR1-DOTOK,i} \in [42.276, 44.167]) \xrightarrow{0.48} Cr357 \\
r_{Cr358} : & (x_{TN-influent,i} \in [0.0, 28.792]) \vee \\
& (x_{Valve-air,i} \in [28.604, 54.777]) \wedge \\
& (x_{Temp-ww,i} \in [8.472, 13.327]) \xrightarrow{0.54} Cr358, \\
r_{Cr360} : & (x_{TN-influent,i} \in [0.0, 28.792]) \vee \\
& (x_{Valve-air,i} \in [28.604, 54.777]) \wedge \\
& (x_{Temp-ww,i} \in (20.928, 21.896]) \xrightarrow{0.83} Cr360
\end{aligned}$$

22.4.1 Interpretación final

Que correspondería a la conceptualización:

- Cr353: “*TN – influent* no es bajo o *Valve – air* es alto y *Q – influent* es alto o *FR1 – DOTOK* no es bajo”
- Cr357: “*TN – influent* no es bajo o *Valve – air* es alto y *Q – influent* no es alto o *FR1 – DOTOK* es bajo”
- Cr358: “*TN – influent* es bajo o *Valve – air* no es alto y *Temp – ww* es bajo”
- Cr360: “*TN – influent* es bajo o *Valve – air* no es alto y *Temp – ww* es alto”

22.4.2 Evaluación de la propuesta

Una vez determinados los conceptos asociados a cada clase se ha procedido a evaluar cual(es) era(n) satisfecho(s) por cada objeto de \mathcal{I} con tal de evaluar la correcta correspondencia entre la muestra y la conceptualización inducida. La Tabla 22.25 muestra los resultados de calcular el soporte $Sup(r)$, ver ecuación (9.1), la cobertura relativa $CovR(r)$, ver ecuación (9.5) y la confianza $p(r)$, ver ecuación (9.2) de cada una de las reglas compuestas inducidas para cada clase de la partición final.

La evaluación de cada regla se hace con respecto al total de objetos de la base de datos, es evidente que habrá inconsistencias al evaluar cada regla por separado ya que los intervalos que conforman cada regla compuesta, no son disjuntos entre una clase y otra debido a la forma en que se construyen los conceptos, ver ecuaciones (10.23),(10.24).

Si se observa la Tabla 22.25, se tiene que el soporte se obtiene dividiendo cada celda de la tercera columna entre el total de objetos de la base de datos (365), la cobertura relativa se obtiene dividiendo cada celda de la cuarta columna entre la correspondiente celda de la

quinta columna y la confianza dividiendo las celdas de la cuarta entre las correspondientes celdas de la tercera columna.

Ruler	Consec.	$\#\{i \in A_C^\xi\}$	$\#\{i \in A_C^\xi \cap i \in C\}$	n_c	$p(r)$	$Sup(r)$	$CovR(r)$
r_{Cr353}	$Cr353$	309	122	122	39,48%	84,66%	100,00%
r_{Cr357}	$Cr357$	101	48	50	47,52%	27,67%	96,00%
r_{Cr358}	$Cr358$	114	62	93	54,39%	31,23%	66,67%
r_{Cr360}	$Cr360$	35	29	100	82,86%	9,59%	29,00%
<i>Media</i>					56,06%		72,92%
<i>Suma</i>		559*	261	365		153,15%**	
$CovG_{global}(\mathbb{R})$							71,5%

Tabla 22.25: Evaluación: Best local concept and partial Close-World Assumption.

(*) Hay inconsistencias, con lo cuál hay objetos que satisfacen el antecedente de más de una regla, es por esto que el número total de objetos que satisfacen el antecedente es mayor al número total de objetos de la base de datos. (**) *Soporte* > 100% debido a las inconsistencias.

Como en todas las clases se cumple que $\#\{i \in A_C^\xi\} > \#\{i \in A_C^\xi \cap i \in C\}$ se puede concluir que hay objetos mal asignados, es decir objetos que no satisfacen el consecuente de la regla, pero que satisfacen el antecedente de ésta. Este valor no se considera en el soporte ni en la cobertura relativa, pero queda de manifiesto en la confianza ya que cuando se obtienen confianzas del 100% significa que todos los objetos que satisfacen el antecedente de la regla, también satisfacen simultáneamente el antecedente y el consecuente de la regla y por lo tanto pertenecen a la clase que la regla asigna. En este caso se puede concluir que el número de objetos correctamente asignados por clase viene dado por $\#\{i \in A_C^\xi \cap i \in C\}$, pero el número de objetos mal asignados por clases no se puede calcular haciendo $\#\{i \in A_C^\xi\} - \#\{i \in A_C^\xi \cap i \in C\}$, como en el caso de la propuesta Best global concept and Close-World Assumption debido a que hay inconsistencias.

También es interesante observar que el número de objetos asignados por el sistema de reglas final $\mathbb{R}(\mathcal{P}_4)$ se puede obtener de $\#\{i \in A_C^\xi\}$ y el valor total viene cuantificado por el soporte, ya que si la suma total de los soportes por clase corresponde al porcentaje de objetos asignados, cuando este es menor que el 100% significa que hay objetos sin asignar, pero cuando es mayor, como en este caso, ver Tabla 22.25, significa que hay inconsistencias.

22.5 Interpretación de \mathcal{P}_4 utilizando Best local-global concept and Close-World Assumption:

1. $\xi = 2$: Así, $\mathcal{P}^{EnW,G}_{Lj3,R2} = \{Cr361, Cr362\}$. La raíz del árbol $Cr363$ tiene 2 hijos:

$$Cr363 \left\{ \begin{array}{l} Cr361 \\ Cr362. \end{array} \right.$$

2. La discretización realizada con el BbD se presenta en el Apéndice H.1
3. El sistema de reglas inducido para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(\mathcal{P}_2^*) = \bigcup_{k=1}^k \mathcal{R}(X_k, \mathcal{P}_2^*)$, se presenta en el Apéndice H.2
4. Como resultado de los pasos anteriores, se consideran los siguientes sistemas de reglas, con reglas seguras $\mathcal{S}_{Cr361}(\mathcal{P}_2^*) \subseteq \mathcal{S}(\mathcal{P}_2^*)$:

$$\mathcal{S}_{Cr361}(\mathcal{P}_2^*) = \{ \begin{array}{ll} r_{3,Cr361}^{NH4-influent} : x_{NH4-influent,i} \in (40.541, 48.477] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (8.371, 8.785] \xrightarrow{1.0} Cr361 , \\ r_{1,Cr361}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 4.94) \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{Valve-air} : x_{Valve-air,i} \in (54.777, 69.898] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{Q-air} : x_{Q-air,i} \in (2030.77, 2201.27] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{h-ww} : x_{h-ww,i} \in (3.058, 3.098] \xrightarrow{1.0} Cr361 , \\ r_{1,Cr361}^{Q-influent} : x_{Q-influent,i} \in [49.706, 50.99) \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{Freq-rec} : x_{Freq-rec,i} \in (43.97, 44.0] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{TN-influent} : x_{TN-influent,i} \in (65.25, 83.792] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{Temp-ww} : x_{Temp-ww,i} \in (21.896, 22.583] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{TOC-influent} : x_{TOC-influent,i} \in (290.212, 355.0] \xrightarrow{1.0} Cr361 , \\ r_{3,Cr361}^{TOC-effluent} : x_{TOC-effluent,i} \in (44.053, 52.57] \xrightarrow{1.0} Cr361 \end{array} \}$$

y $\mathcal{S}_{Cr362}(\mathcal{P}_2^*) \subseteq \mathcal{S}(\mathcal{P}_2^*)$.

$$\mathcal{S}_{Cr362}(\mathcal{P}_2^*) = \{ \begin{array}{ll} r_{1,Cr362}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.972) \xrightarrow{1.0} Cr362 , \\ r_{3,Cr362}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (9.846, 20.282] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 3.297) \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{Valve-air} : x_{Valve-air,i} \in [28.604, 28.934) \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{Q-air} : x_{Q-air,i} \in [697.829, 739.819) \xrightarrow{1.0} Cr362 , \\ r_{3,Cr362}^{Q-influent} : x_{Q-influent,i} \in (85.092, 85.5] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 42.276] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{TN-influent} : x_{TN-influent,i} \in [0.0, 28.792) \xrightarrow{1.0} Cr362 , \\ r_{3,Cr362}^{TN-effluent} : x_{TN-effluent,i} \in (28.933, 34.867] \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{TOC-influent} : x_{TOC-influent,i} \in [0.0, 63.22) \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.0, 3.833) \xrightarrow{1.0} Cr362 , \\ r_{1,Cr362}^{TOC-effluent} : x_{TOC-effluent,i} \in [0.0, 2.014) \xrightarrow{1.0} Cr362 \end{array} \}$$

5. Los Cuadros 22.26 y 22.27 muestran la cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$ y $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr361}^{NH4-influent}$	5	2,91%
$r_{3,Cr361}^{O2-1aerobic}$	1	0,58%
$r_{1,Cr361}^{O2-2aerobic}$	2	1,16%
$r_{3,Cr361}^{Valve-air}$	28	16,28%
$r_{3,Cr361}^{Q-air}$	27	15,70%
$r_{3,Cr361}^{h-ww}$	16	9,30%
$r_{1,Cr361}^{Q-influent}$	6	3,49%
$r_{3,Cr361}^{Freq-rec}$	3	1,74%
$r_{3,Cr361}^{TN-influent}$	9	5,23%
$r_{3,Cr361}^{Temp-ww}$	5	2,91%
$r_{3,Cr361}^{TOC-influent}$	20	11,63%
$r_{3,Cr361}^{TOC-effluent}$	10	5,81%

Tabla 22.26: Cobertura relativa de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr362}^{NH4-influent}$	3	1,55%
$r_{3,Cr362}^{NH4-2aerobic}$	38	19,69%
$r_{1,Cr362}^{O2-1aerobic}$	4	2,07%
$r_{1,Cr362}^{Valve-air}$	5	2,59%
$r_{1,Cr362}^{Q-air}$	2	1,04%
$r_{3,Cr362}^{Q-influent}$	1	0,52%
$r_{1,Cr362}^{FRI-DOTOK}$	5	2,59%
$r_{1,Cr362}^{TN-influent}$	44	22,80%
$r_{3,Cr362}^{TN-effluent}$	14	7,25%
$r_{1,Cr362}^{TOC-influent}$	20	10,36%
$r_{1,Cr362}^{Nitritox-influent}$	4	2,07%
$r_{1,Cr362}^{TOC-effluent}$	1	0,52%

Tabla 22.27: Cobertura relativa de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$.

La regla con mayor cobertura relativa y $p_{sc} = 1$ de $\mathcal{S}_{Cr361}(\mathcal{P}_2^*)$ es $r_{1,Cr362}^{Valve-air}$ con una $CovR(r)=16,28\%$ y la de $\mathcal{S}_{Cr362}(\mathcal{P}_2^*)$ es $r_{1,Cr362}^{TN-influent}$ con una $CovR(r)=22,80\%$:

$$\begin{aligned} r_{3,Cr361}^{Valve-air} : x_{Valve-air,i} &\in (54.777, 69.898] \xrightarrow{1.0} Cr361 \\ r_{1,Cr362}^{TN-influent} : x_{TN-influent,i} &\in [0.0, 28.792) \xrightarrow{1.0} Cr362 \end{aligned}$$

Así,

$$\begin{aligned} A_{Cr361}^{2,Valve-air} &= "x_{Valve-air,i} \in (54.777, 69.898]" \\ A_{Cr362}^{2,TN-influent} &= "x_{TN-influent,i} \in [0.0, 28.792)" \end{aligned}$$

En este caso no hay empate en las coberturas y la variable tampoco coincide, ver ecuaciones (11.1) y (11.2), por tanto:

- $A_{Cr361}^2 = A_{Cr361}^{2, Valve-air} \vee \neg A_{Cr362}^{2, TN-influent}$

$$A_{Cr361}^2 = "x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]"$$

- $A_{Cr362}^2 = A_{Cr362}^{2, TN-influent} \vee \neg A_{Cr361}^{2, Valve-air}$

$$A_{Cr362}^2 = "x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]"$$

6. Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$p_{sCr361} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{172}{321} = 0.54$$

$$p_{sCr362} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{181}{334} = 0.54$$

$$\mathbb{R}(\mathcal{P}_2) = \{ \quad r_{Cr361} : \quad x_{Valve-air,i} \in (54.777, 69.898] \vee \\ x_{TN-influent,i} \in [28.792, 83.792] \quad \xrightarrow{0.54} Cr361,$$

$$r_{Cr362} : \quad x_{TN-influent,i} \in [0.0, 28.792] \vee \\ x_{Valve-air,i} \in [28.604, 54.777] \quad \xrightarrow{0.54} Cr362 \}$$

7. $\xi = 3$: Así, $\mathcal{P}3_{Lj3,R2}^{EnW,G} = \{Cr362, Cr353, Cr357\}$. La clase $Cr361$ es la que se divide en 2 hijos y la clase $Cr362$ es la que ya estaba en la partición anterior:

$$Cr361 \left\{ \begin{array}{l} Cr353 \\ Cr357 \end{array} \right.$$

Así $C_i^3 = Cr353$; $C_j^3 = Cr357$; $C_t^2 = Cr361$.

8. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_3^* = \{Cr353, Cr357\}$, donde $\mathcal{P}_3^* \subseteq \mathcal{P}3_{Lj3,R2}^{EnW,G}$, se presenta en el Apéndice H.3.
9. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_3^*)$, $k \in 1 : K$.
10. Como resultado de los pasos anteriores, se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_3^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_3^*)$ que se se presenta en el Apéndice H.4 y se trabaja con $\mathcal{S}_{Cr357}(\mathcal{P}_3^*) \subseteq \mathcal{S}(\mathcal{P}_3^*)$ el cual es:

$$\mathcal{S}_{Cr357}(\mathcal{P}_3^*) = \{ \quad \begin{aligned} &r_{3,Cr357}^{NH4-influent} : x_{NH4-influent,i} \in (42.511, 48.477] \xrightarrow{1.0} Cr357, \\ &r_{1,Cr357}^{Q-influent} : x_{Q-influent,i} \in [49.706, 51.123] \xrightarrow{1.0} Cr357, \\ &r_{1,Cr357}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [42.276, 44.167] \xrightarrow{1.0} Cr357, \\ &r_{3,Cr357}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (38.333, 53.0] \xrightarrow{1.0} Cr357 \end{aligned} \}$$

y $\mathcal{S}_{Cr353}(\mathcal{P}_3^*) \subseteq \mathcal{S}(\mathcal{P}_3^*)$

$$\mathcal{S}_{Cr353}(\mathcal{P}_3^*) = \{ \begin{array}{ll} r_{1,Cr353}^{NH4-influent} : x_{NH4-influent,i} \in [7.972, 23.23] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (8.262, 9.846] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (7.018, 8.785] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 5.675] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Value-air} : x_{Value-air,i} \in (57.442, 69.898] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Q-air} : x_{Q-air,i} \in (2062.554, 2201.27] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{h-ww} : x_{h-ww,i} \in (3.055, 3.098] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Q-influent} : x_{Q-influent,i} \in (55.666, 85.092] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (47.265, 50.7] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{Freq-rec} : x_{Freq-rec,i} \in (40.633, 44.0] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{TN-influent} : x_{TN-influent,i} \in [28.792, 40.417] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{TN-effluent} : x_{TN-effluent,i} \in (17.837, 28.933] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{Temp-ww} : x_{Temp-ww,i} \in [8.217, 11.235] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{TOC-influent} : x_{TOC-influent,i} \in [63.22, 89.833] \xrightarrow{1.0} Cr353 , \\ r_{1,Cr353}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [3.833, 4.875] \xrightarrow{1.0} Cr353 , \\ r_{3,Cr353}^{TOC-effluent} : x_{TOC-effluent,i} \in (36.476, 52.57] \xrightarrow{1.0} Cr353 \end{array} \}$$

11. Los Cuadros 22.28 y 22.29 muestran la coberturas relativas de las reglas de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$ y $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr357}^{NH4-influent}$	1	2,00%
$r_{1,Cr357}^{Q-influent}$	6	12,00%
$r_{1,Cr357}^{FR1-DOTOK}$	8	16,00%
$r_{3,Cr357}^{Nitritox-influent}$	1	2,00%

Tabla 22.28: Cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr353}^{NH4-influent}$	11	9,02%
$r_{3,Cr353}^{NH4-2aerobic}$	9	7,38%
$r_{3,Cr353}^{O2-1aerobic}$	9	7,38%
$r_{1,Cr353}^{O2-2aerobic}$	12	9,84%
$r_{3,Cr353}^{Value-air}$	11	9,02%
$r_{3,Cr353}^{Q-air}$	22	18,03%
$r_{3,Cr353}^{h-ww}$	20	16,39%
$r_{3,Cr353}^{Q-influent}$	100	81,97%
$r_{3,Cr353}^{FR1-DOTOK}$	98	80,33%
$r_{3,Cr353}^{Freq-rec}$	49	40,16%
$r_{1,Cr353}^{TN-influent}$	16	13,11%
$r_{3,Cr353}^{TN-effluent}$	81	66,39%
$r_{1,Cr353}^{Temp-ww}$	11	9,02%
$r_{1,Cr353}^{TOC-influent}$	13	10,66%
$r_{1,Cr353}^{Nitritox-influent}$	2	1,64%
$r_{3,Cr353}^{TOC-effluent}$	17	13,93%

Tabla 22.29: Cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$.

La regla con mayor cobertura relativa de $\mathcal{S}_{Cr353}(\mathcal{P}_3^*)$ es $r_{3,Cr353}^{Q-influent}$, con una cobertura relativa $CovR = 81,97\%$ y la regla con mayor cobertura relativa de $\mathcal{S}_{Cr357}(\mathcal{P}_3^*)$ es $r_{1,Cr357}^{FR1-DOTOK}$, con una cobertura relativa $CovR = 16\%$.

$$\begin{aligned} r_{3,Cr353}^{Q-influent} : x_{Q-influent,i} &\in (55.666, 85.092] \xrightarrow{1.0} Cr353 \\ r_{1,Cr357}^{FR1-DOTOK} : x_{FR1-DOTOK,i} &\in [42.276, 44.167) \xrightarrow{1.0} Cr357 \end{aligned}$$

Así,
 $A_{Cr353}^{3,Q-influent} = "x_{Q-influent,i} \in (55.666, 85.092]"$

$$A_{Cr357}^{3,FR1-DOTOK} = "x_{FR1-DOTOK,i} \in [42.276, 44.167)"$$

En este caso no hay empate en las coberturas y la variable tampoco coincide, ver ecuaciones (11.1) y (11.2), por tanto:

- $A_{Cr353}^{*3} = A_{Cr353}^{3,Q-influent} \vee \neg A_{Cr357}^{3,FR1-DOTOK}$
- $A_{Cr353}^{*3} = "x_{Q-influent,i} \in (55.666, 85.092]" \vee "x_{FR1-DOTOK,i} \in [44.167, 50.7]"$
- $A_{Cr357}^{*3} = A_{Cr357}^{3,FR1-DOTOK} \vee \neg A_{Cr353}^{3,Q-influent}$
- $A_{Cr357}^{*3} = "x_{Q-influent,i} \in [49.706, 55.666]" \vee "x_{FR1-DOTOK,i} \in [42.276, 44.167]"$

12. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

- $A_{Cr362}^3 = A_{Cr362}^2$
- $A_{Cr362}^3 = "x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]"$
- $A_{Cr353}^3 = A_{Cr361}^2 \wedge A_{Cr353}^{*3}$
- $A_{Cr353}^3 = ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge ("x_{Q-influent,i} \in (55.666, 85.092]" \vee "x_{FR1-DOTOK,i} \in [44.167, 50.7]")$
- $A_{Cr357}^3 = A_{Cr361}^2 \wedge A_{Cr357}^{*3}$
- $A_{Cr357}^3 = ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge ("x_{Q-influent,i} \in [49.706, 55.666]" \vee "x_{FR1-DOTOK,i} \in [42.276, 44.167]")$

Así pues asociando una regla compuesta a cada clase y calculando los p_{sc} de las nuevas reglas:

$$p_{sCr353} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{122}{309} = 0.40$$

$$p_{sCr357} = \frac{\text{card}\{i \in C \text{ tq } A(i)=\text{true}\}}{\text{card}\{A(i)=\text{true}\}} = \frac{48}{101} = 0.48$$

$$\begin{aligned} \mathbb{R}(\mathcal{P}_3) = \{ & \quad r_{Cr362} : \quad x_{TN-influent,i} \in [0.0, 28.792] \vee \\ & \quad x_{Valve-air,i} \in [28.604, 54.777] \xrightarrow{0.54} Cr362 \} \\ r_{Cr353} : & \quad (x_{Valve-air,i} \in (54.777, 69.898] \vee \\ & \quad x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\ & \quad (x_{Q-influent,i} \in (55.666, 85.092] \vee \\ & \quad x_{FR1-DOTOK,i} \in [44.167, 50.7]) \xrightarrow{0.4} Cr353, \\ r_{Cr357} : & \quad (x_{Valve-air,i} \in (54.777, 69.898] \vee \\ & \quad x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\ & \quad (x_{Q-influent,i} \in [49.706, 55.666] \vee \\ & \quad x_{FR1-DOTOK,i} \in [42.276, 44.167]) \xrightarrow{0.48} Cr357 \} \end{aligned}$$

13. $\xi = 4$: Así, $\mathcal{P}_4^{EnW,G}_{Lj3,R2} = \{Cr358, Cr360, Cr353, Cr357\}$. La clase $Cr362$ de $\mathcal{P}_3^{EnW,G}_{Lj3,R2}$ tiene 2 hijos:

$$Cr362 \left\{ \begin{array}{l} Cr358 \\ Cr360 \end{array} \right.$$

Así $C_i^4 = Cr358$; $C_j^4 = Cr360$; $C_t^3 = Cr362$

14. La discretización realizada con el BbD de todas las X_k , $k \in 1 : K$ según la partición restringida $\mathcal{P}_4^* = \{Cr358, Cr360\}$, donde $\mathcal{P}_4^* \subseteq \mathcal{P}_4^{EnW,G}_{Lj3,R2}$, se presenta en el Apéndice H.5.
15. Con el BbIR se obtienen los sistemas de reglas inducidos para todas las variables numéricas que caracteriza ambas clases, $\mathcal{R}(X_k, \mathcal{P}_4^*)$, $k \in 1 : K$.
16. Como resultado de los pasos anteriores, se considera un único sistema de reglas $\mathcal{R}(\mathcal{P}_4^*) = \bigcup_{k=1}^K \mathcal{R}(X_k, \mathcal{P}_4^*)$ que se presenta en el Apéndice H.6 y se trabaja con $\mathcal{S}_{Cr360}(\mathcal{P}_4^*) \subseteq \mathcal{S}(\mathcal{P}_4^*)$ el cual es:

$$\begin{aligned} \mathcal{S}_{Cr360}(\mathcal{P}_4^*) = \{ & \quad r_{1,Cr360}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.79] \xrightarrow{1.0} Cr360, \\ & \quad r_{1,Cr360}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 4.479] \xrightarrow{1.0} Cr360, \\ & \quad r_{1,Cr360}^{Valve-air} : x_{Valve-air,i} \in [28.604, 29.57] \xrightarrow{1.0} Cr360, \\ & \quad r_{3,Cr360}^{h-ww} : x_{h-ww,i} \in (3.039, 3.058] \xrightarrow{1.0} Cr360, \\ & \quad r_{3,Cr360}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (50.692, 50.733] \xrightarrow{1.0} Cr360, \\ & \quad r_{3,Cr360}^{Temp-ww} : x_{Temp-ww,i} \in (20.928, 21.896] \xrightarrow{1.0} Cr360, \\ & \quad r_{3,Cr360}^{TOC-influent} : x_{TOC-influent,i} \in (225.293, 290.212] \xrightarrow{1.0} Cr360, \\ & \quad r_{3,Cr360}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (30.0, 53.708] \xrightarrow{1.0} Cr360, \\ & \quad r_{3,Cr360}^{TOC-effluent} : x_{TOC-effluent,i} \in (42.251, 44.053] \xrightarrow{1.0} Cr360 \} \end{aligned}$$

y $\mathcal{S}_{Cr358}(\mathcal{P}_4^*) \subseteq \mathcal{S}(\mathcal{P}_4^*)$ es:

$$\begin{aligned} \mathcal{S}_{Cr358}(\mathcal{P}_4^*) = \{ & \quad r_{3,Cr358}^{NH4-influent} : x_{NH4-influent,i} \in (34.353, 40.541] \xrightarrow{1.0} Cr358, \\ & \quad r_{3,Cr358}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (2.856, 20.282] \xrightarrow{1.0} Cr358, \\ & \quad r_{3,Cr358}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (6.998, 8.371] \xrightarrow{1.0} Cr358, \\ & \quad r_{1,Cr358}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [4.94, 5.889] \xrightarrow{1.0} Cr358, \\ & \quad r_{3,Cr358}^{Valve-air} : x_{Valve-air,i} \in (51.168, 54.777] \xrightarrow{1.0} Cr358, \\ & \quad r_{3,Cr358}^{Q-air} : x_{Q-air,i} \in (1770.852, 2030.77] \xrightarrow{1.0} Cr358, \end{aligned}$$

$$\begin{aligned}
r_{1,Cr358}^{h-ww} : x_{h-ww,i} \in [2.978, 2.984] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Q-influent} : x_{Q-influent,i} \in [50.99, 63.038] &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 47.2) &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Freq-rec} : x_{Freq-rec,i} \in [23.863, 24.899) &\xrightarrow{1.0} Cr358 , \\
r_{3,Cr358}^{TN-influent} : x_{TN-influent,i} \in (54.792, 65.25] &\xrightarrow{1.0} Cr358 , \\
r_{3,Cr358}^{TN-effluent} : x_{TN-effluent,i} \in (17.788, 34.867] &\xrightarrow{1.0} Cr358 , \\
\mathbf{r}_{1,Cr358}^{Temp-ww} : x_{Temp-ww,i} \in [8.472, 13.327) &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{TOC-influent} : x_{TOC-influent,i} \in [0.0, 38.888) &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.0, 0.542) &\xrightarrow{1.0} Cr358 , \\
r_{1,Cr358}^{TOC-effluent} : x_{TOC-effluent,i} \in [0.0, 11.879) &\xrightarrow{1.0} Cr358 \quad \}
\end{aligned}$$

17. Los Cuadros 22.30 y 22.31 muestran la cobertura relativa de las reglas de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$ y $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$ respectivamente.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{1,Cr360}^{O2-1aerobic}$	28	28,00%
$r_{1,Cr360}^{Valve-air}$	5	5,00%
$r_{3,Cr360}^{h-ww}$	9	9,00%
$r_{3,Cr360}^{FR1-DOTOK}$	3	3,00%
$\mathbf{r}_{3,Cr360}^{Temp-ww}$	29	29,00%
$r_{3,Cr360}^{TOC-influent}$	1	1,00%
$r_{3,Cr360}^{Nitritox-influent}$	3	3,00%
$r_{3,Cr360}^{TOC-effluent}$	1	1,00%

Tabla 22.30: Cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$.

Regla	$\#\{i \in I_s^{k,\xi} \cap i \in C\}$	$CovR(r)$
$r_{3,Cr358}^{NH4-2aerobic}$	54	58,06%
$r_{3,Cr358}^{O2-1aerobic}$	54	58,06%
$r_{1,Cr358}^{O2-2aerobic}$	5	5,38%
$r_{3,Cr358}^{Valve-air}$	3	3,23%
$r_{3,Cr358}^{Q-air}$	9	9,68%
$r_{1,Cr358}^{h-ww}$	6	6,45%
$r_{1,Cr358}^{Q-influent}$	33	35,48%
$r_{1,Cr358}^{FR1-DOTOK}$	34	36,56%
$r_{1,Cr358}^{Freq-rec}$	13	13,98%
$r_{3,Cr358}^{TN-influent}$	19	20,43%
$r_{3,Cr358}^{TN-effluent}$	38	40,86%
$\mathbf{r}_{1,Cr358}^{Temp-ww}$	62	66,67%
$r_{1,Cr358}^{TOC-influent}$	1	1,08%
$r_{1,Cr358}^{Nitritox-influent}$	1	1,08%
$r_{1,Cr358}^{TOC-effluent}$	6	6,45%

Tabla 22.31: Cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$.

La regla con mayor cobertura relativa de $\mathcal{S}_{Cr358}(\mathcal{P}_4^*)$ es $r_{1,Cr358}^{Temp-ww}$, con una cobertura relativa $CovR = 66,67\%$ y la regla con mayor cobertura relativa de $\mathcal{S}_{Cr360}(\mathcal{P}_4^*)$ es

$r_{3,Cr358}^{Temp-ww}$, con una cobertura relativa $CovR = 29\%$.

$$r_{1,Cr358}^{Temp-ww} : x_{Temp-ww,i} \in [8.472, 13.327] \xrightarrow{1.0} Cr358$$

$$r_{3,Cr360}^{Temp-ww} : x_{Temp-ww,i} \in (20.928, 21.896] \xrightarrow{1.0} Cr360$$

Como la variable es la misma. La regla con mayor cobertura relativa de $\mathcal{S}(\mathcal{P}_4^*)$ es $r_{1,Cr358}^{Temp-ww}$, con una cobertura relativa $CovR = 66, 67\%$.

Así,

$$A_{Cr358}^{4,Temp-ww} = "x_{Temp-ww,i} \in [8.472, 13.327]"$$

$$A_{Cr360}^{4,Temp-ww} = \neg A_{Cr358}^{4,Temp-ww} = "x_{Temp-ww,i} \in [13.327, 21.896]"$$

En este caso no hay empate en las coberturas y la variable es la misma y como la cobertura relativa de la regla $r_{1,Cr358}^{Temp-ww}$ es mayor cobertura relativa de la regla $r_{3,Cr360}^{Temp-ww}$, ver ecuaciones (11.3) y (11.4), por tanto:

- $A_{Cr358}^{*4} = A_{Cr358}^{4,Temp-ww} = "x_{Temp-ww,i} \in [8.472, 13.327]"$
- $A_{Cr360}^{*4} = A_{Cr360}^{4,Temp-ww} = \neg A_{Cr358}^{4,Temp-ww} = "x_{Temp-ww,i} \in [13.327, 21.896]"$

18. Integrando el conocimiento extraído en esta iteración al de la iteración anterior:

$$\bullet A_{Cr353}^4 = A_{Cr353}^3 = A_{Cr361}^2 \wedge A_{Cr353}^{*3}$$

$$A_{Cr353}^3 = ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge ("x_{Q-influent,i} \in (55.666, 85.092]" \vee "x_{FR1-DOTOK,i} \in [44.167, 50.7]")$$

$$\bullet A_{Cr357}^4 = A_{Cr357}^3 = A_{Cr361}^2 \wedge A_{Cr357}^{*3}$$

$$A_{Cr357}^3 = ("x_{Valve-air,i} \in (54.777, 69.898]" \vee "x_{TN-influent,i} \in [28.792, 83.792]") \wedge ("x_{Q-influent,i} \in [49.706, 55.666]" \vee "x_{FR1-DOTOK,i} \in [42.276, 44.167]")$$

$$\bullet A_{Cr358}^4 = A_{Cr362}^2 \wedge A_{Cr358}^{*4}$$

$$A_{Cr358}^4 = ("x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]") \wedge "x_{Temp-ww,i} \in [8.472, 13.327]"$$

$$\bullet A_{Cr360}^4 = A_{Cr362}^2 \wedge A_{Cr360}^{*4}$$

$$A_{Cr360}^4 = ("x_{TN-influent,i} \in [0.0, 28.792]" \vee "x_{Valve-air,i} \in [28.604, 54.777]") \wedge "x_{Temp-ww,i} \in [13.327, 21.896]"$$

Así pues asociando una regla compuesta a cada clase

$$\begin{aligned}
\mathbb{R}(\mathcal{P}_4) = \{ & \quad r_{Cr353} : (x_{Valve-air,i} \in (54.777, 69.898] \vee \\
& \quad x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\
& \quad (x_{Q-influent,i} \in (55.666, 85.092] \vee \\
& \quad x_{FR1-DOTOK,i} \in [44.167, 50.7]) \xrightarrow{0.4} Cr353, \\
r_{Cr357} : & (x_{Valve-air,i} \in (54.777, 69.898] \vee \\
& x_{TN-influent,i} \in [28.792, 83.792]) \wedge \\
& (x_{Q-influent,i} \in [49.706, 55.666] \vee \\
& x_{FR1-DOTOK,i} \in [42.276, 44.167]) \xrightarrow{0.48} Cr357 \\
r_{Cr358} : & (x_{TN-influent,i} \in [0.0, 28.792]) \vee \\
& "x_{Valve-air,i} \in [28.604, 54.777]" \wedge \\
& "x_{Temp-ww,i} \in [8.472, 13.327]" \xrightarrow{1.0} Cr358, \\
r_{Cr360} : & ("x_{TN-influent,i} \in [0.0, 28.792]) \vee \\
& "x_{Valve-air,i} \in [28.604, 54.777]" \wedge \\
& "x_{Temp-ww,i} \in [13.327, 21.896]" \xrightarrow{1.0} Cr360
\end{aligned}$$

22.5.1 Interpretación final

Que correspondería a la conceptualización:

- Cr353: “*TN – influent* no es bajo o *Valve – air* es alto y *Q – influent* es alto o *FR1 – DOTOK* no es bajo”
- Cr357: “*TN – influent* no es bajo o *Valve – air* es alto y *Q – influent* no es alto o *FR1 – DOTOK* es bajo”
- Cr358: “*TN – influent* es bajo o *Valve – air* no es alto y *Temp – ww* es bajo”
- Cr360: “*TN – influent* es bajo o *Valve – air* no es alto y *Temp – ww* no es bajo”

22.5.2 Evaluación de la propuesta

Una vez determinados los conceptos asociados a cada clase se ha procedido a evaluar cual(es) era(n) satisfecho(s) por cada objeto de \mathcal{I} con tal de evaluar la correcta correspondencia entre la muestra y la conceptualización inducida. La Tabla 22.32 muestra los resultados de calcular el soporte $Sup(r)$, ver ecuación (9.1), la cobertura relativa $CovR(r)$, ver ecuación (9.5) y la confianza $p(r)$, ver ecuación (9.2) de cada una de las reglas compuestas inducidas para cada clase de la partición final.

La evaluación de la regla se hace con respecto al total de objetos de la base de datos, es evidente que habrá inconsistencias al evaluar cada regla por separado ya que los intervalos que conforman cada reglas compuesta, no son disjuntos entre una clase y otro debido a la forma en que se construyen los conceptos, ver ecuaciones (11.3), (11.4), (11.5), (11.6) y (11.1).

Si se observa la Tabla 22.32, se tiene que el soporte se obtiene dividiendo cada celda de la tercera columna entre el total de objetos de la base de datos (365), la cobertura relativa se obtiene dividiendo cada celda de la cuarta columna entre la correspondiente celda de la

quinta columna y la confianza dividiendo las celdas de la cuarta entre las correspondientes celdas de la tercera columna.

Como en todas las clases se cumple que $\#\{i \in A_C^\xi\} > \#\{i \in A_C^\xi \cap i \in C\}$ se puede concluir que hay objetos mal asignados, es decir objetos que no satisfacen el consecuente de la regla, pero que satisfacen el antecedente de esta. Este valor no se considera en el soporte ni en la cobertura relativa, pero queda de manifiesto en la confianza ya que cuando se obtienen confianzas del 100% significa que todos los objetos que satisfacen el antecedente de la regla, también satisfacen simultáneamente el antecedente y el consecuente de la regla y por lo tanto pertenecen a la clase que la regla asigna. En este caso, también, se puede concluir que el número de objetos correctamente asignados por clase viene dado por $\#\{i \in A_C^\xi \cap i \in C\}$, pero el número de objetos mal asignados por clases no se puede calcular haciendo $\#\{i \in A_C^\xi\} - \#\{i \in A_C^\xi \cap i \in C\}$, como en el caso de la propuesta Best global concept and Close-World Assumption debido a que hay inconsistencias.

También es interesante observar que el número de objetos asignados por el sistema de reglas final $\mathbb{R}(\mathcal{P}_4)$ se puede obtener a partir de $\#\{i \in A_C^\xi\}$ y el valor total viene cuantificado por el soporte, ya que si la suma total de los soportes por clase corresponde al porcentaje de objetos asignados, cuando este es menor que el 100% significa que hay objetos sin asignar. En la Tabla 22.32 se observa que el soporte total es mayor que el 100% con lo cual se comprueba que hay inconsistencias.

Ruler	Consec.	$\#\{i \in A_C^\xi\}$	$\#\{i \in A_C^\xi \cap i \in C\}$	n_c	$p(r)$	$Sup(r)$	$CovR(r)$
$rCr353$	$Cr353$	309	122	122	39,48%	84,66%	100,00%
$rCr357$	$Cr357$	101	48	50	47,52%	27,67%	96,00%
$rCr358$	$Cr358$	114	62	93	54,39%	31,23%	66,67%
$rCr360$	$Cr360$	220	99	100	45,00%	60,27%	99,00%
<i>Media</i>					46,60%		90,42%
<i>Suma</i>		744*	331	365		203,84%**	
<i>CovGlobal(\mathbb{R})</i>							90,7%

Tabla 22.32: Evaluación: Best local-global concept and Close-World Assumption.

(*) Hay inconsistencias, con lo cuál hay objetos que satisfacen el antecedente de más de una regla, es por esto que el número total de objetos que satisfacen el antecedente es mayor al número total de objetos de la base de datos. (**) Soporte > 100% debido a las inconsistencias.

Capítulo 23

Análisis y resultados, planta eslovena

23.1 Interpretación validada por el experto

En esta sección recordamos la interpretación que a juzgar por los expertos que participan en este estudio, es la que mejor caracteriza a las 4 clases provenientes de la partición objetivo $\mathcal{P}4_{Lj3,R2}^{EnW,G} = \{Cr358, Cr360, Cr353, Cr357\}$ y la que nos servirá para validar y comparar las interpretaciones obtenidas utilizando la propuesta metodológica *Best global concept and Close-World Assumption* con las 5 diferentes formas de integrar el conocimiento.

- *Cr353*, represents the plant operation under the high load. In this case influent nitrogen concentrations are high and also influent flow rate is quite high as well. Even though the oxygen concentration in the aerobic tanks are high this can not decrease the effluent nitrogen concentrations. It means that when the plant is overloaded high effluent concentrations at the effluent of the plant can be expected.
- *Cr357*, represents the situation when the influent flow rate is low, that is, when the hydraulic retention time of the plant is high. In this case we get quite low effluent nitrogen concentrations if of course oxygen concentration in the aerobic tank is high enough. It means when the influent flow rate to the plant is low the effluent concentrations of the plant can be obtained at the low level if the oxygen concentration in the aerobic tanks is high.
- *Cr358*, explains the situation when the wastewater temperature is low. In this case nitrogen removal efficiency of the plant is rather low. This is so because microorganisms in the tanks don't work so intensively in cold conditions and therefore higher concentrations at the effluent of the plant can be expected.
- *Cr360*, shows the situation when the wastewater temperature is high. In warmer conditions the microorganisms in the plant work faster, so the effluent nitrogen concentrations can be low even when the oxygen concentrations in the aerobic tanks are quite low.

23.2 Interpretaciones generadas por cada propuesta

A continuación se presenta el resumen de los resultados de las interpretaciones obtenidas tanto con la propuesta Best global concept and Close-World Assumption de construcción del concepto, así como de las 4 alternativas.

23.2.1 Best global concept and Close-World Assumption:

- Cr353: “ $TN - influent$ no es bajo y $Q - influent$ es alto”
- Cr357: “ $TN - influent$ no es bajo y $Q - influent$ no es alto”
- Cr358: “ $TN - influent$ es bajo y $Temp - ww$ es bajo”
- Cr360: “ $TN - influent$ es bajo y $Temp - ww$ no es bajo”

23.2.2 Best local concept and no Close-World Assumption:

- Cr353: “ $Valve - air$ es alto y $Q - influent$ es alto”
- Cr357: “ $Valve - air$ es alto y $FR1 - DOTOK$ es bajo”
- Cr358: “ $TN - influent$ es bajo y $Temp - ww$ es bajo”
- Cr360: “ $TN - influent$ es bajo y $Temp - ww$ es alto”

23.2.3 Best local concept and Close-World Assumption:

- Cr353: “($TN - influent$ no es bajo o $Valve - air$ es alto) y ($Q - influent$ es alto o $FR1 - DOTOK$ no es bajo)”
- Cr357: “($TN - influent$ no es bajo o $Valve - air$ es alto) y ($Q - influent$ no es alto o $FR1 - DOTOK$ es bajo)”
- Cr358: “($TN - influent$ es bajo o $Valve - air$ no es alto) y ($Temp - ww$ es bajo o $Temp - ww$ no es alto)”
- Cr360: “($TN - influent$ es bajo o $Valve - air$ no es alto) y ($Temp - ww$ no es bajo o $Temp - ww$ es alto)”

23.2.4 Best local concept and partial Close-World Assumption:

- Cr353: “($TN - influent$ no es bajo o $Valve - air$ es alto) y $Q - influent$ es alto o $FR1 - DOTOK$ no es bajo”
- Cr357: “($TN - influent$ no es bajo o $Valve - air$ es alto) y ($Q - influent$ no es alto o $FR1 - DOTOK$ es bajo)”
- Cr358: “($TN - influent$ es bajo o $Valve - air$ no es alto) y $Temp - ww$ es bajo”
- Cr360: “($TN - influent$ es bajo o $Valve - air$ no es alto) y $Temp - ww$ es alto”

23.2.5 Best local-global concept and Close-World Assumption:

- Cr353: “($TN - \text{influent}$ no es bajo o $\text{Valve} - \text{air}$ es alto) y ($Q - \text{influent}$ es alto o $FR1 - \text{DOTOK}$ no es bajo)”
- Cr357: “($TN - \text{influent}$ no es bajo o $\text{Valve} - \text{air}$ es alto) y ($Q - \text{influent}$ no es alto o $FR1 - \text{DOTOK}$ es bajo)”
- Cr358: “($TN - \text{influent}$ es bajo o $\text{Valve} - \text{air}$ no es alto) y ($Temp - \text{ww}$ es bajo)”
- Cr360: “($TN - \text{influent}$ es bajo o $\text{Valve} - \text{air}$ no es alto) y $Temp - \text{ww}$ no es bajo”

23.3 Análisis de los resultados

23.3.1 Análisis cualitativo

A modo de resumen se ha construido la Tabla 23.1, con los conceptos generados para cada clase y por cada propuesta.

Método	Clase	TN-influent	Q-influent	Temp-ww	Valve-air	FR1-DOTOK
BG & CWA	Cr353	no bajo	alto	–	–	–
	Cr357	no bajo	no alto	–	–	–
	Cr358	bajo	–	bajo	–	–
	Cr360	bajo	–	no bajo	–	–
BL & no CWA	Cr353		alto		alto	
	Cr357				alto	
	Cr358	bajo		bajo		
	Cr360	bajo		alto		bajo
BL & CWA	Cr353	no bajo	alto		alto	
	Cr357	no bajo	no alto		alto	
	Cr358	bajo		bajo y no alto	no alto	
	Cr360	bajo		alto y no bajo	no alto	bajo
BL & partial CWA	Cr353	no bajo	alto		alto	
	Cr357	no bajo	no alto		alto	
	Cr358	bajo		bajo	no alto	
	Cr360	bajo		alto	no alto	
BL +G & CWA	Cr353	no bajo	alto		alto	
	Cr357	no bajo	no alto		alto	
	Cr358	bajo		bajo	no alto	
	Cr360	bajo		no bajo	no alto	

Tabla 23.1: Resumen de las interpretaciones obtenidas por las diferentes propuestas.

Observando la Tabla 23.1, se puede concluir que la propuesta Best global concept and Close-World Assumption, produce conceptos menos complejos a diferencia de la propuesta *Best local concept and Close-World Assumption* que produce conceptos de una complejidad mayor, lo que se hace evidente en el caso de la temperatura, donde nos encontramos frente a una redundancia conceptual al asociar el concepto (bajo y no alto), lo cual no aporta nada a

la interpretación de las clases, con lo cual se descarta esta propuesta que, desde un punto de vista lógico se comporta peor que las otras..

La propuesta *Best local concept and partial Close-World Assumption* y la propuesta *Best local-global concept and Close-World Assumption* son simplificaciones de *Best local concept and Close-World Assumption* y están a medio camino hacia *Best global concept and Close-World Assumption*, aunque producen conceptos algo más elaborados, incluyendo 2 variables nuevas a la interpretación (Valve-air y FR1-DOTOK) que en *Best global concept and Close-World Assumption* no estaban.

La que se comporta de manera más diferente es *Best local concept and no Close-World Assumption* que caracteriza los extremos de las clases

Si se observa la interpretación de las clases proporcionada por el experto la que más se acerca es *Best local-global concept and Close-World Assumption*, básicamente por la forma en que se comporta la temperatura y a la incorporación de más conceptos debido a la aparición de más variables que en la propuesta *Best global concept and Close-World Assumption* no aparecían.

23.3.2 Análisis cuantitativo

Como ya se ha dicho anteriormente, una vez determinados los conceptos asociados a cada clase, se ha procedido a evaluar cual(es) era(n) satisfecho(s) por cada objeto de \mathcal{I} con tal de evaluar la correcta correspondencia entre la muestra y la conceptualización inducida. En la Tabla 23.2 se muestra un resumen de los resultados de calcular el soporte $Sup(r)$ (ver ecuación (9.1)), la cobertura relativa $CovR(r)$ (ver ecuación (9.5)) y la confianza $p(r)$ (ver ecuación (9.2)) de cada una de las reglas compuestas inducidas para cada clase de la partición final y para las 5 propuestas metodológicas.

En cuanto a las **Inconsistencias** (lo que significa que hay objetos que satisfacen el antecedente de más de una regla) en el caso de la propuesta *Best global concept and Close-World Assumption* y la propuesta *Best local concept and no Close-World Assumption* es evidente que no habrá inconsistencias al evaluar cada regla por separado ya que los intervalos que conforman cada regla compuesta, son disjuntos entre una clase y otra debido a la forma en que se construyen los conceptos, (ver ecuaciones (10.1), (10.3), (10.2), (10.4), (10.5), (10.6) y (10.7), (10.8) para la propuesta *Best global concept and Close-World Assumption* y ver ecuaciones (10.9) y (10.10) para la propuesta *Best local concept and no Close-World Assumption* que lo que hace esta última es caracterizar los extremos de las clases).

En el caso de la propuesta *Best local concept and Close-World Assumption* habrá inconsistencias, por esto algunos objetos satisfacen simultáneamente varias reglas con distinta parte derecha. Esto ocurre debido a que los intervalos que conforman cada regla compuesta no son disjuntos entre cada clase, es decir cada una de las reglas compuestas asociadas a cada clase presentan intersecciones, esto es debido a la forma en que se construyen los conceptos, ver ecuaciones (10.15) y (10.16). Esto mismo ocurre con la propuesta *Best local concept and partial Close-World Assumption*, ver ecuaciones (10.23) y (10.24) utilizadas para la construcción del concepto y la propuesta *Best local-global concept and Close-World Assumption*, ver ecuaciones (11.3), (11.4), (11.5), (11.6) y (11.1).

En la Tabla 23.2 se puede observar que el número de objetos asignados por las reglas a cada clase se puede obtener con $\#\{i \in A_C^\xi\}$.

En cuanto al **Soporte**, se tiene que se obtiene dividiendo $\#\{i \in A_C^\xi\}$ entre el total de objetos de la base de datos (365).

Si se considera la suma de los soportes de todas la reglas cuando éste es menor que el 100% significa que hay objetos de la base de datos sin asignar, es decir, que no satisface ninguna regla y cuando es mayor hay inconsistencias.

Met.	Ruler	Concec.	$\#\{i \in A_C^\xi\}$	$\#\{A_C^\xi \cap i \in C\}$	n_c	$p(r)$	$Sup(r)$	$CovR(r)$
BG & CWA	r_{Cr353}	$Cr353$	220	99	122	45,00%	60,27%	81,15%
	r_{Cr357}	$Cr357$	98	46	50	46,94%	26,85%	92,00%
	r_{Cr358}	$Cr358$	6	6	93	100%	1,64%	6,45%
	r_{Cr360}	$Cr360$	38	33	100	86,84%	10,41%	33,00%
$(\bar{p}(\mathbb{R}))$						69,70%		
<i>Suma</i>			362	184	365		99,18%	
$(CovG_{global}(\mathbb{R}))$								50,4%
BL & no CWA	r_{Cr353}	$Cr353$	27	27	122	100%	7,40%	22,13%
	r_{Cr357}	$Cr357$	1	1	50	100%	0,27%	2,00%
	r_{Cr358}	$Cr358$	6	6	93	100%	1,64%	6,45%
	r_{Cr360}	$Cr360$	3	3	100	100%	0,82%	3,00%
$(\bar{p}(\mathbb{R}))$						100%		
<i>Suma</i>			37	37	365		10,14%	
$(CovG_{global}(\mathbb{R}))$								10,1%
BL & CWA	r_{Cr353}	$Cr353$	309	122	122	39,48%	84,66%	100,00%
	r_{Cr357}	$Cr357$	101	48	50	47,52%	27,67%	96,00%
	r_{Cr358}	$Cr358$	299	91	93	30,43%	81,92%	97,85%
	r_{Cr360}	$Cr360$	220	99	100	45,00%	60,27%	99,00%
$(\bar{p}(\mathbb{R}))$						40,61%		
<i>Suma</i>			929*	360	365		254,52%**	
$(CovG_{global}(\mathbb{R}))$								98,6%
BL & parc. CWA	r_{Cr353}	$Cr353$	309	122	122	39,48%	84,66%	100,00%
	r_{Cr357}	$Cr357$	101	48	50	47,52%	27,67%	96,00%
	r_{Cr358}	$Cr358$	114	62	93	54,39%	31,23%	66,67%
	r_{Cr360}	$Cr360$	35	29	100	82,86%	9,59%	29,00%
$(\bar{p}(\mathbb{R}))$						56,06%		
<i>Suma</i>			559*	261	365		153,15%**	
$(CovG_{global}(\mathbb{R}))$								71,5%
BL +G & CWA	r_{Cr353}	$Cr353$	309	122	122	39,48%	84,66%	100,00%
	r_{Cr357}	$Cr357$	101	48	50	47,52%	27,67%	96,00%
	r_{Cr358}	$Cr358$	114	62	93	54,39%	31,23%	66,67%
	r_{Cr360}	$Cr360$	220	99	100	45,00%	60,27%	99,00%
$(\bar{p}(\mathbb{R}))$						46,60%		
<i>Suma</i>			744*	331	365		203,84%**	
$(CovG_{global}(\mathbb{R}))$								90,7%

Tabla 23.2: Resumen resultados.

Para la propuesta Best global concept and Close-World Assumption en la Tabla 23.2 se puede observar que se asignan 362 objetos de 365, es decir que sólo hay 3 objetos que no han sido asignados y se tiene un promedio del soporte cercano al 100%. Como esta propuesta no genera reglas inconsistentes podemos decir que estamos cerca de una base de conocimiento inducida (o sistema de reglas) completa para la base de datos observada. En el caso de la propuesta *Best local concept and no Close-World Assumption* se puede observar que el promedio del soporte es muy bajo, lo cual implica que el número de objetos sin asignar es muy alto, hay ($365 - 37 = 328$) objetos que no han sido asignados y que por lo tanto no satisfacen

el antecedente de ninguna regla compuesta. Esto significa que El sistema de reglas inducido aporta poca información respecto al dominio.

A diferencia de los 2 casos anteriores, en las propuestas *Best local concept and Close-World Assumption*, *Best local concept and partial Close-World Assumption* y *Best local-global concept and Close-World Assumption* la mayor parte de los objetos presentan inconsistencias, ya que todas estas tienen la suma total de los soportes medios mayores que el 100% (la que menos *Best local concept and partial Close-World Assumption*).

En cuanto a la **Confianza**(columna $p(r)$), si se observa la Tabla 23.2, la confianza se obtiene dividiendo las celdas de la columna $\#\{i \in A_C^\xi \cap i \in C\}$ por las de la columna $\#\{i \in A_C^\xi\}$.

En la propuesta Best global concept and Close-World Assumption hay 3 clases en donde $\#\{i \in A_C^\xi\} > \#\{i \in A_C^\xi \cap i \in C\}$ por tanto se puede concluir que hay objetos mal asignados, es decir objetos que no satisfacen el consecuente de la regla, pero que satisfacen el antecedente de ésta (esta característica es aplicable a todas las propuestas). Este valor no se considera en el soporte ni en la cobertura relativa, pero queda de manifiesto en la confianza ya que cuando se obtienen confianzas del 100% significa que todos los objetos que satisfacen simultáneamente el antecedente y el consecuente de la regla. Con lo cual se puede concluir, en este caso, que el número de objetos correctamente asignados por clase viene dado por $\#\{i \in A_C^\xi \cap i \in C\}$ y el número de objetos mal asignados por clases, en este caso, se puede calcular $\#\{i \in A_C^\xi\} - \#\{i \in A_C^\xi \cap i \in C\}$. En el caso concreto de la propuesta Best global concept and Close-World Assumption hay 184 objetos que satisfacen algunas de las reglas simultáneamente en su antecedente y consecuente, es decir, que usando El sistema de reglas inducido, se asignarían correctamente a la clase correspondiente. Esto representa un 50,4% del total de la muestra analizada.

Representamos en estas tablas las confianza media de cada sistema de reglas final, de acuerdo a lo que hacen otros autores que manejan este tipo de conceptos (Liu 2000). En este caso la media de las confianzas es la segunda mejore de las 5 propuestas.

En la propuesta *Best local concept and no Close-World Assumption* se cumple en todas las clases $\#\{i \in A_C^\xi\} = \#\{i \in A_C^\xi \cap i \in C\}$, lo cual era ya predecible dada la forma cómo se construyen los conceptos, se puede concluir que no hay objetos mal asignados, es decir, todos los objetos satisfacen el antecedente y el consecuente de la reglas. En este caso, el número de objetos correctamente asignados viene dado por $\#\{i \in A_C^\xi \cap i \in C\}$ o por $\#\{i \in A_C^\xi\}$ y el porcentaje total de objetos correctamente asignados se puede obtener dividiendo 37 entre 365, (10,14%). En este caso las confianzas en promedio son las más altas aunque se tienen soportes muy bajos, como ya se dijo anteriormente.

En la propuesta *Best local concept and Close-World Assumption*, en las 4 clases ocurre que $\#\{i \in A_C^\xi\} > \#\{i \in A_C^\xi \cap i \in C\}$ con lo cual se puede concluir que hay objetos mal asignados, es decir objetos que no satisfacen el consecuente de la regla, pero que satisfacen el antecedente de ésta. En este caso el número de objetos correctamente asignados por clase viene dado por $\#\{i \in A_C^\xi \cap i \in C\}$, pero al haber inconsistencias, ya no es cierto que el número de objetos mal asignados por clases se pueda calcular haciendo $\#\{i \in A_C^\xi\} - \#\{i \in A_C^\xi \cap i \in C\}$, como en el caso de la propuesta Best global concept and Close-World Assumption. Las confianzas son algo más bajas que en la propuesta Best global concept and Close-World Assumption debido a que hay muchos objetos que satisfacen el antecedente de cada regla y ello es debido a las inconsistencias.

Algo similar a lo analizado en el párrafo anterior, ocurre en la propuesta *Best local concept and partial Close-World Assumption*, aunque con confianzas algo más altas en los casos de las clases Cr358 y Cr360, y en la propuesta *Best local-global concept and Close-World Assumption* aunque aquí sólo aumenta la confianza en la clase Cr358.

En general ocurre que a mayor confianza menor soporte y la solución mejor será la más equilibrada.

En cuanto a la **Cobertura relativa**, si se observa la Tabla 23.2, la cobertura relativa se obtiene dividiendo cada celda de la columna $\#\{i \in A_C^\xi \cap i \in C\}$ entre la correspondiente celda de la columna $\#\{C\}$.

En estas tablas, en lugar de representar la cobertura relativa media de cada sistema de reglas final, como se hace con la confianza, representamos la media de la cobertura relativa ponderada por el tamaño de cada clase, que coincide con el porcentaje global de objetos de la base de datos que se asignan correctamente, dando una idea mas ajustada de la calidad predictiva del sistema de reglas inducido.

La relación entre cobertura relativa y confianza es también inversamente proporcional, a medida que se pierde confianza se gana cobertura relativa.

En el caso de la propuesta *Best local concept and partial Close-World Assumption* y la propuesta *Best local-global concept and Close-World Assumption* la relación entre cobertura relativa y confianza es similar a la propuesta *Best global concept and Close-World Assumption* y a la propuesta *Best local concept and Close-World Assumption* aunque se puede considerar mas equilibrada que esta última, mas bien parecida a caso de *Best global concept and Close-World Assumption*.

23.4 Conclusiones de la aplicación

- En cuanto a la interpretación de clases.
 1. Si se observa la interpretación de las clases proporcionada por el experto la que más se acerca es *Best local-global concept and Close-World Assumption*, básicamente por la forma en que se comporta la temperatura y a la incorporación de más conceptos dada la aparición de más variables que a diferencia de la propuesta original no existían.
 2. La propuesta *Best local concept and partial Close-World Assumption* y la propuesta *Best local-global concept and Close-World Assumption* son simplificaciones de *Best local concept and Close-World Assumption* y están a medio camino hacia CCCS, aunque producen conceptos algo mas elaborados, incluyendo 2 variables nuevas a la interpretación (Valve-air y FR1-DOTOK) que en *Best global concept and Close-World Assumption* no estaban.
 3. La que se comporta de manera más diferente es *Best local concept and no Close-World Assumption* que caracteriza los extremos de las clases, por lo tanto el soporte de las reglas compuestas asociadas a cada clase es mínimo, a pesar que las confianzas son máximas
- En cuanto a las medidas de calidad de las bases de conocimiento inducidas según las distintas propuestas.
 1. Si se considera el índice de calidad $CovG_{local}$ que representa el porcentaje de objetos de \mathcal{I} correctamente asignados por la base de conocimientos final, la propuesta

Best local concept and Close-World Assumption tiene el valor más alto de las 5 propuestas (98,6%), pero como genera conceptos que presentan redundancia conceptual (como es el caso de la variable *Temp – ww* en;

- Cr358: “(*TN – influent* es bajo o *Valve – air* no es alto) y (*Temp – ww* es bajo o *Temp – ww* no es alto)”
- Cr360: “(*TN – influent* es bajo o *Valve – air* no es alto) y (*Temp – ww* no es bajo o *Temp – ww* es alto)”), lo cual no aporta nada a la interpretación de las clases, se descarta esta propuesta que, desde un punto de vista lógico se comporta peor que la *Best local-global concept and Close-World Assumption*.

2. La propuesta *Best local-global concept and Close-World Assumption* es la que tiene la segunda mejor cobertura global ($CovG_{lobal} = 90,7\%$) lo que confirma la decisión que es la más acertada, tanto cualitativamente como en el valor de este índice de calidad.
3. La suma total de los soportes por clase corresponde al porcentaje de objetos asignados, cuando este es menor que el 100% significa que hay objetos sin asignar y cuando es mayor hay inconsistencias.
4. Cuando no hay inconsistencias ($Sup(r) \leq 100\%$) el número de objetos mal asignados por clase se puede calcular haciendo $\#\{i \in A_C^\xi\} - \#\{i \in A_C^\xi \cap i \in C\}$
5. Si $\#\{i \in A_C^\xi\} > \#\{i \in A_C^\xi \cap i \in C\}$ se puede concluir que hay objetos mal asignados. Este valor queda de manifiesto en la confianza ya que cuando se obtienen confianzas del 100% significa que todos los objetos que satisfacen el antecedente de la regla, también satisfacen simultáneamente el antecedente y el consecuente de la regla y por lo tanto pertenecen a la clase a la que la regla asigna el objeto.
6. La relación entre Soporte ($Sup(r)$) y Confianza ($p(r)$) es inversamente proporcional, a medida que se pierde confianza se gana soporte.
7. La relación entre Cobertura relativa ($CovR(r)$) y Confianza ($p(r)$) es también inversamente proporcional, a medida que se pierde confianza se gana cobertura relativa.

23.5 Resumen

En general si se observa la interpretación de las clases proporcionada por el experto, la que más se acerca es *Best local-global concept and Close-World Assumption*, básicamente por la forma en que se comporta la temperatura y por la incorporación de más conceptos dada la aparición de más variables que en deferencia de la propuesta original no aparecían y por el valor de la cobertura global ($CovG_{lobal}(\mathbb{R}) = 90,7\%$). Si el objetivo primordial es favorecer la riqueza conceptual de la interpretación o grado de interpretabilidad (y/o utilidad) de las clases formadas, la propuesta *Best local-global concept and Close-World Assumption* es la mejor, pero si se considera la capacidad predictiva (permitir para un nuevo objeto (día), predecir la clase (situación típica de la planta) que le corresponde y generar las caracterización e interpretación conceptual correspondiente a esa clase), el hecho que existan inconsistencias genera conflictos, lo que ha se ha estudiado previamente en (Pérez-Bonilla and Gibert 2007a) (para más detalles ver el reporte de investigación (Pérez-Bonilla and Gibert 2008a)), donde se proponen diversos métodos de resolución de conflictos.

Parte IV

Conclusiones y líneas futuras

Capítulo 24

Conclusiones, trabajo futuro y líneas abiertas

Jo

*Jo i una altra persona som un parell,
Molts com nosaltres formen un nombre de gent,
Gran nombre de persones son una multitud,
Una multitud de gent un poble,
Un poble una nació,
Moltes nacions un continent,
Els continents la humanitat.*

Joan Brossa

Esta tesis presenta una metodología de generación de interpretaciones conceptuales asociadas a una partición procedente de una clasificación jerárquica.

La interpretación de clasificaciones es un problema abierto cuya complejidad crece enormemente en contextos de Data Mining, donde el número de clases y el de variables es grande. Éste, de hecho, es uno de los problemas objeto del aprendizaje automático, del cual ID3 (Quinlan 1990) y sus sucesores son exponentes característicos. Si bien existen métodos bien conocidos de inducción de conceptos a partir de clases, éstos están más orientados a optimizar el poder predictivo de la solución más que a optimizar el poder descriptivo. Ello hace que raramente un experto que se enfrenta a un problema de clustering recurra a este tipo de métodos para interpretar los resultados de las clases obtenidas en la profundidad necesaria para diseñar planes de acción o de distribución de recursos asociados a cada clase. Así, la interpretación se aborda habitualmente de forma bastante rudimentaria y no sistemática a base de analizar los listados de los paquetes estadísticos manualmente. Este proceso se torna particularmente complejo cuando aumenta el número de clases o el de variables.

Esta tesis pretende contribuir a la mejora de este proceso, fundamental para comprender el significado de las clases obtenidas y dar soporte efectivo a la posterior toma de decisiones.

La alternativa que parece más prometedora para resolver estas limitaciones es aligerar al experto de este trabajo, mediante el desarrollo de técnicas que a partir de la evidencia empírica, identifiquen las variables más relevantes y formulen conceptos que expresen las particularidades de cada clase y se expresen en una forma de representación conceptual generable automáticamente y directamente comprensible para el experto.

Incorporar procedimientos que trasladen los resultados del análisis (en este caso del *clustering*) a una representación explícita del conocimiento obtenido, se sitúa en la línea de lo que Fayyad propone para los sistemas de KDD, donde la fase de post-proceso de los resultados para generar conocimiento es casi tan importante como el análisis en si mismo. Quizás por su

naturaleza más semántica la generación automática de interpretaciones de una clasificación no se ha tratado formalmente desde el ámbito estadístico, aunque resolverlo es fundamental.

En esta tesis se propone una solución aproximada al problema planteado de construir un sistema de conceptos $\mathcal{A}_{\mathcal{P}_\xi} = \{A_1, A_2, A_3, \dots, A_\xi\}$ que describen las clases de tal forma que:

- $A, A' \in \mathcal{A}_{\mathcal{P}_\xi} \Rightarrow A \neq A'$
- $\forall i \in \mathcal{I}, \quad A_C(i) = \text{true} , \text{ si } C = C(i, \mathcal{P}_\xi), \quad A_C \in \mathcal{A}_{\mathcal{P}_\xi}$
- $\forall i \in \mathcal{I}, \quad A_C(i) = \text{false}, \text{ si } C \neq C(i, \mathcal{P}_\xi), \quad A_C \in \mathcal{A}_{\mathcal{P}_\xi}$

Teniendo en cuenta que existirá cierta incerteza en el modelo, se propone tratar con reglas más genéricas de la forma $r : A_C(i) \xrightarrow{p} C$ donde $p \in [0, 1]$ es la probabilidad con que se cumple r . De este modo las reglas incorporan incerteza bajo una aproximación probabilística.

- Aporta una sistematización al proceso de interpretación de clases procedentes de un cluster jerárquico y supone un avance significativo respecto al estado actual en que la interpretación se realiza de forma manual y más o menos artesanal.
- Asimismo, contribuye a sistematizar y objetivar los mecanismos de interpretación que usan los expertos humanos.
- Los resultados que genera la metodología permiten que el experto pueda comprender más fácilmente las características principales de la clasificación obtenida, debido a que ésta genera conocimiento explícito directamente a partir de las clases.

La metodología que se propone trata de aproximar en un modelo formal el proceso natural que sigue un experto en su fase de interpretación de resultados.

En la propuesta metodológica presentada en esta tesis, se identifican conceptos que interpretan a cada clase y, bajo este punto de vista, ello representa una alternativa a los objetos simbólicos. Estos conceptos son mutuamente excluyentes y se reducen a un conjunto mínimo de variables relevantes en la descripción de cada clase facilitando el proceso de conceptualización del experto. Es decir, la comprensión última del porqué de cada clase, puesto que la interpretación que se propone focaliza solamente en las variables relativas a los rasgos distintivos de la clase y prescinde de aquéllo que, siendo común a otras clases, no contribuye a entender la génesis de la clase y por tanto a identificar la utilidad semántica de la misma.

La propuesta además es aplicable a cualquier partición de objetos, provenga de donde provenga, ya sea fruto de un proceso de clustering automático, como de una propuesta del experto, como del consenso de un grupo expertos y sea cual sea el modelo de formalización utilizado en la representación de datos. La propuesta permite de forma simple y eficaz obtener buenas descripciones evitando el análisis de interacciones complejas, que podrían complicar la complejidad algorítmica.

No obstante, los métodos análisis de datos simbólicos se muestran poco eficaces en matrices muy grandes por cuestiones de complejidad y en este sentido nuestra propuesta de aprovechar la jerarquía subyacente garantiza la eficiencia del proceso.

Así pues las descripciones conceptuales que proponemos, responden también a una descripción intensional de las clases, al igual que los objetos simbólicos, pero son más compactas y mutuamente excluyentes facilitando la conceptualización del experto, además de resultar un cálculo más eficiente ante grandes bases de datos.

Basándonos en trabajos previos que intentan ya resolver partes de este problema se elabora una nueva propuesta más completa, que supere todas las limitaciones observadas en las

propuestas anteriores y consolide una metodología de generación automática de interpretaciones de clases, que además sirva para dar apoyo a la construcción de sistemas inteligentes de soporte a la toma de decisiones.

Si bien la metodología que se propone es general, se ha centrado la aplicación a Estaciones depuradoras de aguas residuales (EDAR) por ser éste uno de los dominios donde las aproximaciones clásicas funcionan peor y porque se encuadran en una de las líneas marco de investigación que se desarrolla en el grupo.

Desde un punto de vista teórico, el interés de esta tesis ha sido presentar una propuesta metodológica híbrida que combine herramientas y técnicas de Estadística e Inteligencia Artificial en forma cooperativa, tal que, a partir de las variables que describen los objetos pertenecientes a cierto dominio, podamos caracterizar las situaciones características (clases resultantes) que se pueden encontrar en él, contribuyendo así al proceso de interpretación conceptual automática de clases procedentes de un cluster.

En términos generales podemos decir que:

- Es claro que hoy, las nuevas tecnologías aumentan significativamente nuestra capacidad de producir, colecciónar y almacenar datos. Enormes cantidades de datos están disponibles para ser analizados y extraer conocimiento en corto tiempo.
- Cuando se trata de dominios poco estructurados, obtener conocimiento útil de conjuntos de datos es una tarea muy difícil y el tamaño de la base de datos se hace secundario. En estos dominios las técnicas clásicas no ofrecen buen comportamiento por si solas por las características inherentes a estos dominios.
- La situación se agrava si el número de clases a interpretar y el número de variables que intervienen es además alto.
- Durante la década pasada, en una gran variedad de dominios de aplicación, surge reiteradamente en la comunidad científica la constatación de que la multidisciplinariedad y la hibridación de técnicas ha de proporcionar mejores aproximaciones a las realidades complejas como los dominios poco estructurados. La combinación de técnicas de análisis de datos (ej. clustering), aprendizaje inductivo (ej. sistemas basados en conocimiento), administración de base de datos y representación gráfica multidimensional, producen beneficios en esta dirección. *Knowledge Discovery and Database (KDD)* y *AI&Stats* se orientan en esa línea.
- Existen diversos software informáticos que ofrecen herramientas para hacer una primera aproximación a algunas de las situaciones mencionadas (ej. Clementine, Intelligent Manager, SPAD, SPSS, WEKA entre otras son algunas de los más famosas hoy en día). Estos softwares presentan principalmente un compendio de técnicas existentes, permitiendo comparación de resultados y la selección del mejor método en cada caso, pero no guían al usuario sobre cuál es la mejor forma de combinar las distintas herramientas para obtener una buena y correcta interpretación de un conjunto de clases.
- La propuesta metodológica que se presenta sigue también éste enfoque transversal y multidisciplinar combinando elementos de la inducción de conceptos en Inteligencia Artificial, lógica proposicional y teoría de probabilidad y contribuye a la concepción genérica de sistema de *KDD* propuesto por Fayyad (Fayyad, Piatetsky-Shapiro, and Smyth 1996), que debe incluir módulos de soporte a la definición del problema (que incluya el conocimiento), recolección de datos, depuración y preprocessamiento, reducción de datos, selección de la técnica de data mining, interpretación y producción del conocimiento descubierto *a posteriori*.

- Que la clasificación obtenida por métodos automáticos tenga una interpretación clara está relacionado con la *utilidad* de una clasificación, que actualmente se utiliza también como criterio de validación y se puede utilizar para decidir si es correcta o no (Gibert, Hernández, and Cortés 1996); evaluarla requiere un mecanismo a posteriori de *comprensión* del significado de las clases. Así esta tesis contribuye también a objetivar los procedimientos de validación de resultados.

El desarrollo de la metodología que se presenta ha requerido a su vez del desarrollo de un marco formal de apoyo sobre el que se construye posteriormente la propuesta final:

- Se ha realizado una tipificación de los sistemas de reglas que se pueden inducir de una partición en función del tipo de reglas que contienen (Reglas no efectivas, Reglas efectivas y Reglas seguras). En todos los casos existen los equivalentes de los sistemas globales que se construyen como unión directa de los inducidos para cada variable porque en este contexto se cumple que las reglas tienen siempre antecedentes simples correspondientes a una sola variable y no solapan con las generadas por otras variables.
- Se han analizado los tipos de valores que puede tomar una variable según el tipo de regla que genera. Sus valores ya sean modalidades o intervalos pueden ser de 5 tipos básicos según su poder caracterizador (Valor *totalmente caracterizador*, Valor *parcialmente caracterizador*, Valor *caracterizador no propio*, Valor *genérico* y Valor *vacio*).
- Si X_K ha sido discretizada por el *BbD* hemos podido formalizar qué tipo de reglas generan sus intervalos en función de que se corresponda con un Valor *totalmente caracterizador*, Valor *parcialmente caracterizador*, Valor *caracterizador no propio*, Valor *genérico* o Valor *vacio*.
- Se ha realizado una propuesta de criterios de calidad de los sistemas de reglas en general, a partir de sus cardinales y las propiedades que de ellos se deducen y se concluye que la mejor situación que puede ocurrir es que el número de reglas seguras tienda al número de clases de la partición, que es lo mismo que decir que el cociente entre el número de reglas seguras y el número de reglas del sistema de reglas completo tienda a 1. Indirectamente eso significa que el mejor caso es aquel en que hay tantos valores propios como clases y de ahí se ha podido formular el criterio:

La mejor propuesta maximiza el número de valores propios y seguidamente el de valores vacíos si hay empate de valores propios.

Estos criterios permiten establecer comparaciones objetivas entre varios sistemas de reglas.

- A continuación se analiza la relación entre los 3 tipos de valores (intervalos) de una variable y los tipos de reglas que genera, primero para una situación genérica, después para el caso particular en que X_k ha sido discretizada por el *BbD* y finalmente para la situación más concreta aún y relevante en esta tesis que X_k venga de *BbD* para particiones binarias.
- Con los criterios definidos para comparar sistemas de reglas se han observado anomalías en los intervalos obtenidos por el *BbD* y ello ha conducido a una revisión del método. A partir de un análisis por casos se detectan situaciones anómalas, básicamente por dos motivos:

- Se hacía una división artificial del recorrido de la clase.
- Se generaban reglas no seguras sólo porque se incluía el extremo de la otra clase como límite de un intervalo.
- Se generaban reglas seguras donde deberían ser vacías porque se aislaban un límite de clase en un intervalo separado de un sólo punto en realidad conexo a otro intervalo colindante.

Se aplican correcciones para evitar estas anomalías. Se observa que la propuesta resultante contempla únicamente 2 tipos de patrones para construir \mathcal{D}^k a partir de Z^k , ellos son:

Si $(M_{C_j}^k < m_{C_i}^k) \text{ o } (M_{C_i}^k < m_{C_j}^k)$ entonces generar un \mathcal{D}^k centro abierto:

$$\begin{aligned} I_1^{k,\xi} &= [z_1^k, z_2^k] \\ I_2^{k,\xi} &= (z_2^k, z_3^k) \\ I_3^{k,\xi} &= [z_3^k, z_4^k] \end{aligned}$$

sino generar un \mathcal{D}^k centro cerrado:

$$\begin{aligned} I_1^{k,\xi} &= [z_1^k, z_2^k] \\ I_2^{k,\xi} &= [z_2^k, z_3^k] \\ I_3^{k,\xi} &= (z_3^k, z_4^k) \end{aligned}$$

Siendo $m_C^k = \min X_K | C = \min_{i \in C} \{x_{ik}\}$ y $M_C^k = \max X_K | C = \max_{i \in C} \{x_{ik}\}$.

- Las consecuencias inmediatas en lo que se refiere a la estructura de los sistemas de reglas son que utilizando el *BbD* revisado:
 - Cuando se intercalan los extremos de la variable en las 2 clases (que es la mayor parte de las veces), en la corrección que se propone:
 - * *se gana un valor propio y aumenta el número de reglas seguras lo que facilitaría encontrar variables totalmente caracterizadoras.*
 - Cuando los mínimos de ambas clases coincidan, en la corrección que se propone:
 - * *se pierde un valor genérico a favor de un vacío, el número de reglas seguras se mantiene y el número de reglas no efectivas aumenta lo que implica una mayor compactación de X_k (en realidad prescindibles).*
 - Cuando las dos clases no solapan. En esta situación:
 - * *se fusionan 2 valores propios para ganar uno vacío y aunque se pierde una regla segura se tiene un sistema de reglas equivalente más compacto.*
- A continuación se determina que los sistemas de reglas se evaluarán en términos de Soporte total ($Supt(\mathcal{R})$), Certeza o confianza media ($\bar{p}(\mathcal{R})$), Cobertura global ($CovGlobal(\mathcal{R})$).
 - En general serán mejores los sistemas de mayor soporte total, mayor cobertura global y mejor certeza media aunque casi nunca se dará este tipo de situaciones ideales.
 - Si la suma total de los soportes por clase es mayor del 100% hay conflictos de interpretación y los conceptos de varias clases no son disjuntos.
 - La suma total de los soportes por clase corresponde al porcentaje de objetos asignados, cuando este es menor que el 100% significa que hay objetos sin asignar y cuando es mayor hay inconsistencias.

- La relación entre Soporte $Supt(\mathcal{R})$ y Confianza $\bar{p}(\mathcal{R})$ es inversamente proporcional, a medida que se pierde confianza se gana soporte.
- La relación entre Cobertura relativa $CovGlobal(\mathcal{R})$ y Confianza $\bar{p}(\mathcal{R})$ es también inversamente proporcional, a medida que se pierde confianza se gana cobertura relativa.
- Se definen heurísticos que hallen los mejores compromisos entre estos parámetros. Se proponen 5 formas de combinar el conocimiento con la hipótesis de mundo cerrado (closed world assumption) *CWA*, de lo que se ha podido concluir:
 - El principio de la parsimonia tan razonable en contextos de modelización con fines predictivos no produce soluciones cercanas a las descripciones del experto y no genera la mejor aproximación metodológica al problema que se quiere resolver.
 - Si se considera la capacidad predictiva (permitir para un nuevo objeto (día), predecir la clase (situación típica de la planta) que le corresponde y generar las caracterización e interpretación conceptual correspondiente a esa clase), el hecho que existan inconsistencias genera conflictos, lo que ha se ha estudiado previamente en (Pérez-Bonilla and Gibert 2007a), en donde se ha probado distintos algoritmos (asignación por clase más votada, por probabilidad máxima, suma de probabilidades y por clase más votada y probabilidad máxima conjuntamente) que, usando las bases de conocimiento inducidas para cada clase, permiten asignar una etiqueta de clase a cada objeto, concluyendo que el mejor algoritmo de asignación para superar las inconsistencias es el de clase más votada y probabilidad máxima conjuntamente considerando en número de objetos correctamente asignados.
 - Como el objetivo primordial es favorecer la riqueza conceptual de la interpretación o grado de interpretabilidad (y/o utilidad) de las clases formadas los criterios objetivos por si solos no apuntan a la descripción que más se acerca a la que manualmente hace el experto.
- Existen diferencias importantes entre lo que deberían construir los criterios de optimabilidad cuando el objetivo es descriptivo o predictivo y podríamos decir que mientras el concepto óptimo para predicción es minimal, el óptimo para descripción tiende a ser maximal, si se toma como referencia el proceso cognitivo que sigue un experto humano para llegar a identificar un conjunto de objetos con una entidad semántica de su corpus de doctrina. Otras investigaciones que se están desarrollando en paralelo en colaboración con el Instituto Guttman (Gibert and Tormos 2008) permiten constatar que éste es, en efecto, el enfoque más adecuado cuando se pretende aportar conocimiento útil para responder a aspectos desconocidos de un dominio poco estructurado que permitan posterior toma de decisiones informada. Por otro lado en (Gibert, Pérez-Bonilla, and Rodriguez 2006) se constata que para fines descriptivos ésta es la mejor aproximación en contraste con otras más clásicas como regresión logística, árboles de decisión o análisis discriminante.

La propuesta metodológica que se presenta se ha aplicado a 2 casos para obtener la interpretación de 2 clasificaciones reales. Las dos proceden de Estaciones Depuradoras de Aguas Residuales, pero de características algo diferentes. Para ambas se disponía de una interpretación previa de clases proporcionada por los expertos.

En la planta catalana, se ha aplicado la propuesta metodológica comparando los criterios Best local *Best local-global concept and Close-World Assumption (BL+G & CWA)* y *Best Global concept and Close-World Assumption (BG & CWA)* a datos de una planta depuradora

de aguas residuales de la costa catalana. La propuesta *Best local-global concept and Close-World Assumption (BL+G &CWA)* produce conceptos algo más elaborados que *Best Global concept and Close-World Assumption (BG &CWA)*. Ésa es también la que más se acerca a la interpretación de las clases proporcionada por el experto, básicamente por la incorporación de más variables en la descriptiva final, acercándose más a lo que el experto hace cuando describe y que, como se ha dicho, se aleja del clásico principio de la parsimonia común en modelización con fines predictivos. *BL+G &CWA* también optimiza la cobertura global ($CovG_{global}(\mathbb{R}) = 59\%$) lo que confirma que es la más acertada, también en cuanto a los criterios objetivos de calidad de las bases de conocimiento inducidas.

En una segunda aplicación se han utilizado datos de una planta situada en Eslovenia. Se ha aplicado la propuesta metodológica con los 5 métodos de integración del conocimiento propuestos en el capítulo §10 para generar la interpretaciones. Al igual que ocurre en el primer caso de estudio, *Best local-global concept and Close-World Assumption (BL+G &CWA)* es la propuesta que más se acerca a la interpretación de las clases proporcionada por el experto, siendo la que, también en este caso, presenta la mejor cobertura global $CovG_{global}(\mathbb{R}(\mathcal{P}_4)) = 90,7\%$ después de la *Best local concept and Close-World Assumption (BL &CWA)*. Aunque esta última presenta mejor cobertura global ($CovG_{global}(\mathbb{R})$) en términos absolutos, presenta peor comportamiento porque genera conceptos que presentan redundancia conceptual y por ello se descarta. Así también en esta aplicación *BL+G &CWA* aparece como la mejor opción cualitativa y cuantitativamente.

De las 5 propuestas, la que se comporta de manera más diferente es *BL &noCWA* que caracteriza los extremos de las clases, por lo tanto el soporte de las reglas compuestas asociadas a cada clase es mínimo, a pesar de que las confianzas son máximas. Las otras 2 propuestas *Best local concept and partial Close-World Assumption (BL &partial-CWA)* y la propuesta *Best local concept and no Close-World Assumption (BL &noCWA)* son simplificaciones de *BL &CWA* y están a medio camino hacia *BG &CWA*, aunque producen conceptos algo más elaborados que esta última, pero siempre en inferior calidad que la *BL+G &CWA*.

La conclusión general parece ser que restringiendo la interpretación a bastantes reglas seguras, aunque tengan soportes bajos identifican partes complementarias del dominio y cuando se toman en cuenta todas las variables, se acaba teniendo una buena descripción de cobertura alta y con reglas seguras.

En nuestra propuesta se utilizan las variables seleccionadas por el experto, no consideramos la correlación entre ellas previo a la clasificación, pues nos interesa extraer todo el conocimiento posible con las variables que el experto promociona. Y es ésta, una de las contribuciones de la propuesta. Sorprendentemente, se identifican conceptos que representan muy bien a cada clase a base de explorar distribuciones condicionadas univariantes.

Esto supone una gran ventaja para el análisis de ISD donde existen interacciones de orden superior entre las variables muy difíciles de modelar. La propuesta permite de forma simple y eficaz obtener buenas descripciones evitando el análisis de interacciones complejas.

El punto clave está en la discretización basada en boxplots que permite reducir a un simple cálculo de extremos locales el análisis de interacciones entre las clases para una cierta variable numérica. La evolución de los rangos de una variable en un conjunto de clases, los intervalos donde esas clases solapan, el subconjunto de clases que solapan en cada intervalo es, en origen, un problema de complejidad combinatoria.

El Boxplot based Discretization, lo resuelve a coste mínimo y genera discretizaciones con máxima asociación a la partición a explicar y donde en cada intervalo en los que se discretiza la variable el conjunto de clases que solapa se mantiene constante. Ello facilita enormemente la inducción de conceptos consistentes a partir de dichas clases.

Trabajo futuro. Hasta aquí se ha construido operativamente una solución aproximada al problema inicial de la forma:

$$\mathcal{A}_{\mathcal{P}_\xi} = \{C : \mathcal{A}_C \quad \forall C \in \mathcal{P}_\xi\}$$

A partir de una construcción jerárquica de,

$$\mathbb{R}(\mathcal{P}_\xi) = \{r \quad tq \quad r : A \xrightarrow{p(r)} C \quad \forall C \in \mathcal{P}_\xi\}$$

Se está ya trabajando en la definición formal de la solución que se propone y en la demostración formal de sus propiedades, entre lo que también se considera:

- La definición de un nuevo índice de desempeño de la metodología que permita cuantificar la calidad de la interpretación en relación a la que realiza el experto. Esta tendrá que tener en cuenta la similitud, en términos objetivos, entre las variables que aparecen en la interpretación y sus valores.
- La implementación computacional de la metodología CCCS en la plataforma *KLASS*. Hasta ahora se ha realizado el seguimiento y la experimentación de forma semiautomática y manualmente.
- Estudiar la posibilidad de usar la metodología *CCCS* en la *clasificación basada en reglas por estado* (Gibert and Rodríguez-Silva 2008).
- Abordar el etiquetado automático de los valores para conceptualizar patrones dinámicos que aparecen en los conceptos finales teniendo en cuenta que se han generado jerárquicamente y por condicionamientos sucesivos por lo que los valores alto o bajo se han de interpretar siempre en relación a la iteración que los generó.
- Valorar si se mejoran los índices de calidad y se enriquece la interpretación al incorporar más variables, considerando las siguientes coberturas relativas mayores y no sólo la "cobertura relativa mayor".
- Introducir variables de diferencia en las medidas repetidas, por ejemplo valorar el comportamiento de (SS-E)-(SS-D) y/o (SSV-E)-(SSV-D), de manera de enriquecer la interpretación, como ya se intentó hacer, de manera artesanal en (Gibert and Roda 2000), y finalmente valorar cuantitativamente si se mejoran los valores de los índices de calidad, esperaríamos que así fuese.
- Estudiar la posibilidad de utilizar la metodología con otro tipo de dominios poco estructurados.

A continuación se presenta una lista de publicaciones con la contribución de cada una.

	Fecha	Conferencia	Reporte de investigación	Proyecto de tesis(DEA) y tesis de Master	Trabajo tutelado		
2004	Junio 2004		[G&PB 04a]	[PB Mth 04]	[G&PB 04b]		
	Noviembre 2004						
	Diciembre 2004						
2005	Mayo 2005		[PB&G 05]	[G&PB 05a] [PB&G 05a]	[PB dea 05]		
	Junio 2005		[G&PB 05b]				
	Septiembre 2005						
	Septiembre 2005		[G&PB 05c]				
	Septiembre 2005						
2006	Abril 2006		[G&PB 06b] [PB&G 06b]	[G&PB 06a]	[G&PB&RS 06]		
	Mayo 2006						
	Julio 2006						
	Noviembre 2006						
2007	Abril 2007		[PB&G&V 07b]	[PB&G 07a] [PB&G 07b] [PB&G 07c]	[PB&G&V 07a]		
	Septiembre 2007		[PB&G 07a]				
	Septiembre 2007						
	Noviembre 2007						
	Diciembre 2007						
2008	Enero 2008		[G&PB 08a] [G&PB 08b]	[PB&G&V 08]	[PB&G&V 08a]		
	Marzo 2008						
	Julio 2008		[G&PB 08c] [G&PB 08d]				
	Septiembre 2008						
	Septiembre 2008						

Publicaciones. En este apartado se listan las publicaciones y trabajos realizados por tipo de publicación (congresos, reportes de investigación y otros trabajos de investigación) y en orden cronológico (comenzando por el más antiguo dentro de cada tipo). Inmediatamente después de la información de cada publicación se encuentra un breve resumen de la contribución que se hace en cada una de estas publicaciones al desarrollo de esta tesis doctoral. Estos trabajos dan soporte a esta tesis doctoral y pretenden mostrar la evolución de la misma.

1. Conferencias y congresos.

Acrónimo: [G&PB 05b]

Fecha: Septiembre 2005

Título: Fuzzy box-plot based induction rules. Towards automatic generation of classes-interpretation.

Autores: Karina Gibert y Alejandra Pérez Bonilla.

In: Procs. Fuzzy Sets In Learning And Data Mining. IV EUSFLAT'2005.

Páginas: 524 a 529

Lugar: Barcelona. España.

ISBN: 84-7643-872-3

Referencia: (Gibert and Pérez-Bonilla 2005b)

Contribución:

- (a) Versión primera de un mecanismo (o método) que genera reglas probabilizadas a partir de 1 variable, 1 partición y la versión original del *Boxplot based discretization* (BbD) y se sitúa en el paradigma difuso.
- (b) Aplicación del método a datos reales procedentes de una planta de tratamiento de aguas residuales (EDAR Cataluña).

Acrónimo:	[G&PB 05c]
Fecha:	Septiembre 2005,
Título:	Ventajas de la estructura jerárquica del <i>clustering</i> en la interpretación automática de clasificaciones.
Autores:	Karina Gibert, Alejandra Pérez Bonilla
In:	III Taller Nacional de Minería de Datos y Aprendizaje. TAMIDA'2005. I Congreso Español de Informática. CEDI'2005.
Páginas:	67 a 76
Volumen:	1
Publicado por:	Thompson-Paraninfo
Lugar:	Granada. España.
ISBN:	84-9732-449-8
Referencia:	(Gibert and Pérez-Bonilla 2005c)

Contribución:

- (a) Primera propuesta de encadenamiento de conceptos aprovechando la estructura jerárquica del clustering de referencia.
- (b) Metodología de caracterización conceptual por condicionamientos sucesivos (CCCS) con el *Boxplot based discretization* (BbD) original.
- (c) Selección manual de una variable por clase e iteración.
- (d) Priorización de variables totalmente caracterizadoras.
- (e) Aplicación de la metodología CCCS a los datos provenientes de la EDAR de Cataluña, España.

Acrónimo:	[G&PB 06b]
Fecha:	Abril 2006,
Título:	Towards automatic generation of interpretation as a tool for modelling decisions.
Autores:	Karina Gibert, Alejandra Pérez Bonilla
In:	Proceedings of III International Conference on Modeling Decisions for Artificial Intelligence. MDAI'2006.
Páginas:	515 a 524
Lugar:	Tarragona, Catalunya. España.
ISBN:	8400-08416-0
Referencia:	(Gibert and Pérez-Bonilla 2006b)

Contribución:

- (a) Primera versión formal de la metodología de caracterización conceptual por condicionamientos sucesivos (CCCS) con el *Boxplot based discretization* (BbD) original y selección manual de variables.
- (b) Necesidad de definir criterios para la elección de la o las variables que entran en la interpretación, así como, estudiar la forma en que se propaga la incerteza.
- (c) Aplicación de la metodología CCCS a los datos provenientes de la EDAR de Cataluña, España.

Acrónimo: [PB&G 06b]

Fecha: Mayo 2006,

Título: El papel del Boxplot based discretization en la Interpretación conceptual de una clasificación jerárquica.

Autores: Alejandra Pérez Bonilla, Karina Gibert

In: Procs. XXIX Congreso Nacional de Estadística e Investigación Operativa. SEIO'2006.

Páginas: 147 a 148

Lugar: Puerto de la Cruz, Tenerife. España.

ISBN: 84-689-8553-8

Referencia: (Pérez-Bonilla and Gibert 2006)

Contribución:

- (a) Presentación, por primera vez, de la metodología CCCS con el *Boxplot based discretization* (BbD) revisado y selección manual de la variable que entra en cada iteración.
- (b) Las bases de conocimiento generadas por el *Boxplot based discretization* (BbD) revisado tienen mayor grado de certeza que las que se generaban con el *Boxplot based discretization* (BbD) original.
- (c) Aplicación a datos provenientes de la EDAR de Cataluña.

Acrónimo: [G&PB 06a]

Fecha: Julio 2006,

Título: Revised boxplot based discretization as a tool for automatic interpretation of classes from hierarchical cluster.

Autores: Karina Gibert, Alejandra Pérez Bonilla

In: Procs. 10th IFCS Conference: Data Science and Classification. in Artificial. IFCS'2006

Páginas: 229 a 237

Titulo Libro: Data Science and Classification.

Series: Studies in Classification, Data Analysis and Knowledge Organization

Publicado por: Springer-Verlag

Lugar: Ljubljana, Slovenia.

ISBN: 978-3-540-34415-5 (Print) 978-3-540-34416-2 (Online)

ISSN: 1431-8814

Referencia: (Gibert and Pérez-Bonilla 2006a)

Contribución:

- (a) Formalización unificada del *Boxplot based discretization* (BbD) revisado.

Acrónimo: [G&PB&RS 06]
 Fecha: Noviembre 2006,
 Título: A Comparative Analysis of different classes-interpretation support techniques.
 Autores: Karina Gibert, Alejandra Pérez Bonilla y Gustavo Rodriguez-Silva
 In: Novè Congrés Català d'Intel·ligència Artificial: Frontiers in Artificial Intelligence and Applications. CCIA'2006.
 Páginas: 37 a 46
 Volumen: 146
 Titulo Libro: Frontiers in Artificial Intelligence and Applications. Series: Artificial Intelligence. Research and Development
 Publicado por: IOS-Press
 Lugar: Perpignan, France.
 ISBN: 84-9732-449-8
 Referencia: (Gibert, Pérez-Bonilla, and Rodriguez 2006)

Contribución:

- (a) Comparación de la metodología de caracterización conceptual por condicionamientos sucesivos (CCCS)(BbD original) con regresión logística, árboles de decisión y análisis discriminante.
 (b) Aplicación EDAR Catalunya, España.

Acrónimo: [PB&G 07a]
 Fecha: Septiembre 2007,
 Título: Análisis de inconsistencias en bases de conocimientos inducidas automáticamente.
 Autores: Alejandra Pérez Bonilla, Karina Gibert
 In: Procs. XXX Congreso Nacional de Estadística e Investigación Operativa. SEIO'2007.
 Páginas: 97
 Lugar: Valladolid, España.
 ISBN: 978-84-690-7249-3
 Referencia: (Pérez-Bonilla and Gibert 2007a)

Contribución:

- (a) Construcción de una base de conocimiento global con las reglas de mayor grado de confianza de cada variable generada por el *Boxplot based induction rules* (BbIR)
 (b) Definición de 4 criterios que permitirán resolver las inconsistencias entre reglas contradictorias.
 (c) El que mejor se comporta es el que hemos denominado *CCCS* que combina los criterios de probabilidad máxima y clase más votada.
 (d) Aplicación EDAR Catalunya, España y EDAR Ljubljana, Slovenia.

Acrónimo: [PB&G 07b]
 Fecha: Septiembre 2007,
 Título: The role of the Boxplot based Discretization in the conceptual interpretation of a hierarchical cluster.
 Autores: Alejandra Pérez Bonilla, Karina Gibert
 In: IV Taller Nacional de Minería de Datos y Aprendizaje. TAMIDA'2007.
 II Congreso Español de Informática. CEDI'2007.
 Páginas: 157 a 166
 Volumen: 1
 Publicado por: Thompson
 Lugar: Zaragoza. España.
 ISBN: 978-84-9732-602-5
 Referencia: (Pérez-Bonilla and Gibert 2007b)

Contribución:

- (a) Comparación formal de la metodología de caracterización conceptual por condicionamientos sucesivos (CCCS) con el *Boxplot based discretization* (BbD) original y revisado.
- (b) El *Boxplot based discretization* (BbD) revisado genera mayor número reglas seguras.
- (c) Aplicación EDAR Catalunya, España.

Acrónimo: [PB&G 07c]
 Fecha: Noviembre 2007,
 Título: Towards automatic generation of conceptual interpretation of clustering. of classes from hierarchical cluster.
 Autores: Alejandra Pérez Bonilla, Karina Gibert
 In: 12th Iberoamerican Congress on Pattern Recognition. CIARP'2007.
 Páginas: 653 a 663
 Volumen: 4756
 Titulo Libro: Progress in Pattern recognition, Image analysis and Application.
 Series: Lecture Notes in Computer Science
 Publicado por: Springer-Verlag
 Lugar: Valparaiso - Viña del Mar, Chile.
 ISBN: 978-3-540-76724-4 (Print)
 ISSN: 0302-9743 (Print) 1611-3349 (Online)
 Referencia: (Pérez-Bonilla and Gibert 2007c)

Contribución:

- (a) Metodología de caracterización conceptual por condicionamientos sucesivos (CCCS) con *Boxplot based discretization* (BbD) revisado.
- (b) Selección automática de la variable por iteración.
- (c) Criterio de elección: mayor cobertura relativa (sólo con reglas seguras).
- (d) Caracterización de *variable totalmente caracterizadora* a partir de la cobertura relativa.

(e) Aplicación EDAR Catalunya, España.

Acrónimo:	[PB&G&V 08]
Fecha:	Julio 2008,
Título:	Automatic generation of conceptual descriptions of classifications in Environmental Domains.
Autores:	Alejandra Pérez Bonilla, Karina Gibert y Darko Vrecko
In:	Proceedings of the iEMSs Fourth Biennial Meeting iEMSs'2008.
Páginas:	1791 a 1798
Volumen:	3
Publicado por:	International Environmental Modelling and Software Society
Lugar:	Barcelona, Catalunya. España.
ISBN:	978-84-7653-074-0
Referencia:	(Pérez-Bonilla, Gibert, and Vrecko 2008)

Contribución:

- (a) Revisión de la metodología de caracterización conceptual por condicionamientos sucesivos (CCCS) con distintos criterios de integración del conocimiento en cada iteración: Best Global rule +CWA (original), Best local rule +noCWA, Best local rule +partialCWA y Best Local&Global rule +CWA.
- (b) La que presenta mejor comportamiento es: Best Local&Global rule +CWA.
- (c) Utilización de medidas de calidad (Soporte, cobertura relativa, cobertura relativa global y confianza) para valorar las bases de conocimiento que interpretan la partición objetivo.
- (d) Definición de la $CovG_{global}$ como principal medida de calidad de la base de conocimiento final que interpreta la partición deseada.
- (e) Aplicación EDAR Ljubljana, Slovenia.

2. Reportes de Investigación.

Acrónimo:	[G&PB 04a]
Fecha:	Noviembre 2004,
Título:	Clasificación de Algunas Plantas Depuradoras de Aguas Residuales de Cataluña.
Autores:	Karina Gibert y Alejandra Pérez Bonilla.
In:	Document de Recerca.
Número:	DR 2004/13.
Lugar:	Departament d'Estadística i Investigació Operativa. UPC. Barcelona, España.
Referencia:	(Gibert and Pérez-Bonilla 2004a)

Contribución:

- (a) Clasificación con diferentes métodos (Clustering jerárquico, clasificación basada en regla), métricas (euclídea, euclídea normalizada) y parámetros de 4 bases de datos provenientes de 3 EDAR de Catalunya.

Acrónimo: [PB&G 05]

Fecha: Mayo 2005,

Título: Metodología de Caracterización Conceptual por Condicionamientos Sucesivos (CCCS).

Autores: Alejandra Pérez Bonilla y Karina Gibert.

In: Document de Recerca.

Número: DR 2005/12.

Lugar: Departament d'Estadística i Investigació Operativa. UPC.
Barcelona, España.

Referencia: (Pérez-Bonilla and Gibert 2005b)

Contribución:

- (a) Los sistemas de reglas que se desprenden del *Boxplot based discretization* (BbD) revisado son sensibles a la forma como se definan los límites de los intervalos.
- (b) Identificación y corrección de casos anómalos en particiones binarias del *Boxplot based discretization* (BbD).

Acrónimo: [G&PB 05a]

Fecha: Junio 2005

Título: Análisis y propiedades de la metodología Caracterización Conceptual por Condicionamientos Sucesivos (CCCS).

Autores: Karina Gibert y Alejandra Pérez Bonilla.

In: Document de Recerca.

Número: DR 2005/14.

Lugar: Departament d'Estadística i Investigació Operativa. UPC.
Barcelona, España.

Referencia: (Gibert and Pérez-Bonilla 2005a)

Contribución:

- (a) Propuesta del *Boxplot based discretization* (BbD) revisado.
- (b) Integración del *Boxplot based discretization* (BbD) revisado en el *Boxplot based induction rules* (BbIR).
- (c) Comparación de los sistemas de reglas antes y después de la revisión.
- (d) Propiedades del *Boxplot based discretization* (BbD) revisado.

Acrónimo: [PB&G 05a]
Fecha: Septiembre 2005
Título: Avantatges de l'estructura jeràrquica del clustering en la interpretació automàtica de classificacions.
Autores: Alejandra Pérez Bonilla y Karina Gibert.
In: Document de Recerca.
Número: DR 2005/15.
Lugar: Departament d'Estadística i Investigació Operativa. UPC.
Barcelona, España.
Referencia: (Pérez-Bonilla and Gibert 2005a)

Contribución:

- (a) Versión en catalán del artículo presentado en el III Taller Nacional de Minería de Datos y Aprendizaje. TAMIDA'2005.

Acrónimo: [PB&G&V 07b]
Fecha: Abril 2007
Título: Knowledge Discovery on Domzale-Kamnik Wastewater Treatment Plant Ljubljana - Slovenia.
Autores: Alejandra Pérez Bonilla, Karina Gibert y Darko Vrecko.
In: Document de Recerca.
Número: DR 2007/03.
Lugar: Departament d'Estadística i Investigació Operativa. UPC.
Barcelona, España.
Referencia: (Pérez-Bonilla, Gibert, and Vrecko 2007b)

Contribución:

- (a) Descripción detallada del funcionamiento y objetivo de la planta piloto eslovena.
- (b) Descriptiva estadística completa (análisis univariante, análisis bivariante, etc.)
- (c) Clasificación con diferentes métodos, métricas y parámetros.

Acrónimo: [PB&G&V 07a]
Fecha: Diciembre 2007
Título: Domzale-Kamnik Wastewater Treatment Plant (Ljubljana - Slovenia). Clustering and Induction Knowledge Base.
Autores: Alejandra Pérez Bonilla, Karina Gibert y Darko Vrecko.
In: Document de Recerca.
Número: DR 2007/09.
Lugar: Departament d'Estadística i Investigació Operativa. UPC.
Barcelona, España.
Referencia: (Pérez-Bonilla, Gibert, and Vrecko 2007a)

Contribución:

- (a) La clasificación basada en reglas de la planta eslovena es la que se comporta mejor.

- (b) Aplicación del *Boxplot based induction rules* (BbIR) con el *Boxplot based discretization* (BbD) revisado a la clasificación basada en reglas y al clustering jerárquico variable a variable.
- (c) Composición de todas las $\mathcal{S}(X_k, \mathcal{P}_\xi^*)$ en sistema de reglas seguro global $\mathcal{S}(\mathcal{P}_\xi^*)$ y construcción de una base de conocimiento final de reglas seguras de todas las variables para la partición objetivo.
- (d) Definición del grado de cobertura relativa y cardinalidad de las reglas.

Acrónimo: [PB&G 08a]

Fecha: Enero 2008

Título: Estudio de la inconsistencia en las bases de conocimiento inducidas por la Metodología de Caracterización Conceptual por Condicionamientos Sucesivos (CCCS).

Autores: Alejandra Pérez Bonilla y Karina Gibert.

In: Document de Recerca.

Número: DR 2008/03.

Lugar: Departament d'Estadística i Investigació Operativa. UPC.
Barcelona, España.

Referencia: (Pérez-Bonilla and Gibert 2008a)

Contribución:

- (a) Definición de 4 algoritmos para estudiar las inconsistencias que se presentan en las bases de conocimiento inducidas por la metodología CCCS.
- (b) Definición de criterios para solución de inconsistencias.
- (c) Aplicación EDAR Catalunya, España y EDAR Ljubljana, Eslovenia.
- (d) Da origen al proceeding *Análisis de inconsistencias en bases de conocimientos inducidas automáticamente*, presentado en XXX Congreso Nacional de Estadística e Investigación Operativa -SEIO'2007.

Acrónimo: [PB&G 08b]

Fecha: Marzo 2008

Título: Inducción de una base de conocimiento utilizando la Metodología CCCS para una estación depuradora de aguas residuales (EDAR - Cataluña)

Autores: Alejandra Pérez Bonilla y Karina Gibert.

In: Document de Recerca.

Número: DR 2008/04.

Lugar: Departament d'Estadística i Investigació Operativa. UPC.
Barcelona, España.

Referencia: (Pérez-Bonilla and Gibert 2008b)

Contribución:

- (a) Análisis comparativo de los resultados del reporte *Clasificación de Algunas Plantas Depuradoras de Aguas Residuales de Cataluña* (DR 2004/13).

- (b) Generación de la Base de conocimiento inducida por el *Boxplot based induction rules* (BbIR) con el *Boxplot based discretization* (BbD) revisado para la mejor clasificación.
- (c) Integración de las reglas seguras de todas las variables, para cada clase de la partición objetivo.
- (d) Calculo de la cobertura relativa y la cardinalidad de todas las reglas seguras para cada clase de la partición objetivo.

Acrónimo: [PB&G 08c]

Fecha: Septiembre 2008

Título: Integración global del conocimiento a partir de una clasificación jerárquica en la conceptualización final de una partición.

Autores: Alejandra Pérez Bonilla y Karina Gibert.

In: Document de Recerca.

Número: DR 2008/08.

Lugar: Departament d'Estadística i Investigació Operativa. UPC.
Barcelona, España.

Referencia: (Pérez-Bonilla and Gibert 2008c)

Contribución:

- (a) Formalización matemática de la metodología de caracterización conceptual por condicionamientos sucesivos (CCCS).
- (b) Presentación de 5 propuestas diferentes (incluida la original -CCCS) para combinar los resultados inducidos de la base de conocimiento KB^ξ con los de la iteración anterior.
- (c) Definición de una medida de calidad de las bases de conocimientos obtenidas para la interpretación, $CovG_{global}$
- (d) Calculo de otras medidas de calidad definida por otros autores. Elección de la que mejor se comporta en cuenta a las mediadas de calidad y a la evaluación del experto.

Acrónimo: [PB&G 08d]

Fecha: Septiembre 2008

Título: Integración global del conocimiento a partir de una clasificación jerárquica en la conceptualización final de una partición - EDAR Catalunya.

Autores: Alejandra Pérez Bonilla y Karina Gibert.

In: Document de Recerca.

Número: DR 2008/10.

Lugar: Departament d'Estadística i Investigació Operativa. UPC.
Barcelona, España.

Referencia: (Pérez-Bonilla and Gibert 2008d)

Contribución:

- (a) Evaluación de las 2 mejores propuestas de las 5 que se evaluaron en el reporte anterior con la base de datos proveniente de la EDAR de Catalunya.
- (b) Calculo de medidas de calidad de la interpretación generada.

3. Trabajos de Investigación y otros.

1. DEA

Acrónimo: **[PB dea 05]**
 Fecha: Junio 2005,
 Título: Metodología de Caracterización Conceptual por Condicionamientos Sucesivos. Una aplicación a sistemas medioambientales.
 Autor: Alejandra A. Pérez Bonilla.
 In: Diploma de estudios Avanzados en Estadística e Investigación operativa y proyecto de tesis. DEIO-UPC.
 Lugar: Barcelona, España.

2. Master Tesis

Acrónimo: **[PB Mth 04]**
 Fecha: Diciembre 2004,
 Título: Estudio Comparativo Entre Métodos de Clasificación para Problemas de Diagnóstico Médico. Estadística v/s Inteligencia Artificial o Estadística e Inteligencia Artificial.
 Autor: Alejandra A. Pérez Bonilla.
 In: Tesis de grado presentada en conformidad a los requisitos para obtener el grado de Magíster en Ciencias de la Ingeniería, mención Ingeniería Industrial. Material de consulta pública biblioteca central y biblioteca especializada del Departamento de Ingeniería Industrial de la Universidad de Santiago de Chile.
 Lugar: Santiago, Chile.

3. Trabajo tutelado de Investigación

Acrónimo: **[GP&B 04b]**
 Fecha: Junio 2004,
 Título: Estudio comparativo entre clasificación jerárquica, clasificación basada en reglas y redes neuronales feed-forward en problemas de diagnóstico médico.
 Autores: Karina Gibert y Alejandra Pérez Bonilla.
 In: Document de Recerca. DR 2004/08.
 Lugar: Departament d'Estadística i Investigació Operativa. UPC. Barcelona, España.
 Referencia: (Gibert and Pérez-Bonilla 2004b)

References

- Adelman, L. (1992). *Evaluating Decision Support and Expert Systems*. John Wiley and Sons, New York, NY.
- Alter, S. L. (1980). *Decision support systems: current practice and continuing challenges*. Addison-Wesley.
- Aluja, T. (1996). *Análisis Factoriales Descriptivos con SPAD-N*.
- Aluja, T. (2001). La Minería de Datos, entre la Estadística y la Inteligencia Artificial. *QÜESTIÓ*. V.15, N°3, 479–498.
- Anderberg, M. R. (1973). *Cluster Analysis for applications*. Academic Press.
- Annichiarico, R. and K. Gibert (2004). Qualitative profiles of disability. *JRRD* 41(6A). California, USA.
- Baek, Fogel, M. E. (1997). *Handbook of Evolutionary Computation*. Oxford NY.
- Baeza-Yates, R. and J. Pino (2006). Towards formal evaluation of collaborative work and its application to information retrieval. *Information Research* 11(4).
- Ball, G. and D. J. Hall (1965). ISODATA, a novel method of data analysis and pattern classification. Technical report, Stanford Research Institute.
- Barthélemy, J.-P., L. B. and B. Monjardet (1986). On the use of ordered sets in problems of comparison and consensus classifications. *Journal of Classification* (3), pages 187–224.
- Bayona, S. (2000). Descriptiva de dades y de classes. PFC Facultat d'Informàtica, UPC.
- Benzécri, J. (1973). *L'analyse des données*. Paris: Dunod. Tome 1: La Taxinomie, Tome 2: L'analyse des correspondances. First Edition.
- Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective functions*. Plenum Press. New York.
- Bishop, C. (1995). *Neural Networks for Pattern Recognition*. (First ed.). Oxford: Clarendon Press.
- Bock, H. (1985). On some significance tests in cluster analysis. *Journal of Classification* (2), pages 77 – 108.
- Bock, H. (1996). *Probability models and hypotheses testing in partitioning cluster analysis*.
- Bock, H. H. and E. Diday (1999). *Analysis of symbolic data*. Berlin, Germany: Springer Verlag.
- Bouldin, D. and D. Davies (1979). A cluster separation measure. *EEE Transactions on Pattern Recognition and Machine Intelligence* 1 (2), 224–227.
- Brachman, R. and T. Anand (1996.). The Process of Knowledge Discovery in Databases: A Human-Centered Approach. *In Advances in Knowledge Discovery and Data Mining*, 65–78.

- Castillejo, X. (1996). Un entorn de treball per a Klass. PFC Facultat d'Informàtica, UPC.
- Cheeseman, P. and R. W. Oldford (eds.) (1994). *Artificial Intelligence and Statistics IV*, Volume 89 of *LNS*. NY, USA: Springer.
- Cheeseman, P. and J. Stutz (1996). Bayesian classification (auto-class): theory and results. pp. 153–180.
- Chieppa, A., K. Gibert, S.-M. M., and I. Gómez-Sebastià (2008). Improving pseudobagging techniques. In T. Alsinet, J. Puyol-Gruart, and C. Torras (Eds.), *Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence*, Volume 184 of *Artificial Intelligence Research and Development*, pp. 161–169. IOS press.
- Comas, J., S. Dzeroski, K. Gibert, I. Roda, and M. Sànchez-Marrè (2001). Knowledge discovery by means of inductive methods in wastewater treatment plant data. *AI Communications. The european journal on artificial intelligence* 14(1), 45–62.
- Cormack, R. (1971). A review of clasification. In *Journal of the Royal Statistical Society (Series A)*, pp. 134: 321–367.
- Cortés, U., M. Sànchez-Marrè, L. Ceccaroni, I. R-Roda, and M. Poch (2000). Artificial intelligence and environmental decision support systems. *Applied Intelligence* 13(1), 77–91.
- Cuadras, C. (1991). *Métodos de análisis multivariante*. Promociones y publicaciones universitarias S.A. Barcelona.
- De Andrés Argente, T. (2002). *Homo Cibersapiens. La Inteligencia Artificial y La Humana*. Eunsa. Ediciones Universidad de Navarra.
- De Rham, C. (1997). La classification hiérarchique selon la méthode des voisins réciproques. *Cahiers d'Analyse des Données* 5(2), 135–144.
- Devijver, P. and J. Kittler (1982). *Pattern Recognition: A Statistical Approach*. (First ed.). London: Prentice Hall.
- Devroye, L. Gyorfi, L. and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. (First ed.). Berlin: Springer Verlang.
- Diday, E. (1971.). La méthode des nuées dynamiques. *Statistics Applications*. 2(19), 19–34.
- Diday, E., P. Brito, and M. Mfoumou (1993.). Modelling probabilistic data by conceptual pyramidal clustering. *Proc. of 4th Int'l Work. on AI&Stats.*, 213–218.
- Diday, E. and K. Gowda (1991). Symbolic clustering using a new dissimilarity measure. *Pattern Recognition* 24(6), 567–578.
- Diday, E. and J. Moreau (1984). Learning hierarchical clustering from examples. In N. . Centre de Rocquencourt, Rapports de Recherche (Ed.), *INRIA*.
- Diday, E. e. (1979). *Optimisation en classification automatique. Tome 1 et 2*. Rocquencourt: Institut National de Recherche en Informatique et en Automatique (INRIA).
- Dillon, W. and M. Goldstein (1984). *Multivariate analysis... .* Wiley. USA.
- Drucker, P. F. (1969). *The Age of Discontinuity*. New York, NY: Harper & Row.
- Drucker, P. F. (1974). *La sociedad postcapitalista*. Norma.
- Druzdzel, M. J. and R. R. Flynn (1999). *Decision Support Systems. Encyclopedia of Library and Information Science*. Addison-Wesley.
- Dubes, R. and A. Jain (1980). *Clustering Methodologies in Exploratory Data Analysis*, Volume 19. Advances in Computers.

- Duda, R. and P. Hart (1973). *Pattern Classification and Scene Analysis*. New York: Wiley and Sons.
- Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics* 4, 95–104.
- Everitt, B. (1981). *Cluster Analysis*. London, England.: Heinemann Educational Books.
- Farré, R., R. Nieuwenhuis, P. Nivela, A. Oliveras, and E. Rodríguez (2008). *Introducción a la lógica*. Notas de clase. FIB. Barcelona.
- Fayyad, U. (1996). From data mining to knowledge discovery: An overview. *Advances in KD and DM, Fayyad, AAAI/MIT..*
- Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth (1996.). From Data Mining to Knowledge Discovery in Databases (a survey). *AI Magazine*. 3(17), 37–54.
- Fisher, D. (1993). Machine Learning. Including a discussion on neural networks. In Florida (Ed.), *4th Int'l Work. on AI&Stats*.
- Forgy, E. (1962). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics* 21, 768–769.
- Fox, J. and S. Das (2000). *Safe and sound*. AAAI Press/The MIT Press.
- Fu, K. (1983). A step toward unification of syntactic and statistical pattern recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*. 5(2), 200–205.
- Fukunaga, K. (1990.). *Introduction to Statistical Pattern Recognition*. Academic Press.
- Gantz, J. F. and et.al. (2007). *The Expanding Digital Universe. A Forecast of Worldwide Information Growth Through 2010*. IDC White Paper.
- Garey, M. R. and D. S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman.
- Geffner, H. (1992). *Default Reasoning: Causal and Conditional Theories*. MIT Press Cambridge, MA, USA.
- Getoor, L., N. Friedman, D. Koller, A. Pfeffer, and B. Taskar (2007). Probabilistic relational models. In L. Getoor and B. Taskar (Eds.), *Introduction to Statistical Relational Learning*. MIT Press.
- Gibert, K. Garc-Rudolph, A. G.-M. A. R.-R. T. B. M. and J. Tormos (2008). Response to tbi-neurorehabilitation through an ai& stats hybrid kdd methodology. *Medical Archives* 62(3).
- Gibert, K. (1991). Klass. Estudi d'un sistema d'ajuda al tractament estadístic de grans bases de dades. Master's thesis, UPC.
- Gibert, K. (1994). *L'ús de la Informació Simbòlica en l'Automatització del Tractament Estadístic de Dominis Poc Estructurats*. In the statistics and operations research phd. thesis., Universitat Politecnica de Catalunya, Barcelona, Spain.
- Gibert, K. (1996a). On the uses and costs of rules-based classification. In A. P. Physica-Verlag (Ed.), *Proceedings of Computational Statistics*, pp. 265–270.
- Gibert, K. (1996b). The use of symbolic information in automation of statistical treatment for ill-structured domains. *AI Communications* 9(1), 36–37.
- Gibert, K. (2004). *Tendencia de la Minería de Datos en España*, Chapter :Técnicas híbridas de Inteligencia Artificial y Estadística para el descubrimiento de conocimiento y la minería de datos, pp. 119–130. Thompson Ed.

- Gibert, K. (2008). *Estadística: Contexto histórico. Introducción a la descriptiva*. DEIO-FME. UPC.
- Gibert, K. and T. Aluja (1998). A computational technique for comparing classifications and its relationship with knowledge discovery. In *International Seminar on New Techniques and Technologies for Statistics*, Italy, pp. 193–198.
- Gibert, K., T. Aluja, and U. Cortés (1998). Knowledge Discovery with Clustering Based on Rules. Interpreting results. In *Principles of Data Mining and Knowledge Discovery*, Volume 1510 of *LNAI*, pp. 83–92. Springer.
- Gibert, K. and U. Cortés (1992.). KLASS: Una herramienta estadística para la creación de prototipos en dominios poco estructurados. *proc. IBERAMIA-92.*, 483–497. Noriega Eds. México.
- Gibert, K. and U. Cortés (1993a). Combining a knowledge based system with a clustering method for an inductive construction of models. In *Proc. 4th Int Work. on AI and Stats.* Florida, USA.
- Gibert, K. and U. Cortés (1993b). On the uses of the expert knowledge for automatic biasing of a clustering method. In *ITI 93. Proceedings of the International Conference on Information Technology Interfaces*, Croatia, pp. 219–224. issn 1330-1012.
- Gibert, K. and U. Cortés (1994). *Combining a knowledge-based system and a clustering method for a construction of models in ill-structured domains*, Volume 89 of *LNS*, pp. 351–360. Springer-Verlag.
- Gibert, K. and U. Cortés (1997). Weighing quantitative and qualitative variables in clustering methods. *Mathware and Soft Computing* 4(3), 251–266.
- Gibert, K. and U. Cortés (1998.a). Clustering based on rules and knowledge discovery in ill-structured domains. *Computación y Sistemas*. 1(4), 213–227.
- Gibert, K. and U. Cortés (1998b). Generación automática de reglas a partir de la caracterización de clases. *Butlletí de l'ACIA*, 14–15.
- Gibert, K. and A. García-Rudolph (2008). Posibilidades de aplicación de la minería de datos para el descubrimiento de conocimiento a partir de la práctica clínica en tecnologías aplicadas al proceso neurorehabilitador: Estrategias para evaluar su eficacia. Volume 19, pp. 93–106. Fund. Ins. Guttmann.
- Gibert, K., A. García-Rudolph, A. García-Molina, T. Roig-Rovira, M. Bernabeu, and J. Tormos (2008). Knowledge discovery on the response to neurorehabilitation treatment of patients with traumatic brain injury through an ai&stats and graphical hybrid methodology. In T. Alsinet, J. Puyol-Gruart, and C. Torras (Eds.), *Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence*, Volume 184 of *Artificial Intelligence Research and Development*, pp. 170–177. IOS press.
- Gibert, K., M. Hernández, and U. Cortés (1996). Classification based on rules: an application to Astronomy. In U. T. Japón (Ed.), *5th. IFCS*, pp. 69–72.
- Gibert, K., J. Izquierdo, G. Holmes, I. Athanasiadis, J. Comas, and M. Sàncchez-Marré (2008). On the role of pre and post-processing in environmental data mining. In M. Sàncchez-Marrè, J. Béjar, J. Comas, A. Rizzoli, and G. Guariso (Eds.), *Proceedings of the iEMSs Fourth Biennial Meeting*, Volume 3, Barcelona-Catalunya, pp. 1937–1958.
- Gibert, K. and R. Nonell (2003). Impact of mixed metrics on clustering. *LNCS 2905*.
- Gibert, K. and R. Nonell (2005). Descriptive statistics with klass. supporting latex documents ellaboration. In *Programme and abstracts: 3rd Word Conference on Computational Statistics & Data Analysis. University of Cyprus*, pp. 90.

- Gibert, K. and R. Nonell (2008). Pre and postprocessing in klass. In M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli, and G. Guariso (Eds.), *Proceedings of the iEMSS Fourth Biennial Meeting*, Volume 3, Barcelona-Catalunya, pp. 1965–1966.
- Gibert, K., R. Nonell, J. M. Velarde, and M. M. Colillas (2004). Kdd with clustering:: Impact of metrics and reporting phase by using klass. In *COMPSTAT Procs.*, pp. 1069–1076. Physica-Verlag.
- Gibert, K., R. Nonell, J. M. Velarde, and M. M. Colillas (2005). Knowledge discovery with clustering: impact of metrics and reporting phase by using klass. *Neural Network World 4/05*, 319–326.
- Gibert, K., L. Oliva, M. Sànchez-Marré, and I. Pinyol (2007). Pseudobagging: Improving class discovery by adapting bagging techniques to clustering algorithms. In f. Ferrer-Troyano, A. Troncoso, and J. Riquelme (Eds.), *Actas del IV Taller de Minería de Datos y Aprendizaje (TAMIDA 2007)*, Zaragoza, pp. 157–166. THOMSON.
- Gibert, K. and A. Pérez-Bonilla (2004a). Clasificación de Algunas Plantas Depuradoras de Aguas Residuales de Cataluña. Research DR 2004/13, Universidad Politécnica de Cataluña, Barcelona, España.
- Gibert, K. and A. Pérez-Bonilla (2004b). Estudio comparativo entre clasificación jerárquica, clasificación basada en reglas y redes neuronales feed-forward en problemas de diagnóstico médico. Research DR 2004/08, Universidad Politécnica de Cataluña, Barcelona, España.
- Gibert, K. and A. Pérez-Bonilla (2005a). Análisis y propiedades de la metodología Caracterización Conceptual por Condicionamientos Sucesivos (CCCS). Research DR 2005/14, Dep. Estadística e Investigación Operativa. Universidad Politécnica de Cataluña, Barcelona, España.
- Gibert, K. and A. Pérez-Bonilla (2005b). Fuzzy Box-plot based Induction Rules. Towards automatic generation of classes-interpretations. In *Fuzzy sets in learning and data mining. 4th EUSFLAT.*, Barcelona, España, pp. 524–529.
- Gibert, K. and A. Pérez-Bonilla (2005c). Ventajas de la estructura jerárquica del clustering en la interpretación automática de clasificaciones. In *III Spanish Workshop on Data Mining and Learning*, Granada, pp. 67–76.
- Gibert, K. and A. Pérez-Bonilla (2006a). Revised boxplot based discretization as a tool for automatic interpretation of classes from hierarchical cluster. In *Data Science and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, Ljubljana, Slovenia, pp. 229–237. Springer-Verlag.
- Gibert, K. and A. Pérez-Bonilla (2006b). Towards automatic generation of interpretation as a tool for modelling decisions. In URV (Ed.), *Procs. of III International Conference on Modeling Decisions for Artificial Intelligence*, Tarragona, Catalunya. España, pp. 515–524.
- Gibert, K., A. Pérez-Bonilla, and G. Rodriguez (2006). A Comparative Analysis of different classes-interpretation support techniques. In M. Polit, T. Talbert, B. López, and J. Meléndez (Eds.), *Frontiers in Artificial Intelligence and Applications.*, Volume 146 of *Artificial Intelligence. Research and Development*, pp. 37–46. IOS press.
- Gibert, K. and I. Roda (2000). Identifying characteristic situations in wastewater treatment plants. In *Workshop in Binding Environmental Sciences and Artificial Intelligence*, Volume 1, pp. 1–9.

- Gibert, K., I. Rodríguez-Roda, and U. Cortés (2004). Identifying characteristic situations in wastewater treatment plants with kdisd. *Applied Intelligence* (4), 319–324.
- Gibert, K. and G. Rodríguez-Silva (2008). *Identification of More Characteristic Dynamic Patterns in a WWTP by CIBRxE, Iberoamerican Congress in Pattern Recognition*, Volume 5197 of *Lecture Notes in Computer Science*, pp. 372–380. Springer Berlin.
- Gibert, K. and A. Salvador (2000). Aproximación difusa a la identificación de situaciones características en el tratamiento de aguas residuales. In *X Congreso Español sobre tecnologías y lógica fuzzy*, España, pp. 497–502.
- Gibert, K. and Z. Sonicki (1997). Classification Based on Rules and Medical Research. In R. Curto (Ed.), *VIII International Symposium on Applied Stochastic Models and Data Analysis*, Italy, pp. 181–186. ASMDA 97.
- Gibert, K. and Z. Sonicki (1999). Classification based on rules and thyroids dysfunctions. *Applied Stochastic Models in Business and Industry* 15(4), 319–324.
- Gibert, K., Z. Sonicki, and J. Martin (2002). Impact of data encoding and thyroids dysfunctions. *Studies in Health Technology and Informatics* 90, 494–498.
- Gibert, K., Z. Sonicki, and J. C. Martín (2001). Impact of data encoding and thyroids dysfunctions. *Technology and Informatics* 90, 494–503.
- Gibert, K., J. Spate, M. Sànchez-Marrè, J. Comas, and I. Athanasiadis (2008). Data Mining for Environmental Systems. In *State of the art and Futures in Environmental Modelling and Software. IDEA Series* (Jackeman, A. J. and Rizzoli, A. and Voinov, A. and Chen, S. (eds)).
- Goldberg, A. (1988). *A History of Personal Workstations*. Addison-Wesley Publishing Co.
- Gordon, A. (1980). *Classification*. Chapman & Hall, London.
- Gordon, A. D. (1994). Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis*, V.18: 561–581.
- Gordon, A. D. (1996). Hierarchical classification. *Clustering and Classification*, 65–122.
- Gowda, K. C. and E. Diday (1992). Symbolic clustering using a new similarity measure. *IEEE Transaction on systems, mans and cybernetics* 22(2)(2), 368–378.
- Gower, J. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53, 315–328.
- Gower, J. (1967). A comparison of some methods of cluster analysis. *Biometrics* 23(4), 623–37.
- Gower, J. (1971). A General coefficient of similarity and some of its properties. *Biometrics* 27, 857–874.
- Haagsma, I. and R. Johanns (1994). *Environmental Systems*, Chapter Decision support systems: an integrated approach, pp. 205–212. Chicago Linguistic Society.
- Haddawy, P., A. Restificar, B. Geisler, and J. Miyamoto (2003). Preference elicitation via theory refinement. *Journal of Machine Learning Research* 4, 2003.
- Halkidi, M., Y. Batistakis, and M. Vazirgiannis (2001). On clustering validation techniques. *Journal of Intelligent Information Systems* 17, 107–145.
- Hand, D. J. (1996). Classification and computers: shifting the focus. In *COMPSTAT: Proceedings in Computational Statistics*, pp. 77–88. Physica-Verlag.
- Hartigan, J. (1975). *Clustering Algorithms*. London (England): John Wiley & Sons.

- Hotelling, H. (1931). The generalization of student's ratio. *Annals of Mathematical Statistics* (2), 360–378.
- Hubert, L. (1987). *Assignment Methods in Combinatorial Data Analysis*. Marcel Dekker, New York.
- Hughes, G. and M. Cresswell (1968). *An Introduction to MODal Logic*. London, eds.
- Huh, M. Y. and K. Song (2002). Davis: A java-based data visualization. *Computational Statistics* 17(3), 411–423.
- Hvala, I. (2004). SMAC - SMArt Control of wastewater systems. Slovene T.I.P. Report Period: 2002-2004 September, Department of Systems and Control, Jozef Stefan Institute. EVK1-CT-2000-00056. Homepage: www.smac.dk.
- Ian H. Witten, E. F. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann. Reviews: Review by J. Geller (SIGMOD Record, Vol. 32:2, March 2002), Review by E. Davis (AI Journal, Vol. 131:1-2, September 2001), Review by P.A. Flach (AI Journal, Vol. 131:1-2, September 2001).
- Ichino, M. and H. Yaguchi (1989). Generalized Minkowski metrics for mixed features. *Trans. IEICE Japó* J72-A(2), 398–405. (en japonès).
- Ichino, M. and H. Yaguchi (1994). Generalized Minkowski Metrics for Mixed feature-type data analysis. *IEEE Transaction on systems, man and cybernetics* 22(2), 146–153. April.
- Jain, A., R. Dubes, and C. Chen (1987). Bootstrap Techniques for error estimation. *IEEE Trans. Pattern Analysis and Machine Intelligence*. 9, 628–633.
- Jardine, N. and R. Sibson (1971.). Choice of Methods for Automatic Classification. *The Computer Journal* 14 (4), 404–406.
- Jeiss., F. (1999). *MultiAgent Systems: an Introduction to Distributed Artificial intelligence*. Addison Wesley.
- Kantrowitz, M. (1994.). Milestones in the Development of Artificial Intelligence 1994. web.
- Keen, P. G. W. (1978). *Decision support systems: an organizational perspective*. Reading, Mass, Addison-Wesley.
- Kocijan, J. Vrecko, D. H. N. C. B. (2004.). Feedback control aspects of nitrogen removal in wastewater treatment. *Electrotechnical Conference. MELECON 2004. Proceedings of the 12th IEEE Mediterranean* 1., 375–378.
- Kohonen, T. (1995). Self-Organizing Maps. *Springer Series in Information Sciences*. 30.
- Kolmogórov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitrechnung*. Ergebnisse der Mathematik, Berlin.
- Kolmogórov, A. (1956). *Foundations of the Theory of Probability*. New York: Chelsea Publishin Company, second english edition.
- Kruskal, W. and L. Goodman (1954). Measures of associations for cross-validations. *J. Am. Stat. Assoc.* 49, 732–764.
- Lapointe, F. and P. Legendre (1990). A statistical framework to test the consensus of two nested classifications. *Systematic Zoology*. (39), pages 1–13.
- Lapointe, F. and P. Legendre (1995). Comparison tests for dendrograms: A comparative evaluation. *Journal of Classification* (12), pp. 265–282.
- Lean, G. and D. Hinrichsen (1994). *Wastewater engineering treatament. Disposal and reuse*. World Wildlife Fund (U.S.). Harpercollins, April.

- Lebart, A., A. Morineau, and J. Fenelon (1985). *Tratamiento estadístico de datos*. Marcombo.
- Lebart, L. (1990). *Traitemet statistique des données*. Paris: DUNOD.
- Lebart, L., A. Morineau, and T. Lambert (1994). *SPAD.N : manual de referencia, versión 2.5. Sistema compatible para el análisis de datos*. Saint-Mandé: CISIA. traducido por: T. Aluja , E.Ibáñez.
- Lee, S.-C. and M. Y. Huh (2003, October). A measure of association for complex data. *Computational Statistics & Data Analysis* 44(1-2), 211–222.
- Liu, B. Hsu, W. C. S. M. Y. (2000). Analyzing the subjective interestiness of association rules. *IEEE Intelligent Systems*, 47–55.
- López de Mántaras, R. (1990). *Approximate reasoning models*. Ellis Horwood series in AI.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observation. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.*, pp. 281–297.
- Mahalanobis, P. (1936). *Proceedings National Institute of Science*. India, 12 - 236-244. traducido por: T. Aluja , E.Ibáñez.
- Mamdani, E., , and G. Gaines (1981). *Fuzzy reasoning and its Applications*. Mamdani-Gains eds.
- Mansell, R. (2000). *Mobilizing the Information Society: Strategies for Growth and Opportunity*. Oxford University Press.
- Marjeta Strazar, P. C. and O. Burica (2006). Porocilo O Delu Centralne Cistilne Naprave Domzale. Kamnik D.O.O. V Letu. Technical report, Department of Systems and Control, Jozef Stefan Institute.
- Márquez, J. and J. Martín (1997). La clasificación automática en las ciencias de la salud. PFC. Facultat de Matemàtiques i Estadística, UPC.
- McCarthy, J. J. (1983). *Papers from the Parasession on the Interplay of Phonology, Morphology, and Syntax*, Chapter Phonological features and morphological structure, pp. 135–161. Chicago.: Chicago Linguistic Society.
- McCutcheon, A. (1987). Latent class analysis. *Quantitative aplications in the social sciences* (64).
- McDermott, J. (1982). R1: A rule-based configurer of computer systems.
- McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. (First ed.). New York: Wiley and Sons.
- Metcalf and Eddy (2003). *Wastewater engineering treatament. Disposal and reuse*. McGraw-Hill. 4th Ed. revised by George Tchobanoglous, Franklin L. Burton NY.US.
- Michalski, R. (1980). Knowledge acquisition through conceptual clustering: A theoretical framework and algorithm for partitioning data... *IJPAIS* 4, 219–243.
- Michalski, R. and R. Stepp (1983.). Learning from Observation Conceptual Clustering. In *Machine Learning: An artificial intelligence approach.*, 331–363. Morgan Kaufmann.
- Milligan., G. (1996). *Clustering validation: Results and implications for applied analyses*.
- Milligan, G. W. and M. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* (50), 159–179.
- Minsky, M. (1961.). Steps Toward Artificial Intelligence. *Marvin Minsky, Proc. IRE* 49, 8–30.

- Morales, S. (2008). Transfer learning for bayesian networks. In S. Berlin-Heidelberg (Ed.), *Advances in Artificial Intelligence. IBERAMIA 2008*, Volume 5290 of *Lecture Notes in Computer Science*. Springer.
- Nagy, G. (1968). State of the art in pattern recognition. *Proc. IEEE.* 56, 836–862.
- Nakhaeizadeh, G. (1996). Classification as a subtask of Data Mining experiences from some industrial projects. In *IFCS, v-I*, Kobe, JAPAN, pp. 17–20.
- Nilsson, N. (1986). Probabilistic logic. *Artificial Intelligence Journal* 71-88, 2003.
- Oliveras Castellà, Josep. Dades heterogènies amb classificació automàtica. Implementació i comparativa de mètriques mixtes. PFC.
- Pérez-Bonilla, A. (2005). Metodología de Caracterización Conceptual por Condicionamientos Sucesivos. Una aplicación a sistemas medioambientales. Master's thesis, Dep. Estadística e Investigación Operativa. Universidad Politécnica de Cataluña.
- Pérez-Bonilla, A. and K. Gibert (2005a). Avantatges de l'estructura jeràrquica del *clustering* en la interpretació automàtica de classificacions. Research DR 2005/15, Barcelona, España.
- Pérez-Bonilla, A. and K. Gibert (2005b). Metodología de Caracterización Conceptual por Condicionamientos Sucesivos (CCCS). Research DR 2005/12, Dep. Estadística e Investigación Operativa. Universidad Politécnica de Cataluña, Barcelona, España.
- Pérez-Bonilla, A. and K. Gibert (2006). El papel del Boxplot based discretization en la Interpretación conceptual de una clasificación jerárquica. In *In Procs. XXIX Congreso Nacional de Estadística e Investigación Operativa (SEIO2006)*, Puerto de la Cruz, Tenerife. España, pp. 147–148.
- Pérez-Bonilla, A. and K. Gibert (2007a). Análisis de inconsistencias en bases de conocimientos inducidas automáticamente. In *In Procs. XXX Congreso Nacional de Estadística e Investigación Operativa (SEIO2007)*, Valladolid, Spain, pp. 97.
- Pérez-Bonilla, A. and K. Gibert (2007b). The role of the Boxplot based Discretization in the conceptual interpretation of a hierarchical cluster. In f. Ferrer-Troyano, A. Troncoso, and J. Riquelme (Eds.), *Actas del IV Taller de Minería de Datos y Aprendizaje (TAMIDA 2007)*, Zaragoza, pp. 157–166. THOMSON.
- Pérez-Bonilla, A. and K. Gibert (2007c). Towards automatic generation of conceptual interpretation of clustering. In L. Rueda, D. Mery, and J. Kittler (Eds.), *Progress in Pattern recognition, Image analysis and Application.*, Volume 4756 of *Lecture Notes in Computer Science*, Valparaíso - Viña del Mar. Chile, pp. 653–663. Springer-Verlag.
- Pérez-Bonilla, A. and K. Gibert (2008a). Estudio de la inconsistencia en las bases de conocimiento inducidas por la Metodología de Caracterización Conceptual por Condicionamientos Sucesivos (CCCS). Research DR 2008/03, Dep. Estadística e Investigación Operativa. Universidad Politécnica de Cataluña, Barcelona, España.
- Pérez-Bonilla, A. and K. Gibert (2008b). Inducción de una base de conocimiento utilizando la Metodología CCCS para una estación depuradora de aguas residuales (EDAR - Cataluña). Research DR 2008/04, Dep. Estadística e Investigación Operativa. Universidad Politécnica de Cataluña, Barcelona, España.
- Pérez-Bonilla, A. and K. Gibert (2008c). Integración global del conocimiento a partir de una clasificación jerárquica en la conceptualización final de una partición . Research DR 2008/08, Dep. Estadística e Investigación Operativa. Universidad Politécnica de Cataluña, Barcelona, España.

- Pérez-Bonilla, A. and K. Gibert (2008d). Integración global del conocimiento a partir de una clasificación jerárquica en la conceptualización final de una partición - EDAR Catalunya. Research DR 2008/10, Dep. Estadística e Investigación Operativa. Universidad Politécnica de Cataluña, Barcelona, España.
- Pérez-Bonilla, A., K. Gibert, and D. Vrecko (2007a). Domzale-Kamnik Wastewater Treatment Plant (Ljubljana - Slovenia). Clustering and Induction Knowledge Base. Research DR 2007/09, Dep. Estadística e Investigación Operativa. Universidad Politécnica de Cataluña, Barcelona, España.
- Pérez-Bonilla, A., K. Gibert, and D. Vrecko (2007b). Knowledge Discovery on Domzale-Kamnik Wastewater Treatment Plant Ljubljana - Slovenia. Research DR 2007/03, Dep. Estadística e Investigación Operativa. Universidad Politécnica de Cataluña, Barcelona, España.
- Pérez-Bonilla, A., K. Gibert, and D. Vrecko (2008). Automatic generation of conceptual descriptions of classifications in environmental domains. In M. Sàncchez-Marrè, J. Béjar, J. Comas, A. Rizzoli, and G. Guariso (Eds.), *Proceedings of the iEMSS Fourth Biennial Meeting*, Volume 3, Barcelona-Catalunya, pp. 1791–1798.
- Poole, D. (1997). The independent choice logic for modelling multiple agents under uncertainty. *Artificial Intelligence* 94, 7–56.
- Power, D. J. (2002). *Decision support systems: concepts and resources for managers*. Westport, Conn. and Quorum Books.
- Quinlan, J. (1990). Learning logical definitions from relations. *Machine Learning*. 5(3), 239–266.
- Quinlan, J. (1993). C4.5, programs for machine learning. *Machine Learning*.. Morgan Kauffman.
- Ralambondrainy, H. (1988). *A clustering method for nominal data and mixture of numerical and nominal data. Clasification and Related Methods of Data Analysis*. H.H.Bock, Elsevier Science Publishers, B.V. (North-Holland).
- Ralambondrainy, H. (1995a). A conceptual version of K-means algorithm. *Pattern Recognition Letters*. 16, 1147–1157.
- Ralambondrainy, H. (1995b). *A conceptual version of the K-means algorithm*. Lifetime Learning Publications, Belmont, California.
- Raya, A. (2007, Jul). Incorporació de la classificació basada en regles a java-klass. PFC. Facultat d'Informàtica, UPC.
- Reiter, R. (1978a). On closed world databases. In H. Gallaire and J. Minker (Eds.), *Logic and Data Bases*, pp. 119–140. New York: Plenum.
- Reiter, R. (1978b). On reasoning by default. In *Proceedings TINALP-2*, University of Illinois, Urbana-Champaign, pp. 210–218. Association for Computational Linguistics.
- Ripley, B. (1996). *Pattern Recognition and Neural Networks*. (First ed.). Cambridge: Cambridge University Press.
- Robertson, C. K., D. L. McCracken, and A. Newell (1979). The zog approach to man-machine communication. Research CMU-CS-79-148, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.
- Rodas, J., K. Gibert, and J. Rojo (2001). Electroshock Effects Identification Using Classification Techniques. *Springer's Lecture Notes of Computer Science Series Crespo, Maojo and Martin (Eds.)*, 238–244. Second International Symposium, ISMDA 2001.

- Rodas, J., J. Gramajo, and K. Gibert (2000). AI versus Statistics : Some Common Topics. Research DR 2000-13, Technical University of Catalonia, Barcelona, Spain.
- Rodas Osollo, J. E. (2003). *Knowledge Discovery in repeated and very short serial measures with a blocking factor*. Programa de doctorado: Inteligencia artificial, Universitat Politecnica de Catalunya.
- Rodríguez, D. (1999). Anàlisis de los datos de una depuradora de aguas utilizando clasificación basada en reglas. PFC. Facultat de Matemàtiques i Estadística, UPC.
- Romesburg, H. (1990). *Cluster Analysis for Researchers*. Malabar, FL. USA.: Krieger Publishing Company.
- Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20(1), 53–65.
- Roux, M. (1985.). Algorithmes de classification. Paris: Masson, Paris, France.
- Ruiz-Shulcloper, J., M. Chac-Kantún, and J. Martínez-Trinidad (1997). Bases conceptuales para una teoría de objetos simbólicos. *Computación y Sistemas* 1, 13–20.
- Ruiz-Shulcooper *et al.*, J. (1996). Data analysis between sets of objects. In *8th ICSRIC*, Volume III, Baden Baden, pp. 85–81.
- Sànchez-Marrè, M. (1995). *An Integrated Supervisory Multi-level Architecture for Waste-Water Treatment Plants*. Ph. D. thesis, UPC.
- Schuhfried, G. (1992). *Wiener Testsystem. Vienna Reaction Unit, Basic Program*. Development and production of scientific equipment. Mödling, Austria.
- Schultz, J. and L. Hubert (1976). Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychologie* 29, 190–241.
- Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27(July), 379–423.
- Shoham, Y. (1993). Agent-Oriented Programming. *Artificial Intelligence.*, 60–1, 51–92.
- Shortlife, E. (1976). *MYCIN: A rule-based computer program for advising physicians regarding antimicrobial therapy selection*. Ph. D. thesis, Stanford University, USA, USA.
- Simon, H. A. (1960). *The New Science of Management Decision*. New York: Harper & Row.
- Sànchez-Marrè, M. and et.al. (2006). Uncertainty management, spatial and temporal reasoning and validation of intelligent environmental decision support systems. In *Proceedings of the iEMSS 2006*. Internet: <http://www.iemss.org/iemss2006/sessions/all.html>.
- Sànchez-Marrè, M. and et.al. (2008). Towards a framework for the development of intelligent environmental decision support systems. In *Proceedings of the iEMSS 2008*. Internet:<http://www.iemss.org/iemss2008/uploads/Main/Vol1-iEMSS2008-Proceedings.pdf>.
- Sojda, R. (2002). *Ph. D. Thesis: Artificial intelligence based decision support for trumpeter swan management*. Fort collins, colorado., Colorado State University, USA.
- Sokal, R. and P. Sneath (1963). *Principles of numerical taxonomy*. San Francisco. Freeman.
- Sokal, R. R. and C. Michener (1958). A statistical method for evaluating systematic relationship . *Univ. Kansas Sci. Bull.* 38, 1409–1438.
- Stare, A., N. Hvala, and D. Vrecko (2006). Modeling Identification, and Validation of Models for Predictive Ammonia Control in a Wastewater Treatment Plant. A Case Study. *ISA Transactions* 45(2), 159–174.

- Stehr, N. (Ed.) (1994). *Knowledge Societies*. Sage.
- Tubau, X. (1999, octubre). Sobre el comportament de les mètriques mixtes en algorismes de Clustering. PFC. Facultat d'Informàtica, UPC.
- Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley.
- Vázquez, F. and K. Gibert (2001). Generación automática de reglas difusas en dominios poco estructurados con variables numéricas. In *IXth CAEPIA v. I*, pp. 143–152.
- Vázquez, F. and K. Gibert (2002). Robustness of class prediction depending on reference partition in ill-structured domains. In *XVIII Iberoamerican Conference on Artificial Intelligence, Workshop de Minería de Datos y Aprendizaje (IBERAMIA2002)*, Sevilla, pp. 13–22.
- Vázquez Torres, F. (2002). Caracterización e Interpretación Automática de Descripciones Conceptuales en Dominios Poco Estructurados usando Variables Numéricas. Master's thesis, Facultad de Informática de Barcelona. Universidad Politécnica de Cataluña.
- Visauta, B. (1998). *Análisis Estadístico con SPSS para WINDOWS*. Mc.Graw Hill. (Vol II. Análisis Multivariante).
- Volle, M. (1985). Analyse des données. Ed. Economica, Paris, France.
- Vrecko, D. and e. Hvala, N. (2006.). Improvement of ammonia removal in activated sludge process with feedforward-feedback aeration controllers. *Water Science & Technology IWA Publishing 53. No 4-5*, 125–132.
- Ward, J. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Statis. Ass.* 58.
- Zadeh, L. (1965). Fuzzy Sets. *Information and Control*, 338–353.
- Zadeh, L. (1973). The Concept of a linguistic variable and its application to approximate reasoning. *Memorandum ERL-M 411*, 101–105.
- Zadeh, L. (1993). The role of fuzzy logic and soft computing in the conception and design of intelligent systems. *8th Austrian Artificial Intelligence Conference, LNAI 695*. 695, 1–5.

Parte V

Anexos

Anexo A

KLASS

A.1 Introducción a *KLASS*

KLASS es una herramienta de clustering orientada a dominios poco estructurados. Como los métodos estadísticos clásicos tienen pobre desempeño en estos dominios, surge la idea de trabajar en una nueva dirección: usar restricciones declarativas para solventar las deficiencias detectadas en los métodos tradicionales con el fin de enriquecer el clustering. ***KLASS*** usa el método de *clustering basado en reglas* para hacer clasificación. El algoritmo de agrupamiento básico es el de los vecinos recíprocos enlazados, ***KLASS*** nos permite trabajar con las siguientes métricas:

- euclidiana
- euclidiana estandarizada
- χ^2 (chi-cuadrada)
- métricas mixtas, introducidas por Gibert (Gibert 1994)
- Gower (más detalles en (Gower 1971) y (Gower 1967))
- Ralambondrainy (más detalles en (Ralambondrainy 1995a))

Además, al realizar pruebas reales, se observó que en las primeras etapas de cualquier investigación, ***KLASS*** actúa más bien como una herramienta de adquisición de conocimiento que como una herramienta de simple resolver problemas.

De ahí que, ***KLASS*** cumple un doble propósito:

- Primero, implementa el método de *clustering basado en reglas*.
- Segundo, es una herramienta de apoyo a la adquisición de conocimiento basado en la combinación de bases de conocimiento con métodos estadísticos que abre el camino a la generación automática de reglas para un sistema de diagnóstico basado en conocimiento.

El método de clustering basado en reglas es particularmente útil para *dominios poco estructurados* (Gibert 1994). En general, los expertos tienen alguna información extra acerca de la estructura del dominio. Esta información declarativa puede ser usada para enriquecer el clustering automático. Esta es la base del método que se implementa en ***KLASS***.

De acuerdo con la naturaleza de los datos, δ tomará una forma u otra. En esta tesis se trabaja con *KLASS*, ver Anexo A, que implementa un algoritmo de clasificación ascendente jerárquica que se enmarca en el esquema general que se acaba de presentar. En concreto se trata del algoritmo de los vecinos recíprocos encadenados (De Rham 1997) que describimos a continuación.

A.1.1 Vecinos recíprocos encadenados

En estos contextos identificar cuáles son las parejas de elementos más próximos (o de vecinos recíprocos en este último caso) en cada iteración requiere la definición de una métrica sobre el espacio de las variables que permita calcular la distancia entre dos individuos.

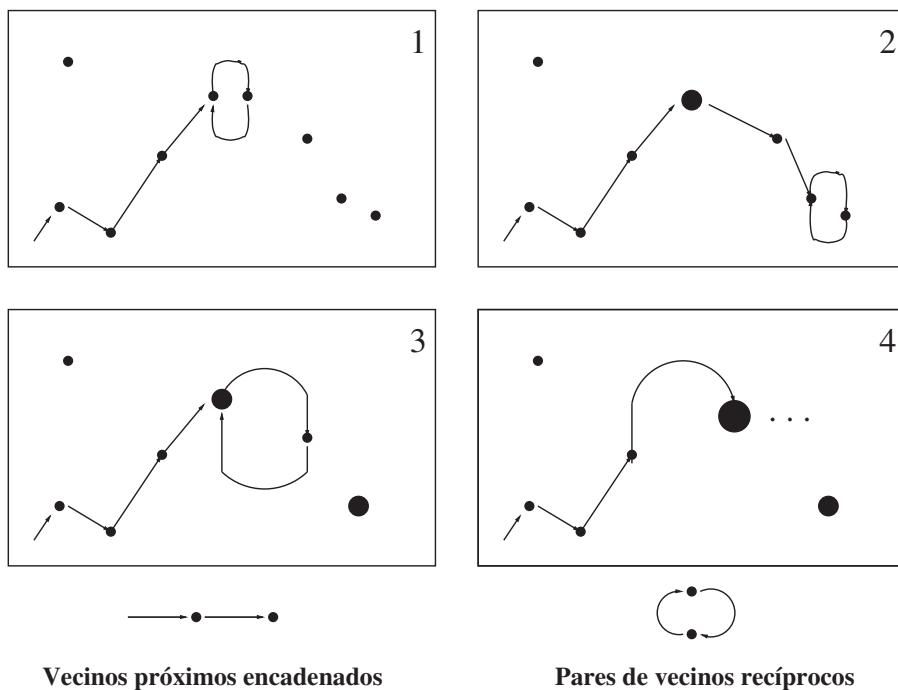


Figura A.1: El proceso de los vecinos recíprocos encadenados.

El algoritmo de los *vecinos recíprocos encadenados* utiliza un concepto propio para determinar cuáles son los individuos que se agregan:

Son vecinos recíprocos los individuos i, i' si i es el objeto más próximo a i' en la muestra, y i' es a su vez el más próximo a i . De este modo, en la clasificación por vecinos recíprocos, siempre se agregarán parejas de vecinos recíprocos.

La principal propiedad de este método es que el resultado no depende del orden como se procesan los datos (ni del orden como se producen las agregaciones) porque se está trabajando con un criterio global sobre todos los datos.

Existen muchas implementaciones del algoritmo de los *vecinos recíprocos*. En esta tesis se utiliza la de los *vecinos recíprocos encadenados* que es mas barato que otras versiones porque su complejidad es menor que $O(n^2)$.

la Figura A.1 ilustra cómo en este algoritmo se produce un encadenamiento de objetos que lleva del objeto más cercano al siguiente más cercano hasta que se bucla en un lazo. El lazo es precisamente la expresión gráfica de las parejas de vecinos recíprocos. Cuando se halla uno, se produce una agregación con la consecuente creación de una nueva clase. Es frecuente representar en forma de árbol la secuencia de agregaciones de un proceso así. Estos árboles

reciben el nombre de *dendogramas*, ver §5.3.

En estos contextos identificar cuáles son las parejas de elementos más próximos (o de vecinos recíprocos en este último caso) en cada iteración requiere la definición de una métrica sobre el espacio de los atributos que permita calcular la distancia entre dos individuos.

A.1.2 Criterio de Agregación

El criterio de agregación es aquél que selecciona en cada paso de la clasificación la pareja de puntos, de individuos o subclases, que se han de fusionar en una nueva clase (Völle 1985).

En esta tesis se han usado 2 criterios para realizar las clasificaciones de las secciones §16 y §21; el criterio del centroide (Sokal and Michener 1958) y el criterio de Ward (Ward 1963), aunque en las clasificaciones elegidas para ser interpretadas se ha usado siempre el criterio de Ward (Ward 1963) por tener mejores propiedades.

El criterio de Ward decide qué dos elementos (ya sea objetos individuales o clases) deben fusionarse cada vez basándose, no en una noción pura de distancia, sino en el concepto de inercia.

La inercia total de un conjunto de puntos \mathcal{I} respecto su centro de gravedad \bar{i} es constante.

$$M^2(\mathcal{I}/\bar{i}) = \sum_{i \in \mathcal{I}} m_i d^2(i, \bar{i}) \quad (\text{A.1})$$

siendo m_i la masa del punto i o, en nuestro contexto, el efectivo de la clase que representa.

En cada paso del proceso de clasificación se agregarán los dos elementos más homogéneos (o próximos), cosa que nos permite considerar que el aumento de inercia juega un papel de pseudodistancia entre clases.

Desde el punto de vista de los vecinos recíprocos, no hace falta más que sustituir la matriz de distancias entre clases \mathcal{D} por una matriz donde se recoge el aumento de inercia que se produciría si se efectuase la fusión de cada dos clases, y operara en la forma ordinaria. Bajo el criterio de Ward, pues, la matriz de distancias de partida sería $\mathcal{D} = (d_{ed})$, con $d_{ed} = \Delta(\mathcal{C}_e, \mathcal{C}_d)$.

En (Gibert 1994) se llega a la expresión que permite la inicialización de la matriz de pseudodistancias \mathcal{D}

$$\mathcal{D} = (d_{ed}), \text{ donde } d_{ed} = \frac{1}{2} d^2(\mathcal{C}_e, \mathcal{C}_d),$$

siendo $d^2(\mathcal{C}_e, \mathcal{C}_d)$ la distancia mixta entre los centros de gravedad de las clases \mathcal{C}_e y \mathcal{C}_d .

En un primer paso, se agregan los puntos i, i' que tengan la menor pseudodistancia entre sí. En los pasos sucesivos, se identificará la nueva agregación después de actualizar la matriz \mathcal{D} en la forma conveniente.

De hecho, en (Gibert 1994) se demuestra que el aumento de inercia que produce la fusión de $\mathcal{C} = \{\mathcal{C}_e, \mathcal{C}_d\}$ con cualquier otro elemento i se puede expresar también como función del aumento de inercia que se hubiera producido agregando i a cualquiera de los componentes de \mathcal{C} . Esta información está en la \mathcal{D} actual, tal como convenía. La recurrencia buscada es:

$$d_{\mathcal{C}_i} = \frac{1}{m} ((m_e + m_i)d_{ei} + (m_d + m_i)d_{di} - m_id_{ed}), m = m_i + m_e + m_d \quad (\text{A.2})$$

Con esta relación se actualizarán las pseudodistancias de un paso a otro, y en la matriz \mathcal{D} se tienen las cantidades que determinan directamente qué nodos se agregan en el siguiente paso.

A.1.3 El representante de la clase

KLASS representa las clases que se van formando a través de un elemento *prototipo*. La idea de prototipo \bar{i}_C de un conjunto de individuos $C = \{i_1 \dots i_{n_C}\} \subseteq \mathcal{I}$ lleva a pensar en un individuo, real o no, que sintetice las características de sus representantes.

La explicitación de un elemento que represente prototípicamente cada una de las clases formadas por **Klass** permite:

- Por un lado, tratar de forma homogénea individuos propiamente dichos y las clases, las cuales se identificaran con su representante.
- Proporcionar una descripción conceptual (o prototípica) de las clases formadas, lo que va a ser fundamental en el proceso de interpretar los resultados finales desde un punto de vista cognitivo y ver si las clases encontradas tienen *significado* o no.

Considerese la descripción de \bar{i}_C , $\bar{x}_C = (\bar{x}_{C1} \dots \bar{x}_{CK})$. El cálculo de \bar{x}_C se hace componente a componente, y hace falta estudiar por separado el caso de las componentes numéricas y el de las categóricas.

Para variables numéricas, en (Gibert 1994) justifica que este representante sea la media aritmética de los componentes de la clase, lo que coincide con el centro de gravedad de la clase en esa variable.

Para variables categóricas, la propuesta de (Gibert 1994) es el *objeto extendido* que se define a continuación.

A.1.4 Objetos compactos y objetos extendidos

Dada una clase $C = \{i_1 \dots i_{n_C}\}$ y la variable cualitativa X_k que toma valores en el conjunto $\mathcal{D}_k = \{c_1^k \dots c_{n_k}^k\}$, la k -ésima componente del representante de C , \bar{i}_C es:

$$\bar{x}_{Ck} = ((f_C^{k_1} \quad c_1^k) \dots (f_C^{k_{n_k}} \quad c_{n_k}^k)), \quad f_C^{k_j} = \frac{\text{card}\{i \in C : x_{ik} = c_j^k\}}{n_C}$$

No hace falta decir que si X_k toma un mismo valor en toda la clase (c_s^k), el representante de clase será también $\bar{x}_{Ck} = c_s^k$ (lo que sería equivalente a $\bar{x}_{Ck} = ((0 \ c_1^k) \dots (1 \ c_s^k) \dots (0 \ c_{n_k}^k))$).

La propuesta es representar al prototipo de una variable categórica detallando la forma cómo los elementos de la clase se distribuyen en las diferentes modalidades de la variable, y hacerlo mediante proporciones por razones técnicas. En la referencia original aparece una descripción más detallada de por qué se hace de esta manera.

Nombrando $I_C^{k_j}$ al número de individuos de la clase C que toman valor c_j^k para la variable X_k , se pueden denotar las componentes del vector \bar{x}_{Ck} como

$$f_C^{k_j} = \frac{I_C^{k_j}}{\sum_{j=1}^{n_k} I_C^{k_j}} = \frac{I_C^{k_j}}{n_C}$$

A partir de ahora, para simplificar la notación, se representaran las componentes cualitativas del centro de gravedad de una forma equivalente a la anterior:

$$\bar{x}_{Ck} = ((f_C^{k_1} \quad c_1^k) \dots (f_C^{k_{n_k}} \quad c_{n_k}^k)) \equiv (f_C^{k_1} \dots f_C^{k_{n_k}})$$

Con esta definición de prototipo de una clase para las variables cualitativas, utilizada por primera vez en (Gibert 1991), aparece un nuevo tipo de objeto que puede tener componentes vectoriales en algunas variables categóricas.

Un objeto cualquiera i puede tener, para la variable categórica X_k , un valor del tipo $(f_i^{k_1}, \dots, f_i^{k_{n_k}})$ donde $f_i^{k_j}, (j = 1 : n_k)$ es la proporción de objetos elementales de la clase representada por i sobre la categoría c_j^k . Bajo esta representación, los valores posibles de las variables categóricas serán un subconjunto de $\mathcal{D}_k \cup [0, 1]^{n_k}$, dado que las $f_i^{k_j}$ son proporciones, y siempre satisfacen

1. $f_i^{k_j} \geq 0, \forall i$
2. $\sum_{j=1}^{n_k} f_i^{k_j} = 1$

Por tanto, en (Gibert 1991) se define como *valor en forma extendida* a todo valor vectorial de una variable cualitativa, y *valor en forma compacta* al que es representable mediante un símbolo.

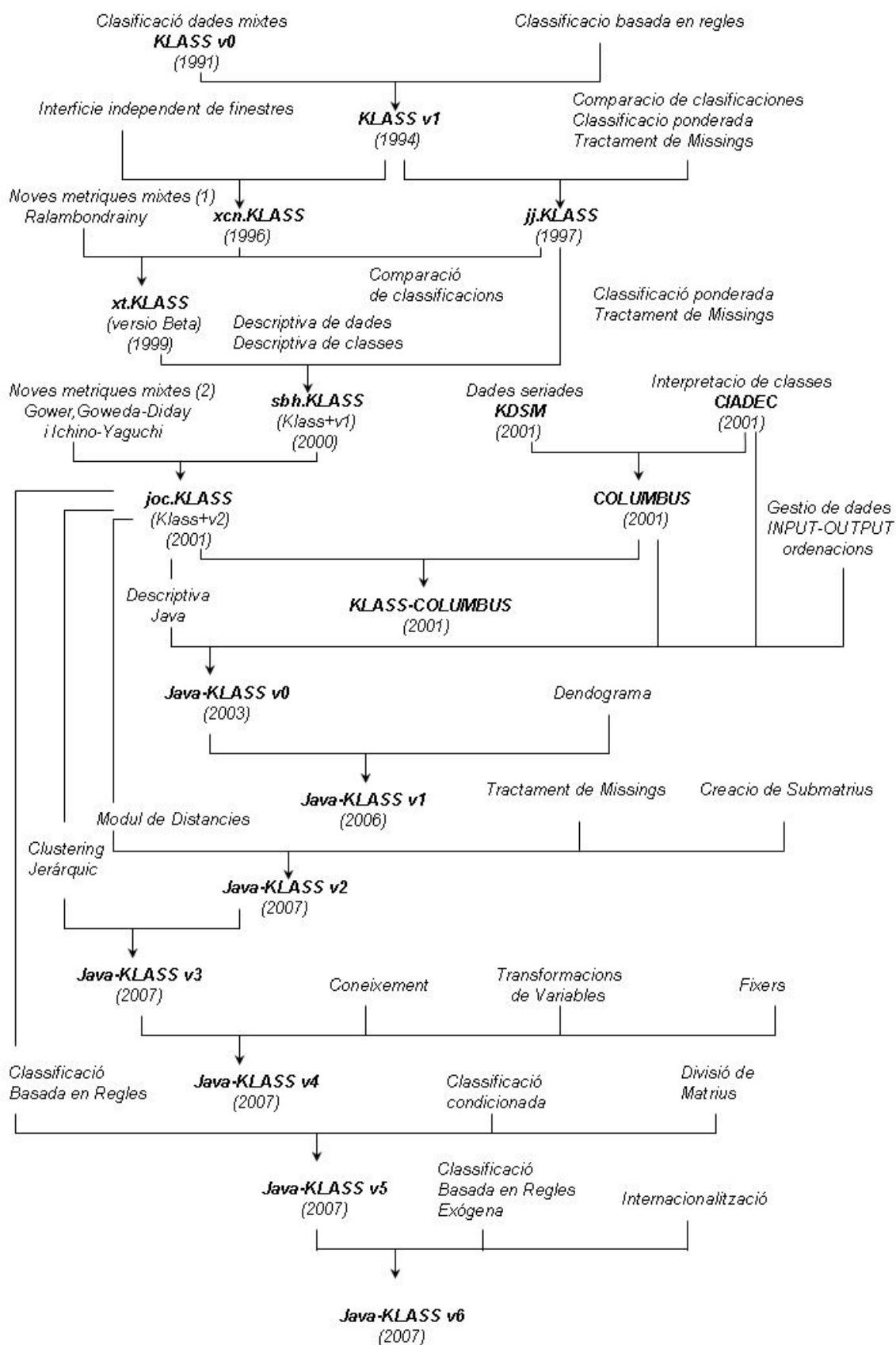
Se considera *objeto compacto* a todo aquél que tiene valores compactos en todas sus variables categóricas, mientras que los *objetos extendidos* presentan valores extendidos al menos en una de las variables categóricas.

A.2 Evolución de la plataforma KLASS

- Feb. 1991 **KLASS v0**. Tesina Karina Gibert. “KLASS. Estudi d'un sistema d'ajuda al tractament estadístic de grans bases de dades”. Clasifica matrices de datos heterogéneas con la distancia mixta. (Gibert 1991)
- Nov. 1994 **KLASS v1**. Tesis Karina Gibert. “L'ús de la informació simbòlica en l'automatització del tractament estadístic de dominis poc estructurats”. Es una ampliación de **KLASS v0**. Incorpora la clasificación basada en reglas. (Gibert 1994)
- Jul. 1996 **KLASS v1.1**. PFC Xavier Castillejo. Incorpora a **KLASS.v1** una interfaz de ventanas independiente con un sistema que facilita el uso de KLASS desde SUN y desde PC a usuarios que desconocen Lisp y UNIX. Denominaremos **xcn.KLASS** al núcleo Lisp de esta nueva versión y **xcn.i** a la interfaz C. (Castillejo 1996)
- Oct. 1997 **jj.KLASS**. PFC Juan José Marquez y Juan Carlos Martín. Incorpora a la versión **KLASS.v1** nuevas opciones para el tratamiento de missings, la posibilidad de trabajar con objetos ponderados e implementa un test no paramétrico de comparación de clasificaciones (Márquez and Martín 1997).
- Set. 1999 **KLASS v1.2**. PFC Xavier Tubau (versión β). Incorpora a la versión **xcn.KLASS** el módulo de comparación de clasificaciones de **jj.KLASS**, la métrica mixta de Ralambondrainy (Ralambondrainy 1995a) (Ralambondrainy 1988) y prepara la formulación de tres más para su posterior implementación. Denominaremos **xt.KLASS** al núcleo Lisp de esta nueva versión y **xt.i** a la interfaz C asociada. (Tubau 1999)
- 1999-2000 **KLASS+ v1**. PFC Sílvia Bayona. Fusión definitiva de la versión **xt.KLASS** y **jj.KLASS**. Incorpora además un módulo nuevo de análisis descriptiva de datos, así como de las clases resultantes, reorientando KLASS hacia un propósito más general y menos especializado. Denominaremos **sbh.KLASS** al núcleo Lisp de esta nueva versión y **sbh.i** a la interfaz C asociada. (Bayona 2000)
- 2000-2002 **KLASS+ v2**. PFC Josep Oliveras. Agrega a **sbh.KLASS** las métricas mixtas pendientes (Gower (Gower 1971) (Gower 1966) (Gower 1967), Gowda-Diday (Gowda and Diday 1992) (Diday and Gowda 1991) e Ichino-Yaguchi (Ichino and Yaguchi

1994) (Ichino and Yaguchi 1989)). Denominaremos **joc.KLASS** a esta nueva versión. (Oliveras Castellà, Josep)

- 2000-2003 **jr.KLASS+**. Tesis doctoral Jorge Rodas. Integra **KLASS+ v.2** y **Columbus**, que se presenta más adelante. (Rodas Osollo 2003)
- 2000-2003 Investigación Anna Salvador y Fernando Vázquez. Desarrollo de **CIADEC**, que se presenta más adelante. (Gibert and Salvador 2000) (Vázquez and Gibert 2002)
- 2002-2003 **Java-KLASS v0**. PFC M^a del Mar Colillas. Versión Java del módulo de análisis descriptiva e integración con **CIADEC** y **Columbus**.
 - 2003-2005 **Java-KLASS v0.22**. Colaboración con Mar Colillas. Ampliación del módulo de análisis descriptiva e introducción de herramientas de gestión de datos (*definición de ordenaciones en los informes, posibilidad de varias matrices de objetos en el sistema simultáneamente, cambio de matriz activa*).
 - 2005-2006 **Java-KLASS v1.0**. Colaboración con Mar Colillas. Incluye la lectura y visualización de dendogramas aislados, así como la generación de particiones a partir de ellos.
- 2006-2007 **Java-KLASS v2.0**. PFC Jose Ignacio Mateos. Ampliación de **Java-KLASS** con un módulo de cálculo de distancias para diferentes tipos de matrices de datos, incluyendo las que combinan información cualitativa y cuantitativa, tratamiento de missings y creación de submatrices.
- 2006-2007 **Java-KLASS v3.0**. PFC Roberto Tuda. Incluye un módulo de clasificación automática por métodos jerárquicos, utilizando todas las distancias implementadas a la v2.0 y una opción para estudiar agregaciones de objetos paso a paso. Se crea la opción de poder seleccionar el directorio de trabajo por defecto. Se le agrega la opción de añadir y grabar objetos con peso.
- 2006-2007 **Java-KLASS v4.0**. PFC Laia Riera Guerra. Introducción, gestión y evaluación de Bases de Conocimiento. Ampliación de **Java-KLASS** con un módulo de transformación de variables que permite discretizaciones, recodificaciones y cálculos aritméticos con variables numéricas. Por último, esta versión incluye la definición de submatrices via filtros lógicos sobre los objetos, la edición de metainformación de las variables de la matriz, eliminación de variables e importación de ficheros en formato .dat estándar.
- 2007 **Java-KLASS v5.0**. PFC Andreu Raya. Incluye la clasificación condicionada, la clasificación basada en reglas y funcionalidades de división de la base de Datos y de gestión de árboles de clasificación (*o dendogramas*) asociados a las diferentes matrices de datos.
- 2007 **Java-KLASS v6.0**. Trabajo Investigación Tutelada Alejandro García. Clasificación basada en reglas exógena. Internacionalización y Localización a tres idiomas (Catalán, Inglés, Castellano). Fusión de matrices
- 2008. Tesis doctoral Alejandra Pérez. Caracterización por condicionamientos sucesivos, metodología que induce automáticamente conceptos asociados a las clases descubiertas
- 2008. Tesis doctoral Gustavo Rodríguez. Clasificación basada en reglas por estados que permite análisis de sistemas dinámicos.



A.2.1 Satélites de KLASS

Columbus es un satélite de **KLASS+** desarrollado en Java (también existe una versión en C) para implementar el método KDSM ('Knowledge Discovery in Serial Measures'). Es una herramienta para el descubrimiento de conocimiento en dominios poco estructurados que contienen series repetidas de medidas.

Las tareas más importantes implementadas por **Columbus** son:

- Pretratamiento de los datos para identificar los paquetes de series repetidas.
- Representación gráfica de las series repetidas y de las series cruzadas con una clasificación.
- Conexión con **KLASS+** para realizar tareas de cluster jerárquico y clasificación basada en reglas que forman parte de la metodología KDSM. La comunicación se realiza mediante una interfaz gráfica.
- Interpretación de clases utilizando variables activas e ilustrativas y módulos de inducción de reglas.

De acuerdo con la filosofía **KLASS** los dos satélites realizan la producción de gráficos generando código **LATEX** y conectando con el visualizador por DVI.

A.3 Aplicaciones Clustering basado en reglas

Los dominios en los que esta metodología se ha aplicado y ha producido resultados satisfactorios son variados. A continuación se introducen brevemente algunos de los casos en que se ha trabajado. En todos ellos se dan las características de los *dominios poco estructurados* y, en todos, la clasificación sin utilizar la base de conocimiento *a priori* producía al menos dos clases de significado dudoso, de difícil, por no decir imposible, interpretación:

1. La clasificación de esponjas de mar (Gibert 1994), cuya taxonomía es motivo de controversia entre los espongiólogos. La ubicación de ciertas especies según la *clasificación basada en reglas* fue objeto de discusión, proporcionando argumentos objetivos para darles género.
2. La identificación de poblaciones estelares (Gibert, Hernández, and Cortés 1996). Se trabajó con datos de estrellas de la Vía Láctea tomados de la base de datos del satélite Hipparcos; los resultados mejoraban sensiblemente usando conocimiento sobre el *halo* y el *disco* de la Galaxia.
3. Disfunciones de la tiroides (Gibert and Sonicki 1999). Se trabajó con datos de pacientes de un hospital croata y se vió cómo introducir conocimiento parcial sobre los diagnósticos clásicos producía una subdivisión más específica de utilidad clínica. Con estos mismos datos se vió cómo codificar todas las variables y repetir el análisis utilizando solamente variables cualitativas producía una sensible pérdida de información no deseable (Gibert, Sonicki, and Martín 2001), lo que para nosotros supone un argumento fuerte en contra de esta práctica.
4. Plantas depuradoras de aguas residuales (Gibert and Salvador 2000), (Gibert, Rodríguez-Roda, and Cortés 2004). Con datos de distintas plantas del territorio catalán de distinta estructura y función, utilizar información sobre el estado en que tiene que estar el agua que se ha de verter al río, permite identificar con claridad las situaciones más características que se operan en la planta, lo que contribuye a facilitar el control de la misma.

5. Discapacidades en ancianos (Annichiarico and Gibert 2004). Con pacientes de un hospital de Roma se ha pasado un nuevo test diseñado por la OMS para medir la discapacidad de un individuo. La *clasificación basada en reglas* ha permitido realizar una propuesta de ontología para la discapacidad, que no estaba todavía establecida, y que ha resultado obedecer a criterios funcionales, mas que diagnósticos, lo que resulta muy beneficioso para la concepción geriátrica del paciente.
6. Comportamiento urbanístico de municipios del área metropolitana de Barcelona a partir de las viviendas construidas de diversos tipos. Utilizar información sobre la política de protección oficial ha permitido identificar áreas de crecimiento. Con estos datos (Gibert, Nonell, Velarde, and Colillas 2005), entre todas las métricas mixtas accesibles en **KLASS** la que producía clases más fácilmente interpretables era la *métrica mixta* (Gibert and Cortés 1997).
7. Identificación de perfiles en poblaciones dependientes con trastorno mental grave y discapacidad intelectual. El uso de la *clasificación basada en reglas* con información experta sobre las diferencias entre las personas más comprometidas y las más autónomas en estos colectivos ha permitido definir patrones de dependencia que dan soporte al sistema de ayudas económicas legislado por la nueva *Ley de la Dependencia*, que ha entrado en vigor recientemente.

Anexo B

Análisis de casos, BbD revisado

B.1 Introducción

En este anexo presentamos el análisis de los 12 casos restantes que hemos detectado.

Por comodidad y limpieza de los sistemas de reglas hemos omitido escribir las reglas no efectivas en los sistemas de reglas completos y reducidos, es por eso que a partir de ahora los sistemas de reglas completos no mostrarán las reglas con $p_{sc} = 0$ ó $\mathbf{p}_{sc} = 0$, salvo que sea necesario para evidenciar la aparición de una regla no efectiva cuando se ha realizado la revisión del patrón de los intervalos.

Debido a que es fácil deducir, a partir de los sistemas de reglas completos y reducidos, los sistemas de reglas efectivas y los sistemas de reglas seguras, también limitaremos el análisis solo a los sistemas de reglas completos, y en algunos casos, a los sistemas de reglas reducidos y de reglas seguras.

B.2 Análisis por caso

B.2.1 Caso 2, $M_{C_i}^k < m_{C_j}^k$

El caso 2, que es el simétrico del caso 1. Se presenta como otro de los casos extremos. la Figura B.1 presenta un boxplot múltiple que corresponde a esta situación, junto al sistema de intervalos propuesto en este trabajo.

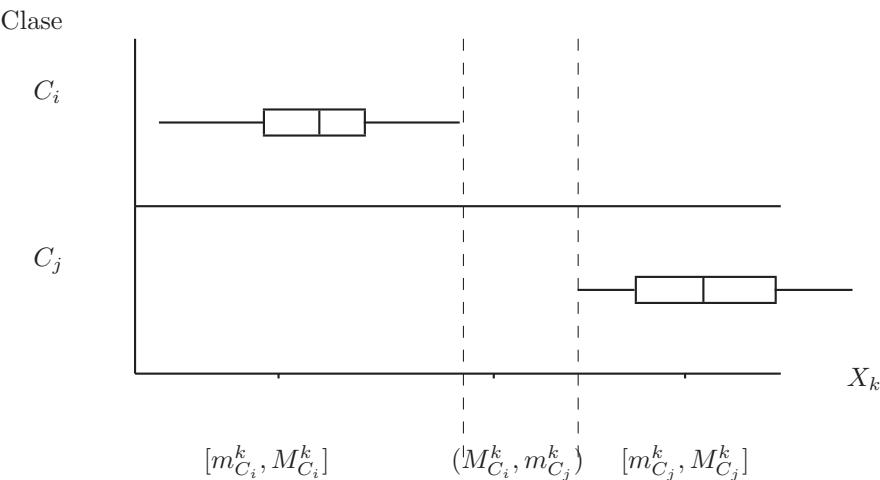


Figura B.1: Caso 2: Boxplot based Discretization.

Por un razonamiento parecido al del caso anterior (Caso 1), en las Tablas B.1 y B.2 se pueden ver los sistemas de intervalos y reglas generados tanto con la propuesta original como con la revisada para el sistema de reglas completo y el sistema de reglas reducido.

En nuestra propuesta al cerrar los intervalos $I_1^{k,2}$ y $I_3^{k,2}$ en sus dos extremos y dejamos abierto el intervalo $I_2^{k,2}$ por ambos lados conseguimos que el intervalo del centro $I_2^{k,2} = \emptyset$, dando lugar a dos reglas de probabilidad 0. Con ello tenemos un sistema de reglas reducido más compacto que la propuesta de (Vázquez and Gibert 2001) con sólo 2 reglas de probabilidad 1 que separan claramente las 2 clases.

Al igual que el caso 1, esto identifica una variable totalmente caracterizadora.

En la Figura B.1 se ha indicado cómo se formularían los intervalos bajo esta propuesta.

En el caso de la propuesta original el sistema de reglas reducido es igual al sistema de reglas seguras e igual al sistema de reglas efectivas:

$$\Re^*(X_k, \mathcal{P}_2) = \mathcal{S}(X_k, \mathcal{P}_2) = \Re(X_k, \mathcal{P}_2)$$

Pero en el caso de la propuesta revisada se cumple que el sistema de reglas efectivas es igual al sistema de reglas seguras:

$$\Re(X_k, \mathcal{P}_2) = \mathcal{S}(X_k, \mathcal{P}_2)$$

Y como en el caso anterior obtenemos un sistema de reglas efectivas y también un sistema de reglas seguras mas compacto que en la propuesta original.

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\Re(X_k, \mathcal{P}_2)$
Previa	$I_1^{k,2} = [m_{C_i}^k, M_{C_i}^k]$ $I_2^{k,2} = (M_{C_i}^k, m_{C_j}^k]$ $I_3^{k,2} = (m_{C_j}^k, M_{C_j}^k]$	$r_{1,i}^k : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1} i \in C_i$ $r_{1,j}^k : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=0} i \in C_j$ $r_{2,j}^k : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}=1} i \in C_j$ $r_{2,i}^k : x_{ik} \in I_2^{k,2} \xrightarrow{p_{1i}=0} i \in C_i$ $r_{3,j}^k : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3j}=1} i \in C_j$ $r_{3,i}^k : x_{ik} \in I_3^{k,2} \xrightarrow{p_{1i}=0} i \in C_i$
Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\Re^*(X_k, \mathcal{P}_2)$
Previa	$I_1^{k,2} = [m_{C_i}^k, M_{C_i}^k]$ $I_2^{k,2} = (M_{C_i}^k, m_{C_j}^k]$ $I_3^{k,2} = (m_{C_j}^k, M_{C_j}^k]$	$r_{1,i}^k : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1} i \in C_i$ $r_{2,j}^k : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}=1} i \in C_j$ $r_{3,j}^k : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3j}=1} i \in C_j$

Tabla B.1: Caso 2: Relación entre Intervalos y Reglas (Propuesta original).

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathcal{R}(X_k, \mathcal{P}_2)$
Revisada	$I_1^{k,2} = [m_{C_i}^k, M_{C_i}^k]$	$r_{1,i}^k : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1} i \in C_i$
	$I_2^{k,2} = (M_{C_i}^k, m_{C_j}^k)$	$r_{1,j}^k : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=0} i \in C_j$
	$I_3^{k,2} = [m_{C_j}^k, M_{C_j}^k]$	$r_{2,j}^k : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2c}=0} i \in C_j$
Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathcal{R}^*(X_k, \mathcal{P}_2)$
Revisada	$I_1^{k,2} = [m_{C_i}^k, M_{C_i}^k]$ $I_2^{k,2} = (M_{C_i}^k, m_{C_j}^k)$ $I_3^{k,2} = [m_{C_j}^k, M_{C_j}^k]$	$r_{1,i}^k : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1} i \in C_i$ $r_{2,j}^k : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2c}=0} i \in C_j$ $r_{3,j}^k : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3j}=1} i \in C_j$

Tabla B.2: Caso 2: Relación entre Intervalos y Reglas (Propuesta revisada).

B.2.2 Caso 3, $m_{C_i}^k = m_{C_j}^k \wedge M_{C_i}^k = M_{C_j}^k$

El caso 3, se presenta como otro de los casos extremos. la Figura B.2 presenta un boxplot múltiple que corresponde a esta situación. En este caso coinciden los mínimos y máximos de ambas clases para la variable en estudio, es decir los rangos son iguales:

$$r_k^{C_i} = [m_{C_i}^k, M_{C_i}^k] ; r_k^{C_j} = [m_{C_j}^k, M_{C_j}^k] \text{ y } r_k^{C_i} = r_k^{C_j}$$

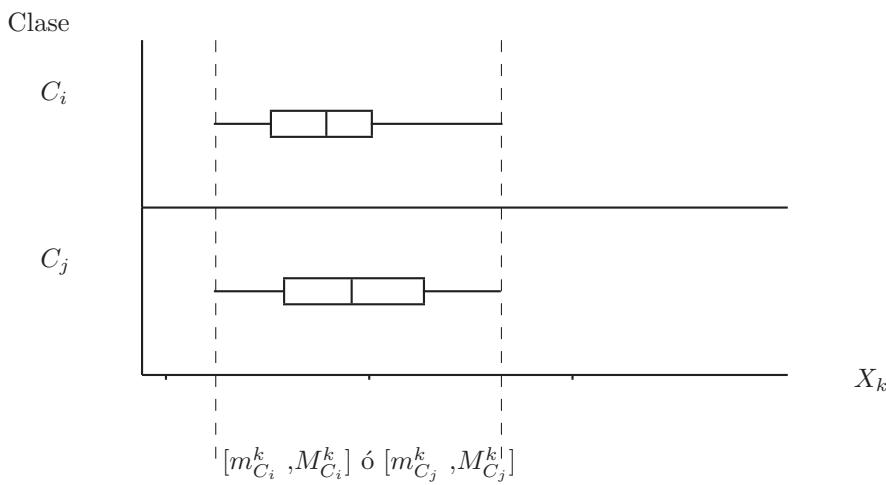


Figura B.2: Caso 3: Boxplot based Discretization.

Según la propuesta presentada en (Vázquez and Gibert 2001) y (Vázquez and Gibert 2002), el sistema de intervalos inducido por \mathcal{P}_2 sobre X_k sería:

$$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k] \text{ o bien } I_1^{k,2} = [m_{C_j}^k, m_{C_i}^k]$$

$$I_2^{k,2} = (m_{C_i}^k, M_{C_j}^k] \text{ o bien } I_2^{k,2} = (m_{C_i}^k, M_{C_i}^k] \text{ ó } I_2^{k,2} = (m_{C_j}^k, M_{C_j}^k] \text{ ó } I_2^{k,2} = (m_{C_j}^k, M_{C_i}^k]$$

$$I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k] \text{ o bien } I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k].$$

El intervalo 3 es un intervalo vacío que genera solamente reglas no efectivas.

Por ser $m_{C_j}^k = m_{C_i}^k$, $I_1^{k,2}$ tiene un único punto $m_{C_i}^k$, y como $M_{C_j}^k = M_{C_i}^k$, el intervalo $I_3^{k,2} = \emptyset$, siempre, lo que genera un sistema de reglas completo $\mathfrak{R}(X_k, \mathcal{P}_2)$, con 2 pares de reglas con probabilidades complementarias (2 reglas por cada intervalo $I_1^{k,2}$ e $I_2^{k,2}$), como se muestra a continuación:

$$\begin{aligned} \mathfrak{R}(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathfrak{p}_{1j}} i \in C_j, \\ & r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathfrak{p}_{1i}=1-\mathfrak{p}_{1j}} i \in C_i, \\ & r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathfrak{p}_{2i}} i \in C_i, \\ & r_4 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathfrak{p}_{2j}=1-\mathfrak{p}_{2i}} i \in C_j \quad \} \end{aligned}$$

Este es el primer caso que se plantea donde el sistema de reglas reducido no es idéntico a $\mathfrak{R}(X_k, \mathcal{P}_2)$. El sistema de reglas reducido lo formarían las reglas que presenten una mayor probabilidad para cada intervalo. Es decir, para el caso que existan reglas complementarias, se eliminará la de menor probabilidad y será parte del sistema de reglas reducido la de mayor probabilidad.

Entonces el sistema de reglas reducido sería:

$$\begin{aligned} \mathfrak{R}^*(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\max\{\mathfrak{p}_{1j}, \mathfrak{p}_{1i}\}} i \in C, \quad \text{siendo, } \mathcal{C} = \begin{cases} C_i & \text{si } \mathfrak{p}_{s1} \geq \mathfrak{p}_{s2} \\ C_j & \text{si } \mathfrak{p}_{s1} < \mathfrak{p}_{s2} \end{cases} \\ & r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{\max\{\mathfrak{p}_{2j}, \mathfrak{p}_{2i}\}} i \in C \quad \} \end{aligned}$$

Nuestra propuesta es redefinir el sistema de intervalos de la siguiente forma:

$$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k) \text{ o bien } I_1^{k,2} = [m_{C_j}^k, m_{C_i}^k)$$

$$I_2^{k,2} = [m_{C_i}^k, M_{C_j}^k] \text{ o bien } I_2^{k,2} = [m_{C_i}^k, M_{C_i}^k] \text{ ó } I_2^{k,2} = [m_{C_j}^k, M_{C_j}^k] \text{ ó } I_2^{k,2} = [m_{C_j}^k, M_{C_i}^k]$$

$$I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k] \text{ o bien } I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$$

Ahora los intervalos 2 y 3 son un intervalo vacío que genera solamente reglas no efectivas. En la Figura B.2 se ha indicado cómo se formularán los intervalos bajo esta propuesta.

Si cerramos, a la derecha y a la izquierda el intervalo $I_2^{k,2}$, generamos un único intervalo no vacío, ya que al ser $m_{C_i}^k = m_{C_j}^k$, $I_1^{k,2} = \emptyset$ y $M_{C_j}^k = M_{C_i}^k$, $I_3^{k,2} = \emptyset$.

No existen reglas de probabilidad 1, en este caso. Sin embargo debemos mencionar que existen 2 reglas con probabilidades complementarias en el sistema de reglas.

A partir de aquí el sistema de reglas inducido por \mathcal{P}_2 sobre X_k sería:

$$\begin{aligned} \mathcal{R}(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathfrak{p}_{2j}} i \in C_j, \\ & r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathfrak{p}_{2i}=1-\mathfrak{p}_{2j}} i \in C_i \quad \} \end{aligned}$$

puesto que $I_1^{k,2}$ e $I_3^{k,2}$ por ser vacíos generan reglas de probabilidad 0.

Si queremos un sistema de reglas reducido este lo formaría una única regla y sería la regla que presente una mayor probabilidad. Es decir, para este caso, el total de objetos será asignado a la clase que tenga el mayor número de elementos, ya que es ésta la que generará la probabilidad más grande.

$$\mathcal{R}^*(X_k, \mathcal{P}_2) = \{r_1 : x_{ik} \in I_2^{k,2} \xrightarrow{\max\{p_{2i}, p_{2j}\}} i \in C\} \quad \text{siendo, } \mathcal{C} = \begin{cases} C_i & \text{si } p_{2i} \geq p_{2j} \\ C_j & \text{si } p_{2i} < p_{2j} \end{cases}$$

Por lo tanto no tenemos ninguna regla de probabilidad 1 que separe claramente las 2 clases, ni siquiera parcialmente. Esto dará lugar a una variable claramente no caracterizadora.

B.2.3 Caso 4, $m_{C_i}^k > m_{C_j}^k \wedge M_{C_i}^k > M_{C_j}^k \wedge m_{C_i}^k < M_{C_j}^k$

El caso 4, se presenta como uno de los casos que podríamos llamar comunes. la Figura B.3 presenta un boxplot múltiple que corresponde a esta situación.

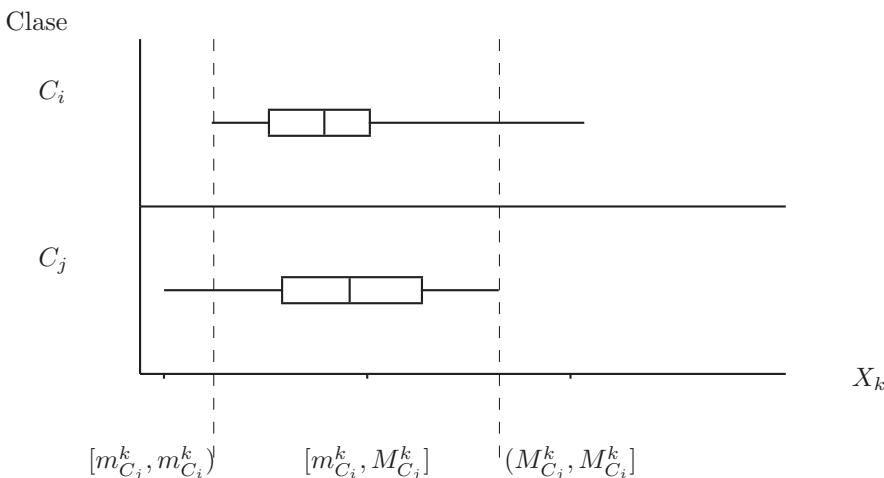


Figura B.3: Caso 4: Boxplot based Discretization.

La Propuesta de (Vázquez and Gibert 2001) para este caso da lugar a un sistema de intervalos I^k de la siguiente forma:

$$I_1^{k,2} = [m_{C_j}^k, m_{C_i}^k]$$

$$I_2^{k,2} = (m_{C_i}^k, M_{C_j}^k]$$

$$I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$$

Lo que genera el siguiente sistemas de reglas completo $\mathfrak{R}(X_k, \mathcal{P}_2)$:

$$\begin{aligned} \mathfrak{R}(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}} i \in C_i, \\ & r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}=1-p_{1i}} i \in C_j, \\ & r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}} i \in C_j, \\ & r_4 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i, \\ & r_5 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3i}=1} i \in C_i \} \end{aligned}$$

Y un sistema de reglas reducido $\mathfrak{R}^*(X_k, \mathcal{P}_2)$ de la siguiente forma:

$$\begin{aligned}\mathfrak{R}^*(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\max\{\mathbf{p}_{1j}, \mathbf{p}_{1i}\}} i \in C, \\ & r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{\max\{\mathbf{p}_{2j}, \mathbf{p}_{2i}\}} i \in C, \quad \text{siendo, } \mathcal{C} = \begin{cases} C_i & \text{si } p_{2i} \geq p_{2j} \\ C_j & \text{si } p_{2i} < p_{2j} \end{cases} \\ & r_3 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathbf{p}_{3i}=1} i \in C_i \quad \}\end{aligned}$$

Nuestra propuesta es redefinir el sistema de intervalos I^k de la siguiente forma:

$$I_1^{k,2} = [m_{C_j}^k, m_{C_i}^k)$$

$$I_2^{k,2} = [m_{C_i}^k, M_{C_j}^k]$$

$$I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$$

Al cerrar el intervalo del centro $I_2^{k,2}$ por ambos lados y dejar abierto por la derecha $I_1^{k,2}$ y abierto por la izquierda $I_3^{k,2}$ el sistema de reglas $\mathcal{R}(X_k, \mathcal{P}_2)$ inducido por \mathcal{P}_2 sobre X_k sería:

$$\begin{aligned}\mathcal{R}(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathbf{p}_{1j}=1} i \in C_j, \\ & r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathbf{p}_{2j}} i \in C_j, \\ & r_4 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathbf{p}_{2i}=\mathbf{p}_{2j}} i \in C_i, \\ & r_5 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathbf{p}_{3i}=1} i \in C_i \quad \}\end{aligned}$$

Solamente moviendo 2 puntos de los intervalos extremos al del centro aparecen 2 reglas de probabilidad 1 que separan parcialmente las 2 clases, con lo que obtenemos un sistema de reglas más compacto y aumentamos el número de reglas seguras, lo que además parece mejor modelo de lo que está mostrando el boxplot múltiple.

Queda claro que según cómo se definan los límites de los intervalos de I^k , el sistema de reglas inducido a partir de él producirá más o menos reglas seguras.

Y el nuevo sistema de reglas reducido $\mathcal{R}^*(X_k, \mathcal{P}_2)$ quedaría de la siguiente forma:

$$\begin{aligned}\mathcal{R}^*(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathbf{p}_{1j}=1} i \in C_j, \\ & r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{\max\{\mathbf{p}_{2j}, \mathbf{p}_{2i}\}} i \in C, \quad \text{siendo, } \mathcal{C} = \begin{cases} C_i & \text{si } p_{2i} \geq p_{2j} \\ C_j & \text{si } p_{2i} < p_{2j} \end{cases} \\ & r_3 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathbf{p}_{3i}=1} i \in C_i \quad \}\end{aligned}$$

Y obtendríamos un sistema de reglas seguras $\mathcal{S}(X_k, \mathcal{P}_2)$ como se muestra a continuación:

$$\begin{aligned}\mathcal{S}(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathbf{p}_{1j}=1} i \in C_j, \\ & r_3 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathbf{p}_{3i}=1} i \in C_i \quad \}\end{aligned}$$

Esto identifica una variable parcialmente caracterizadora. En la Figura B.3 se ha indicado cómo se formularán los intervalos bajo esta propuesta.

B.2.4 Caso 5, $m_{C_i}^k < m_{C_j}^k \wedge M_{C_i}^k < M_{C_j}^k \wedge m_{C_j}^k < M_{C_i}^k$

El caso 5, se presenta como otro de los casos que podríamos llamar comunes. Este caso es simétrico del caso 4. la Figura B.4 presenta un boxplot múltiple que corresponde a esta situación.

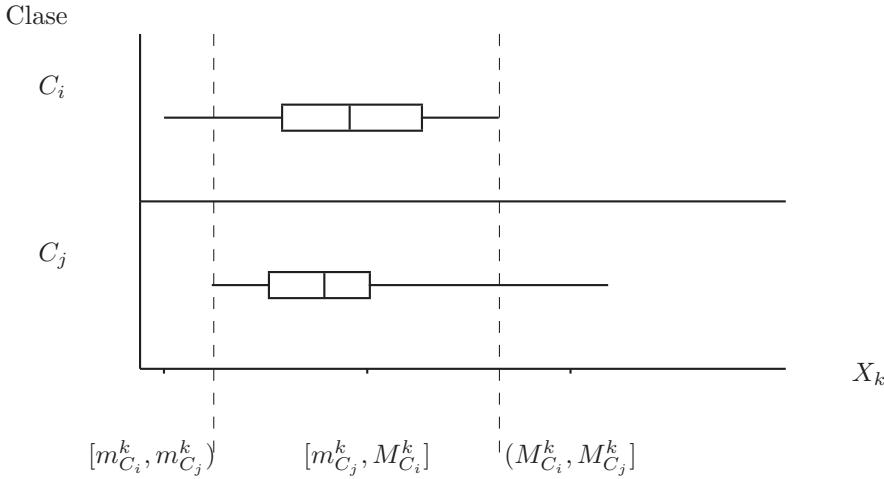


Figura B.4: Caso 5: Boxplot based Discretization.

Entonces, por un razonamiento parecido al anterior, en las Tablas B.3 y B.4 se pueden ver los sistemas de intervalos y reglas generados con ambas propuestas metodológicas.

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathfrak{R}(X_k, \mathcal{P}_2)$
Previa	$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k]$ $I_2^{k,2} = (m_{C_j}^k, M_{C_i}^k]$ $I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k]$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}} i \in C_j$ $r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1-p_{1j}} i \in C_i$ $r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}} i \in C_j$ $r_4 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i$ $r_5 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3j}=1} i \in C_j$

Tabla B.3: Caso 5: Relación entre Intervalos y Reglas (Propuesta original).

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathcal{R}(X_k, \mathcal{P}_2)$
Revisada	$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k)$ $I_2^{k,2} = [m_{C_j}^k, M_{C_i}^k]$ $I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k]$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1} i \in C_i$ $r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}} i \in C_j$ $r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i$ $r_4 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3j}=1} i \in C_j$

Tabla B.4: Caso 5: Relación entre Intervalos y Reglas (Propuesta revisada).

Si, para nuestra propuesta, Tabla B.4, cerramos el intervalo del centro $I_2^{k,2}$ por ambos

lados y dejamos abierto por la derecha $I_1^{k,2}$ y abierto por la izquierda $I_3^{k,2}$ obtenemos un sistema de reglas más compacto y aumentamos el número de reglas seguras que separan parcialmente las 2 clases.

Esto identifica una variable parcialmente caracterizadora. En la Figura B.4 se ha indicado cómo se formularán los intervalos bajo esta propuesta.

B.2.5 Caso 6, $m_{C_i}^k < m_{C_j}^k \wedge M_{C_i}^k > M_{C_j}^k$

El caso 6, se presenta como otro de los casos comunes. El rango de $X_k | C_2$ está totalmente incluído en el rango de $X_k | C_1$ lo que implica que el intervalo del centro contiene totalmente a la clase C_2 y solapa con C_1 . la Figura B.5 presenta un boxplot múltiple que corresponde a esta situación.

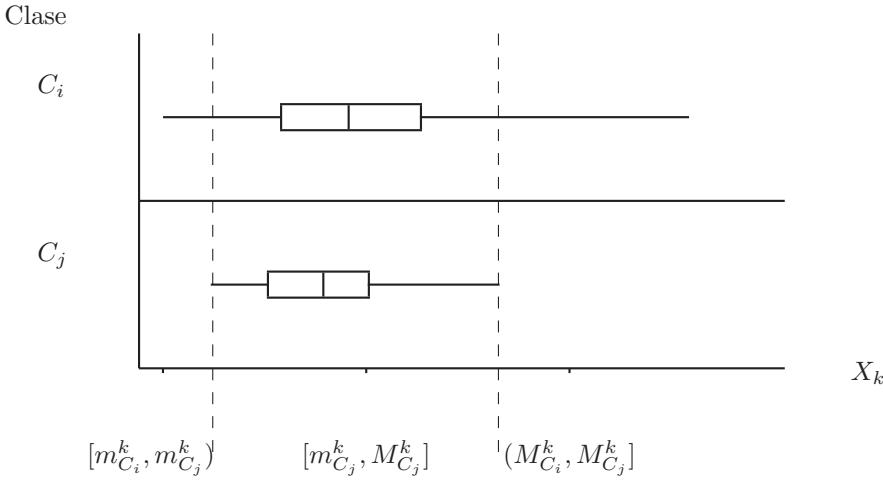


Figura B.5: Caso 6: Boxplot based Discretization.

Entonces, por un razonamiento parecido al anterior, en las Tablas B.5 y B.6 se pueden ver los sistemas de intervalos y reglas generados con ambas propuestas metodológicas.

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathfrak{R}(X_k, \mathcal{P}_2)$
Previa	$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k]$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathfrak{p}_{1j}} i \in C_j$
	$I_2^{k,2} = (m_{C_j}^k, M_{C_j}^k]$	$r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathfrak{p}_{1i}=1-\mathfrak{p}_{1j}} i \in C_i$
	$I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$	$r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathfrak{p}_{2j}} i \in C_j$ $r_4 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathfrak{p}_{2i}=1-\mathfrak{p}_{2j}} i \in C_i$ $r_5 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathfrak{p}_{3i}=1} i \in C_i$

Tabla B.5: Caso 6: Relación entre Intervalos y Reglas (Propuesta original).

Nuestra propuesta es cerrar el intervalo del centro $I_2^{k,2}$, por ambos lados y dejar abierto por la derecha $I_1^{k,2}$ y abierto por la izquierda $I_3^{k,2}$.

Ahora tendremos un sistema más compacto con 2 reglas de probabilidad 1 que separan parcialmente las 2 clases. Esto identifica una variable parcialmente caracterizadora. En la Figura B.5 se ha indicado cómo se formularán los intervalos bajo esta propuesta.

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathcal{R}(X_k, \mathcal{P}_2)$
Revisada	$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k]$ $I_2^{k,2} = [m_{C_j}^k, M_{C_j}^k]$ $I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1} i \in C_i$ $r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}} i \in C_j$ $r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i$ $r_4 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3i}=1} i \in C_i$

Tabla B.6: Caso 6: Relación entre Intervalos y Reglas (Propuesta revisada).

La configuración alternativa que mostramos a continuación:

$$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k]$$

$$I_2^{k,2} = (m_{C_j}^k, M_{C_j}^k)$$

$$I_3^{k,2} = [M_{C_j}^k, M_{C_i}^k]$$

No produciría ninguna regla segura, puesto que $I_1^{k,2}$, $I_2^{k,2}$ e $I_3^{k,2}$ generarían 2 reglas de probabilidades complementarias cada uno, con lo que obtendríamos un sistemas de reglas con 6 reglas, mucho más extenso y menos seguro que $\mathfrak{R}(X_k, \mathcal{P}_2)$ y $\mathcal{R}(X_k, \mathcal{P}_2)$.

B.2.6 Caso 7, $m_{C_i}^k > m_{C_j}^k \wedge M_{C_i}^k < M_{C_j}^k$

El caso 7, se presenta como otro de los casos que podríamos llamar como comunes.

Este caso es simétrico del caso 6. El intervalo del centro contiene totalmente a la clase 1. la Figura B.6 presenta un boxplot múltiple que corresponde a esta situación.

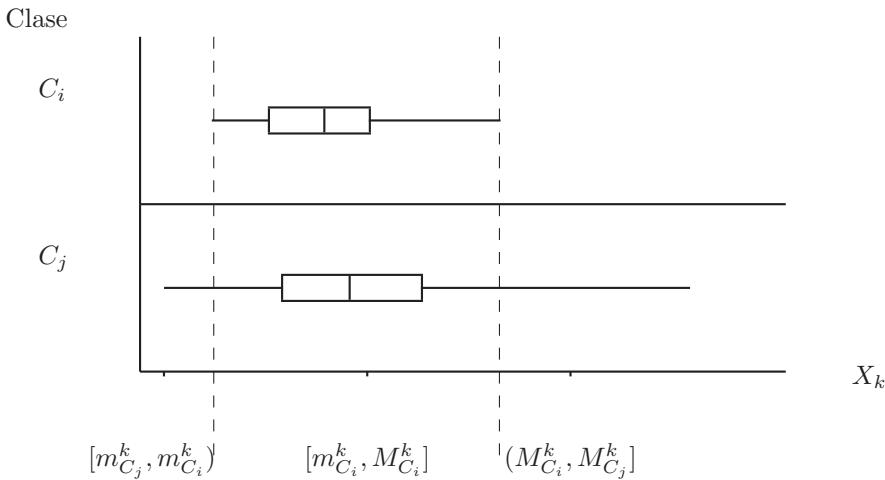


Figura B.6: Caso 7: Boxplot based Discretization.

Entonces, por un razonamiento parecido al del caso 6, en las Tablas B.7 y B.8 se pueden ver los sistemas de intervalos y reglas generados con ambas propuestas metodológicas.

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathfrak{R}(X_k, \mathcal{P}_2)$
Previa	$I_1^{k,2} = [m_{C_j}^k, m_{C_i}^k]$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}} i \in C_j$
	$I_2^{k,2} = (m_{C_i}^k, M_{C_i}^k]$	$r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{1i}=1-p_{1j}} i \in C_i$
	$I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k]$	$r_3 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{2j}} i \in C_j$ $r_4 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i$ $r_5 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3j}=1} i \in C_j$

Tabla B.7: Caso 7: Relación entre Intervalos y Reglas (Propuesta original).

Nuestra propuesta es cerrar el intervalo del centro $I_2^{k,2}$, por ambos lados y dejar abierto por la derecha $I_1^{k,2}$ y abierto por la izquierda $I_3^{k,2}$.

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathcal{R}(X_k, \mathcal{P}_2)$
Revisada	$I_1^{k,2} = [m_{C_j}^k, m_{C_i}^k)$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}=1} i \in C_j$
	$I_2^{k,2} = [m_{C_i}^k, M_{C_i}^k]$	$r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}} i \in C_j$
	$I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k]$	$r_3 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i$ $r_4 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3j}=1} i \in C_j$

Tabla B.8: Caso 7: Relación entre Intervalos y Reglas (Propuesta revisada).

Ahora tendremos un sistema más compacto con 2 reglas de probabilidad 1 que separan parcialmente las 2 clases. Esta variable se define como parcialmente caracterizadora. En la Figura B.6 se ha indicado cómo se formularán los intervalos bajo esta propuesta.

B.2.7 Caso 8, $m_{C_i}^k = m_{C_j}^k \wedge M_{C_i}^k < M_{C_j}^k$

El caso 8 se caracteriza también, como los casos 6 y 7, porque $r_k^{C_1} \subset r_k^{C_2}$.

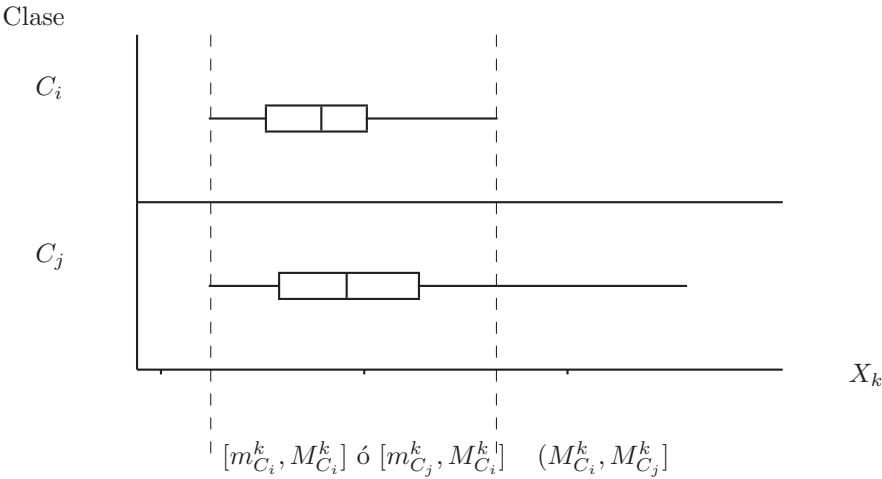


Figura B.7: Caso 8: Boxplot based Discretization.

El intervalo del centro contiene totalmente a la clase 1, pero uno de los extremos de ambas distribuciones coincide.

la Figura B.7 presenta un boxplot múltiple que corresponde a esta situación.

Entonces en las Tablas B.9 y B.10 se pueden ver los sistemas de intervalos y reglas generados con ambas propuestas metodológicas.

El intervalo $I_1^{k,2}$ es un intervalo que sólo contiene 1 punto, $m_{C_i}^k$ según la propuesta presentada en (Vázquez and Gibert 2001) y (Vázquez and Gibert 2002).

Nuestra propuesta es cerrar el intervalo $I_2^{k,2}$, por ambos lados y dejar abierto por la derecha $I_1^{k,2}$ con lo cual, en este caso generaríamos sólo 2 intervalos no vacíos, ya que el intervalo $I_1^{k,2} = \emptyset$, pues sus puntos son absorbidos por el intervalo $I_2^{k,2}$. Y finalmente dejar abierto por la izquierda $I_3^{k,2}$, así conseguiremos 1 regla segura y un sistema de reglas más compacto.

Ahora tendremos un sistema de reglas e intervalos más compactos, que separa parcialmente las 2 clases. Esto identifica una variable como parcialmente caracterizadora. En la Figura B.7 se ha indicado cómo se formularán los intervalos bajo esta propuesta.

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathcal{R}(X_k, \mathcal{P}_2)$
Previa	$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k]$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathfrak{p}_{1j}} i \in C_j$
	$I_2^{k,2} = (m_{C_j}^k, M_{C_i}^k]$	$r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathfrak{p}_{1i}=1-\mathfrak{p}_{1j}} i \in C_i$
	$I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k]$	$r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathfrak{p}_{2j}} i \in C_j$ $r_4 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathfrak{p}_{2i}=1-\mathfrak{p}_{2j}} i \in C_i$ $r_5 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathfrak{p}_{3j}=1} i \in C_j$

Tabla B.9: Caso 8: Relación entre Intervalos y Reglas (Propuesta original).

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathcal{R}(X_k, \mathcal{P}_2)$
Revisada	$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k)$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1c}=0} i \in C$
	$I_2^{k,2} = [m_{C_i}^k, M_{C_i}^k] \text{ ó } I_2^{k,2} = [m_{C_j}^k, M_{C_i}^k]$	$r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}} i \in C_j$
	$I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k]$	$r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i$ $r_4 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3j}=1} i \in C_j$

Tabla B.10: Caso 8: Relación entre Intervalos y Reglas (Propuesta revisada).

B.2.8 Caso 9, $m_{C_i}^k = m_{C_j}^k \wedge M_{C_i}^k > M_{C_j}^k$

El caso 9 es el simétrico del caso 8, y se caracteriza porque $r_k^{C_2} \subset r_k^{C_1}$. la Figura B.8 presenta un boxplot múltiple que corresponde a esta situación.

La clase 2 está totalmente contenida en el intervalo del centro, coincidiendo uno de los extremos de ambas distribuciones.

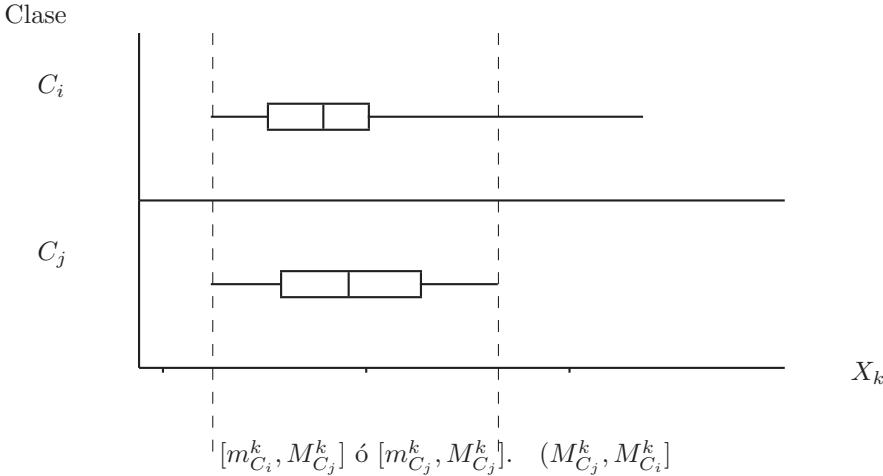


Figura B.8: Caso 9: Boxplot based Discretization.

Entonces, por un razonamiento parecido al del caso 8, en las Tablas B.11 y B.12 se pueden ver los sistemas de intervalos y reglas generados con ambas propuestas metodológicas.

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathfrak{R}(X_k, \mathcal{P}_2)$
Previa	$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k]$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}} i \in C_j$
	$I_2^{k,2} = (m_{C_j}^k, M_{C_j}^k]$	$r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1-p_{1j}} i \in C_i$
	$I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$	$r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}} i \in C_j$ $r_4 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i$ $r_5 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3i}=1} i \in C_i$

Tabla B.11: Caso 9: Relación entre Intervalos y Reglas (Propuesta original).

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathcal{R}(X_k, \mathcal{P}_2)$
Revisada	$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k)$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1c}=0} i \in C$
	$I_2^{k,2} = [m_{C_i}^k, M_{C_j}^k] \text{ ó } I_2^{k,2} = [m_{C_j}^k, M_{C_i}^k]$	$r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}} i \in C_j$
	$I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$	$r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i$ $r_4 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3i}=1} i \in C_i$

Tabla B.12: Caso 9: Relación entre Intervalos y Reglas (Propuesta revisada).

El intervalo $I_1^{k,2}$ es un intervalo que sólo contiene 2 puntos, que son los mínimos de cada clase.

Nuestra propuesta es cerrar el intervalo $I_2^{k,2}$, por ambos lados y dejar abierto por la derecha $I_1^{k,2}$, con lo cual el intervalo $I_1^{k,2} = \emptyset$, pues sus puntos son absorbidos por el intervalo $I_2^{k,2}$. Y finalmente dejar abierto por la izquierda $I_3^{k,2}$.

Ahora tendremos sistemas de reglas e intervalos más compactos, con una regla de probabilidad 1 que separa parcialmente las 2 clases. Aquí se identifica una variable parcialmente caracterizadora. En la Figura B.8 se ha indicado cómo se formularán los intervalos bajo esta propuesta.

B.2.9 Caso 10, $m_{C_i}^k > m_{C_j}^k \wedge M_{C_i}^k = M_{C_j}^k$

la Figura B.9 presenta un boxplot múltiple que corresponde a esta situación.

Entonces, por un razonamiento parecido a los realizados anteriormente, en las Tablas B.13 y B.14 se pueden ver los sistemas de intervalos y reglas generados con ambas propuestas metodológicas.

En este caso sólo se generarían 2 intervalos no vacíos ($I_3^{k,2} = \emptyset$), y no se producen reglas seguras (el intervalo $I_2^{k,2}$ contiene totalmente a la clase 1).

Nuestra propuesta es cerrar el intervalo $I_2^{k,2}$, por ambos lados y dejar abierto por la derecha $I_1^{k,2}$ con lo cual se mantienen los 2 intervalos. Ahora tendremos sistemas de reglas con una regla de probabilidad 1 que separa parcialmente las 2 clases.

Esto identifica una variable parcialmente caracterizadora. En la Figura B.9 se ha indicado cómo se formularán los intervalos bajo esta propuesta.

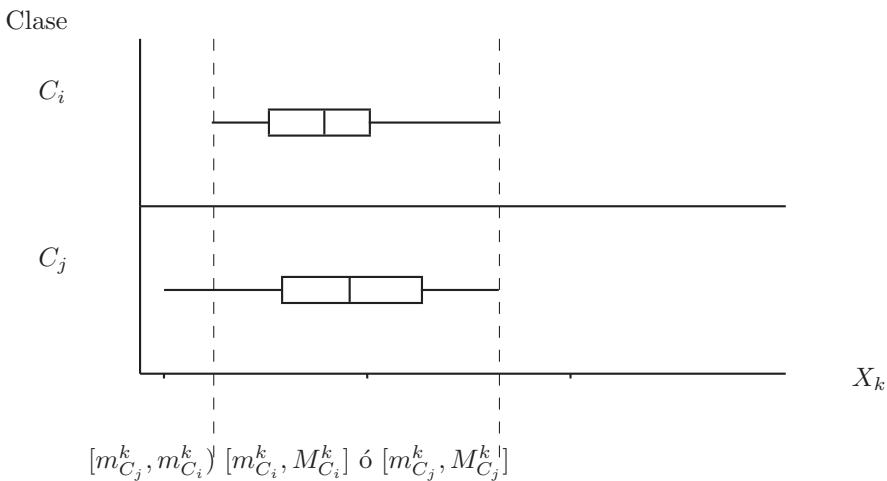


Figura B.9: Caso 10: Boxplot based Discretization.

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathfrak{R}(X_k, \mathcal{P}_2)$
Previa	$I_1^{k,2} = [m_{C_j}^k, m_{C_i}^k]$ $I_2^{k,2} = (m_{C_i}^k, M_{C_i}^k] \text{ ó } I_2^{k,2} = (m_{C_i}^k, M_{C_j}^k]$ $I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k] \text{ ó } I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathbf{p}_{1j}} i \in C_j$ $r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathbf{p}_{1i}=1-\mathbf{p}_{1j}} i \in C_i$ $r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathbf{p}_{2j}} i \in C_j$ $r_4 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathbf{p}_{2i}=1-\mathbf{p}_{2j}} i \in C_i$ $r_5 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathbf{p}_{3c}=0} i \in C$

Tabla B.13: Caso 10: Relación entre Intervalos y Reglas (Propuesta original).

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathcal{R}(X_k, \mathcal{P}_2)$
Revisada	$I_1^{k,2} = [m_{C_j}^k, m_{C_i}^k]$ $I_2^{k,2} = [m_{C_i}^k, M_{C_j}^k]$ ó $I_2^{k,2} = [m_{C_i}^k, M_{C_i}^k]$ $I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k]$ ó $I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}=1} i \in C_j$ $r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}} i \in C_j$ $r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i$ $r_4 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3c}=0} i \in C$

Tabla B.14: Caso 10: Relación entre Intervalos y Reglas (Propuesta revisada).

B.2.10 Caso 11, $m_{C_i}^k < m_{C_j}^k \wedge M_{C_i}^k = M_{C_j}^k$

Este caso es el simétrico del caso 10. la Figura B.10 presenta un boxplot múltiple que corresponde a esta situación.

Entonces, por un razonamiento parecido a los realizados anteriormente, en las Tablas B.15 y B.16 se pueden ver los sistemas de intervalos y reglas generados con ambas propuestas metodológicas.

En este caso sólo se generarían 2 intervalos no vacíos ($I_3^{k,2} = \emptyset$), y no se producen reglas de probabilidad 1 (el intervalo $I_2^{k,2}$ contiene totalmente a la clase 2).

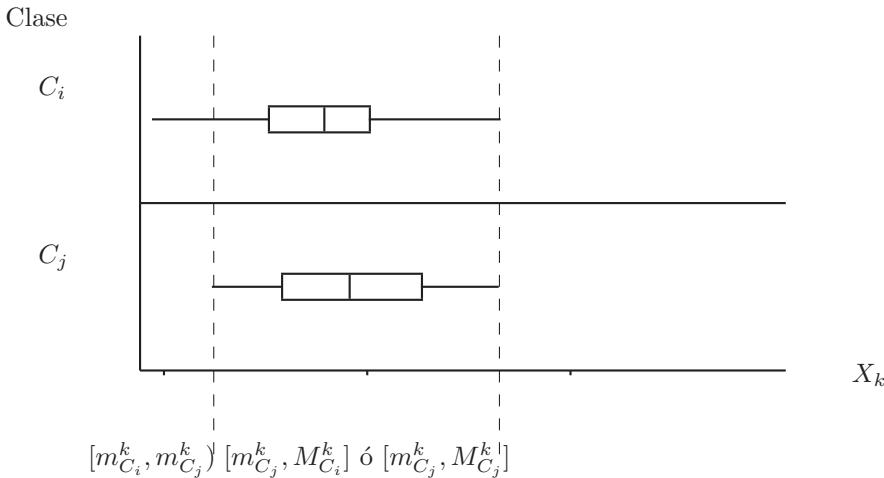


Figura B.10: Caso 11: Boxplot based Discretization.

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathfrak{N}(X_k, \mathcal{P}_2)$
Previa	$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k]$ $I_2^{k,2} = (m_{C_j}^k, M_{C_j}^k]$ ó $I_2^{k,2} = (m_{C_j}^k, M_{C_i}^k]$ $I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k]$ ó $I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}} i \in C_j$ $r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1-p_{1j}} i \in C_i$ $r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}} i \in C_j$ $r_4 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i$ $r_5 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3c}=0} i \in C$

Tabla B.15: Caso 11: Relación entre Intervalos y Reglas (Propuesta original).

Como ya hemos visto anteriormente, según cómo se definamos los límites de los intervalos, el sistema de reglas inducido a partir de él producirá más o menos reglas seguras y en consecuencia, más o menos valores (o intervalos) propios de una clase.

Nuestra propuesta es cerrar el intervalo $I_2^{k,2}$, por ambos lados y dejar abierto por la derecha $I_1^{k,2}$ con lo cual se mantienen los 2 intervalos, pero ahora tendremos sistemas de reglas con una regla de probabilidad 1 que separa parcialmente las 2 clases.

Propuesta	Sistema de Intervalos \mathcal{D}^k	Sistema de Reglas $\mathcal{R}(X_k, \mathcal{P}_2)$
Revisada	$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k)$ $I_2^{k,2} = [m_{C_j}^k, M_{C_j}^k] \text{ ó } I_2^{k,2} = [m_{C_j}^k, M_{C_i}^k]$ $I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k] \text{ ó } I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$	$r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1} i \in C_i$ $r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}} i \in C_j$ $r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}=1-p_{2j}} i \in C_i$ $r_4 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3c}=0} i \in C$

Tabla B.16: Caso 11: Relación entre Intervalos y Reglas (Propuesta revisada).

Esto identifica una variable como parcialmente caracterizadora. En la Figura B.10 se ha indicado cómo se formularán los intervalos bajo esta propuesta.

B.2.11 Caso 12, $M_{C_i}^k = m_{C_j}^k$

El caso 12, se presenta como otro de los casos extremos. la Figura B.11 presenta un boxplot múltiple que corresponde a esta situación.

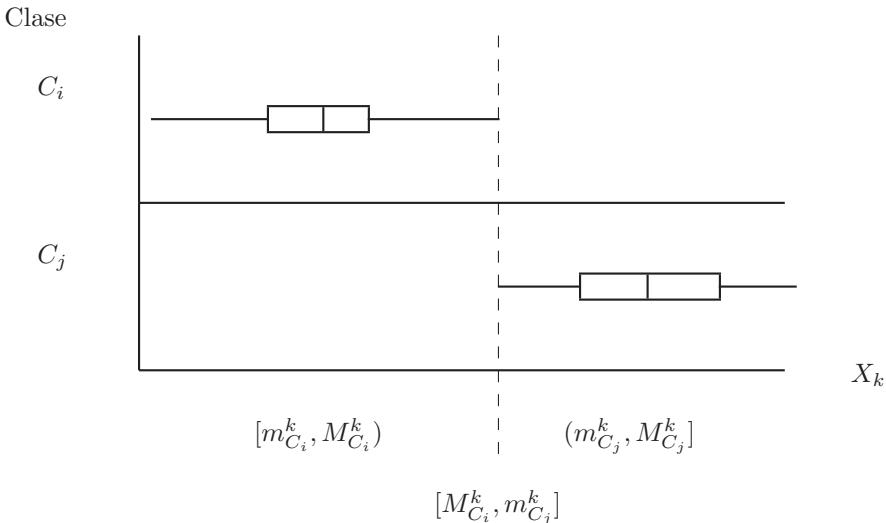


Figura B.11: Caso 12: Boxplot based Discretization.

Según la propuesta presentada en (Vázquez and Gibert 2001) y (Vázquez and Gibert 2002), el sistema de intervalos inducido por \mathcal{P}_2 sobre X_k sería:

$$I_1^{k,2} = [m_{C_i}^k, M_{C_i}^k] \text{ o bien } I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k]$$

$$I_2^{k,2} = (M_{C_i}^k, m_{C_j}^k]$$

$$I_3^{k,2} = (m_{C_j}^k, M_{C_j}^k] \text{ o bien } I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k]$$

Lo que genera el siguiente sistemas de reglas completo $\mathfrak{R}(X_k, \mathcal{P}_2)$:

$$\begin{aligned} \mathfrak{R}(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathfrak{p}_{1i}} i \in C_i, \\ & r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathfrak{p}_{1j}=1-p_{1i}} i \in C_j, \\ & r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathfrak{p}_{2i}=0} i \in C_i, \\ & r_4 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathfrak{p}_{2j}=0} i \in C_j, \\ & r_5 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathfrak{p}_{3j}=1} i \in C_j, \\ & r_6 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathfrak{p}_{3i}=0} i \in C_i \end{aligned} \}$$

Teniendo en cuenta que las reglas con probabilidad cero ($p_{sc} = 0$) son prescindibles, el sistema de reglas efectivas $\mathfrak{Re}(X_k, \mathcal{P}_2)$ es:

$$\begin{aligned} \mathfrak{R}(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathfrak{p}_{1i}} i \in C_i, \\ & r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathfrak{p}_{1j}=1-p_{1i}} i \in C_j, \\ & r_4 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathfrak{p}_{3j}=1} i \in C_j \end{aligned} \}$$

En este caso sólo se generarían 2 intervalos no vacíos, ya que $I_2^{k,2} = \emptyset$, y se produce 1 regla de probabilidad 1 a partir del intervalo $I_3^{k,2}$ y 2 reglas con probabilidades complementarias a partir del intervalo $I_1^{k,2}$.

Y el sistema de reglas reducido $\mathfrak{R}^*(X_k, \mathcal{P}_2)$ sería:

$$\begin{aligned} \mathfrak{R}^*(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\text{Max}\{\mathfrak{p}_{1j}, \mathfrak{p}_{1i}\}} i \in C, \quad \text{siendo, } \mathcal{C} = \begin{cases} C_i & \text{si } \mathfrak{p}_{1i} \geq \mathfrak{p}_{1j} \\ C_j & \text{si } \mathfrak{p}_{1i} < \mathfrak{p}_{1j} \end{cases} \\ & r_2 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathfrak{p}_{3j}=1} i \in C_j \end{aligned} \}$$

Nuestra propuesta es definir el sistema de intervalos inducidos por \mathcal{P}_2 sobre X_k , de la siguiente forma:

$$I_1^{k,2} = [m_{C_i}^k, m_{C_j}^k) \text{ o bien } I_1^{k,2} = [m_{C_i}^k, M_{C_i}^k)$$

$$I_2^{k,2} = [M_{C_i}^k, m_{C_j}^k] \text{ o bien } I_2^{k,2} = [M_{C_i}^k, M_{C_i}^k] \text{ ó } I_2^{k,2} = [m_{C_j}^k, m_{C_j}^k]$$

$$I_3^{k,2} = (m_{C_j}^k, M_{C_j}^k] \text{ o bien } I_3^{k,2} = (M_{C_i}^k, M_{C_j}^k]$$

Si generamos 3 intervalos, no vacíos, cerrando el intervalo $I_2^{k,2}$ por ambos lados y dejamos abierto por la derecha $I_1^{k,2}$ y por la izquierda $I_3^{k,2}$, conseguimos un sistema de reglas menos compacto que la propuesta de (Vázquez and Gibert 2001), pero con 2 reglas de probabilidad 1 que separan casi totalmente las 2 clases, ya que el nuevo intervalo $I_2^{k,2}$, sólo contendrá el punto $m_{C_j}^k = M_{C_i}^k$. En la Figura B.11 se ha indicado cómo se formularán los intervalos bajo esta propuesta.

El sistema de reglas completo $\mathfrak{R}(X_k, \mathcal{P}_2)$, omitiendo las reglas no efectivas, es:

$$\mathcal{R}(X_k, \mathcal{P}_2) = \mathcal{R}e(X_k, \mathcal{P}_2) = \{ \begin{array}{l} r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1} i \in C_i, \\ r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}} i \in C_i, \\ r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}=1-p_{2i}} i \in C_j, \\ r_4 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3j}=1} i \in C_j \end{array} \}$$

El sistema de reglas reducido $\mathcal{R}^*(X_k, \mathcal{P}_2)$ es:

$$\mathcal{R}^*(X_k, \mathcal{P}_2) = \{ \begin{array}{l} r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1i}=1} i \in C_i, \\ r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{\text{Max}\{p_{2j}, p_{2i}\}} i \in C, \quad \text{siendo, } \mathcal{C} = \begin{cases} C_i & \text{si } p_{2i} \geq p_{2j} \\ C_j & \text{si } p_{2i} < p_{2j} \end{cases} \\ r_3 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3j}=1} i \in C_j \end{array} \}$$

Y finalmente el sistema de reglas seguras $\mathcal{S}(X_k, \mathcal{P}_2)$ es:

$$\mathcal{S}(X_k, \mathcal{P}_2) = \{ \begin{array}{l} r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p=1} i \in C_i, \\ r_2 : x_{ik} \in I_3^{k,2} \xrightarrow{p=1} i \in C_j \end{array} \}$$

B.2.12 Caso 13, $M_{C_j}^k = m_{C_i}^k$

El caso 13, se presenta como otro de los casos que podríamos llamar extremos. Este caso es el simétrico del caso 12. la Figura B.12 presenta un boxplot múltiple que corresponde a esta situación.

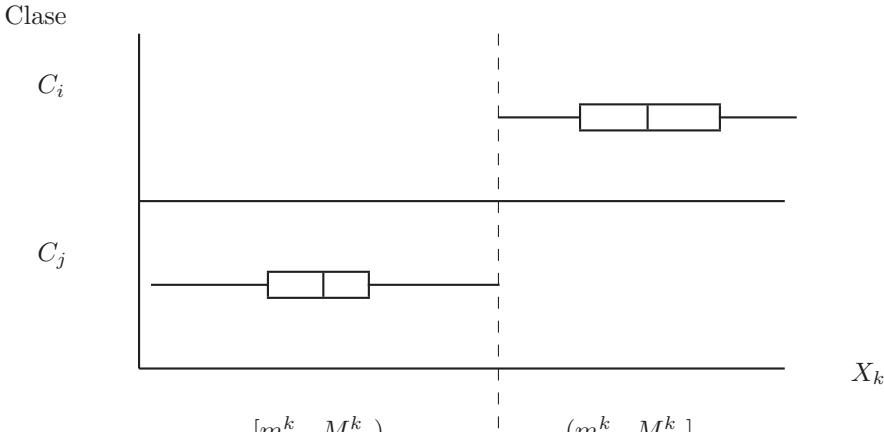


Figura B.12: Caso 13: Boxplot based Discretization.

Según la propuesta presentada en (Vázquez and Gibert 2001) y (Vázquez and Gibert 2002), el sistema de intervalos inducido por \mathcal{P}_2 sobre X_k sería:

$$I_1^{k,2} = [m_{C_j}^k, M_{C_j}^k] \text{ o bien } I_1^{k,2} = [m_{C_i}^k, M_{C_i}^k]$$

$$I_2^{k,2} = (M_{C_j}^k, m_{C_i}^k]$$

$$I_3^{k,2} = (m_{C_i}^k, M_{C_i}^k] \text{ o bien } I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$$

Lo que genera el siguiente sistemas de reglas completo $\mathfrak{R}(X_k, \mathcal{P}_2)$:

$$\begin{aligned} \mathfrak{R}(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathbf{p}_{1i}} i \in C_i, \\ & r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathbf{p}_{1j}=1-\mathbf{p}_{1i}} i \in C_j, \\ & r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathbf{p}_{2i}=0} i \in C_i, \\ & r_4 : x_{ik} \in I_2^{k,2} \xrightarrow{\mathbf{p}_{2j}=0} i \in C_j, \\ & r_5 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathbf{p}_{3j}=0} i \in C_j, \\ & r_6 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathbf{p}_{3i}=1} i \in C_i \end{aligned} \}$$

Teniendo en cuenta que las reglas con probabilidad cero ($p_{sc} = 0$) son prescindibles, el sistema de reglas efectivas $\mathfrak{Re}(X_k, \mathcal{P}_2)$ es:

$$\begin{aligned} \mathfrak{Re}(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathbf{p}_{1i}} i \in C_i, \\ & r_2 : x_{ik} \in I_1^{k,2} \xrightarrow{\mathbf{p}_{1j}=1-\mathbf{p}_{1i}} i \in C_j, \\ & r_3 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathbf{p}_{3i}=1} i \in C_i \} \end{aligned}$$

En este caso sólo se generarían 2 intervalos no vacíos, ya que $I_2^{k,2} = \emptyset$, y se produce 1 regla de probabilidad 1 a partir del intervalo $I_3^{k,2}$ y 2 reglas con probabilidades complementarias a partir del intervalo $I_1^{k,2}$.

Y el sistema de reglas reducido $\mathfrak{R}^*(X_k, \mathcal{P}_2)$, omitiendo las reglas no efectivas, es:

$$\begin{aligned} \mathfrak{R}^*(X_k, \mathcal{P}_2) = \{ & r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{\text{Max}\{\mathbf{p}_{1j}, \mathbf{p}_{1i}\}} i \in C, \quad \text{siendo, } \mathcal{C} = \begin{cases} C_i & \text{si } \mathbf{p}_{1i} \geq \mathbf{p}_{1j} \\ C_j & \text{si } \mathbf{p}_{1i} < \mathbf{p}_{1j} \end{cases} \\ & r_2 : x_{ik} \in I_3^{k,2} \xrightarrow{\mathbf{p}_{3i}=1} i \in C_i \} \end{aligned}$$

Aquí también podemos decir que según como se definan los límites de los intervalos inducidos por \mathcal{P}_2 sobre X_k , el sistema de reglas inducido a partir de él producirá más o menos reglas seguras.

Nuestra propuesta es definir el sistema de intervalos inducidos por \mathcal{P}_2 sobre X_k , de la siguiente forma:

$$I_1^{k,2} = [m_{C_j}^k, m_{C_i}^k] \text{ o bien } I_1^{k,2} = [m_{C_j}^k, M_{C_j}^k)$$

$$I_2^{k,2} = [M_{C_j}^k, m_{C_i}^k] \text{ o bien } I_2^{k,2} = [M_{C_j}^k, M_{C_j}^k] \text{ o } I_2^{k,2} = [m_{C_i}^k, M_{C_i}^k]$$

$$I_3^{k,2} = (m_{C_i}^k, M_{C_i}^k] \text{ o bien } I_3^{k,2} = (M_{C_j}^k, M_{C_i}^k]$$

Si generamos 3 intervalos no vacíos, cerrando el intervalo $I_2^{k,2}$ por ambos lados y los dejamos abierto por la derecha $I_1^{k,2}$ y por la izquierda $I_3^{k,2}$, conseguiríamos un sistema de reglas menos compacto que la propuesta de (Vázquez and Gibert 2001), pero con 2 reglas de probabilidad 1 que separan casi totalmente las 2 clases, ya que el nuevo intervalo $I_2^{k,2}$, sólo contendrá 2 puntos. En la Figura B.12 se ha indicado cómo se formularán los intervalos bajo esta propuesta.

El sistema de reglas completo $\mathcal{R}(X_k, \mathcal{P}_2)$, omitiendo las reglas no efectivas, es:

$$\mathcal{R}(X_k, \mathcal{P}_2) = \mathcal{R}e(X_k, \mathcal{P}_2) = \{ \begin{array}{l} r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}=1} i \in C_j, \\ r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2i}} i \in C_i, \\ r_3 : x_{ik} \in I_2^{k,2} \xrightarrow{p_{2j}=1-p_{2i}} i \in C_j, \\ r_4 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3i}=1} i \in C_i \end{array} \}$$

El sistema de reglas reducido $\mathcal{R}^*(X_k, \mathcal{P}_2)$ es:

$$\mathcal{R}^*(X_k, \mathcal{P}_2) = \{ \begin{array}{l} r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p_{1j}=1} i \in C_j, \\ r_2 : x_{ik} \in I_2^{k,2} \xrightarrow{\text{Max}\{p_{2i}, p_{2j}\}} i \in C, \quad \text{siendo, } \mathcal{C} = \begin{cases} C_i & \text{si } p_{2i} \geq p_{2j} \\ C_j & \text{si } p_{2i} < p_{2j} \end{cases} \\ r_3 : x_{ik} \in I_3^{k,2} \xrightarrow{p_{3i}=1} i \in C_i \end{array} \}$$

Y el sistema de reglas seguras $\mathcal{S}(X_k, \mathcal{P}_2)$ es:

$$\mathcal{S}(X_k, \mathcal{P}_2) = \{ \begin{array}{l} r_1 : x_{ik} \in I_1^{k,2} \xrightarrow{p=1} i \in C_j, \\ r_2 : x_{ik} \in I_3^{k,2} \xrightarrow{p=1} i \in C_i \end{array} \}$$

Anexo C

Análisis descriptivo de los datos, planta catalana

C.1 Análisis Univariante

C.1.1 Variables de entrada

Variable FE-E. Pretratamiento con hierro (ppm=mg/l).

Tenemos una media de hierro a la entrada de la Planta de 48,3 mg por litro con una desviación de 13,934, con unos valores que van de 0 a 89,9. Existen tres missings para esta variable (datos n° 61: 31-X-95, 122: 31-XI-95 y 153: 31-I-96), aunque el segundo corresponde al caso inicialmente comentado es destacable el encontrar éstos a final de mes.

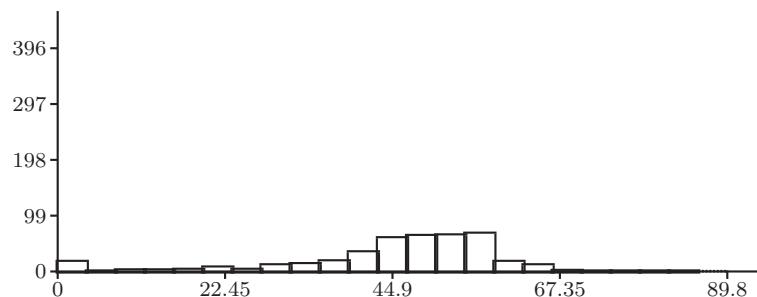


Figura C.1: Histograma de la variable FE-E.

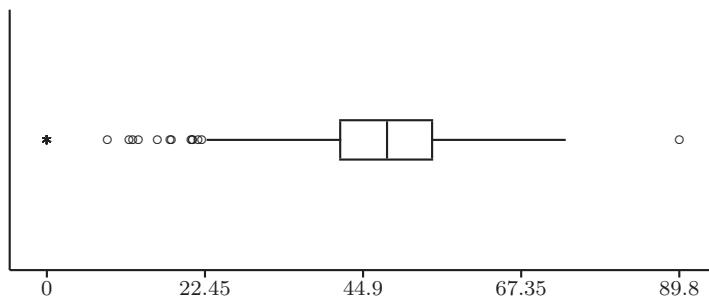


Figura C.2: Boxplot de la variable FE-E.

En el Histograma de la variable , que se muestra en la Figura C.1, se observan claramente los grupos de valores extraños, aunque el comportamiento de esta variable es bastante simétrico.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	3
Número de observaciones útiles	393
Media	45.5303
Mediana	48.3
Primer quartil (Q1)	41.8
Tercer quartil (Q3)	54.55
Mínimo	0
Màximo	89.8
Variància	193.6542
Desviación típica	13.916
Quasi-Desviación típica	13.9337
Coeficiente de variación	0.3056

Se observa la independencia de los datos. Tanto la media como la variabilidad de los datos se mantiene constante, aunque observamos la presencia de 5 valores o grupos de valores que dan los picos que se observan claramente:

- 19-IX-95 al 25-IX-95 (del 19 al 25): Valor 0 para todos ellos.
- 25-XII-95 al 30-XII-95 (del 116 al 121): Valor 0 para todos ellos.
- 18-I-96 (dato nº 140): Valor de 89,8; casi el doble de lo normal.
- 1-II-96 al 4-II-96 (del 154 al 157): Valor 0 para todos ellos.
- 30-IX-96 (dato nº 396): Valor 8,6.

Variable PH-E. PH (unidades de PH).

El PH a la entrada de la Planta es de 7,6172, en media, con una desviación de 0,1464, con unos valores que van de 7,2 a 8, por lo que corresponde a una variable con muy poca variabilidad y un rango de variación también muy pequeño. Existen 71 missings para esta variable distribuidos uniformemente a lo largo del tiempo, lo que nos obliga a trabajar con 325 datos.

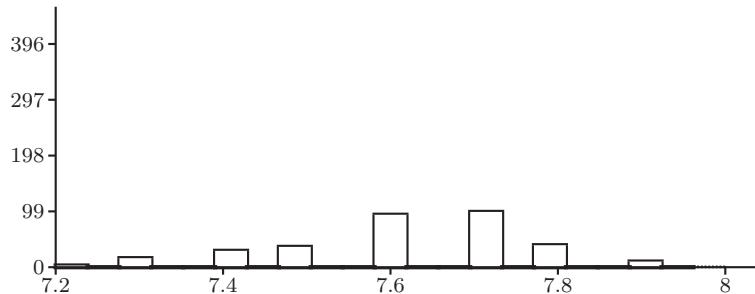


Figura C.3: Histograma de la variable PH-E.

En el Histograma de la variable se observa que la mayor parte de valores están concentrados en un pequeño intervalo, tal y como esperábamos, y su distribución es bastante simétrica y próxima a una normal, que se muestra en la Figura C.3

No se observa la independencia de los datos, dándose repetidos cambios de media en pequeños intervalos de tiempo. La variabilidad de los datos se mantiene constante y, en este caso, no aparece ningún dato que signifique una gran diferencia respecto a los anteriores; recordemos que existe muy poca variabilidad en esta variable.

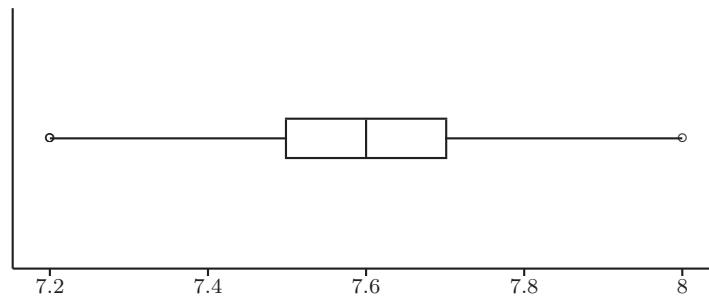


Figura C.4: Boxplot de la variable PH-E.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	71
Número de observaciones útiles	325
Media	7.6172
Mediana	7.6
Primer quartil (Q1)	7.5
Tercer quartil (Q3)	7.7
Mínimo	7.2
Máximo	8
Variància	0.0213
Desviación típica	0.1459
Quasi-Desviación típica	0.1461
Coeficiente de variación	0.0191

Variable SS-E. Sólidos en suspensión (mg de sólidos por litro de agua).

En el Histograma de la variable , que se muestra en la Figura C.5, se observa que la mayor parte de valores están concentrados en dos o tres intervalos, y su distribución es aplanada a la derecha, es decir, predominan los valores pequeños (debido a los saltos a valores elevados antes comentados).

Tenemos una media, a la entrada de la Planta, de 212 mg de sólidos suspendidos con una desviación de 87,09 y unos valores que van de 62 a 655. Existen 70 missings para esta variable, lo que nos obliga a trabajar con 326 datos. Estos 70 missings coinciden con los de la variable anterior, a excepción del dato nº 346 (11-VIII-96) que en este caso si tiene valor.

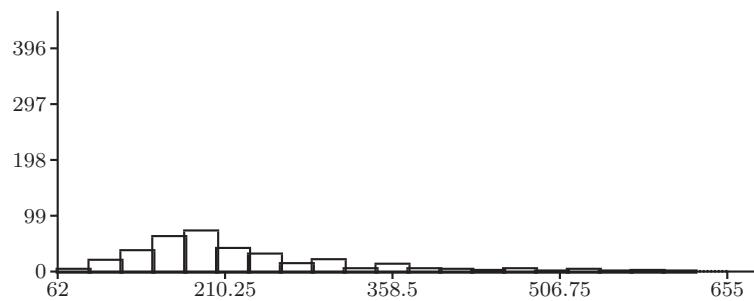


Figura C.5: Histograma de la variable SS-E.

Se observa la independencia de los datos. Tanto la media como la variabilidad de los datos se mantiene constante, aunque aparecen datos de valor elevado que se dan repentinamente y no de forma gradual. Se observa gran número de éstos casos aunque el más pronunciado se da en la parte central de los datos; nº 210 (28-III-96) que corresponde al valor máximo.

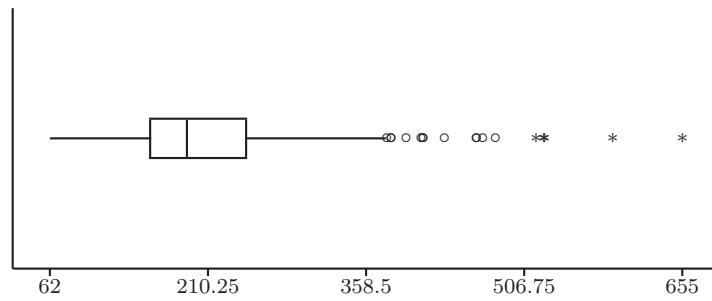


Figura C.6: Boxplot de la variable SS-E.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	70
Número de observaciones útiles	326
Media	212.0031
Mediana	190.5
Primer quartil (Q1)	157
Tercer quartil (Q3)	245
Mínimo	62
Màximo	655
Variància	7,561.5356
Desviación típica	86.9571
Quasi-Desviación típica	87.0908
Coeficiente de variación	0.4102

Variable SSV-E. Sólidos volátiles en suspensión (mg de sólidos/l).

La media de sólidos volátiles suspendidos a la entrada de la Planta es de 159,04 mg con una desviación de 64,87, con unos valores que van de 19 a 593 mg. Existen 70 missings para esta variable, lo que nos obliga a trabajar con 326 datos. Estos 70 missings coinciden con los de la variable anterior.

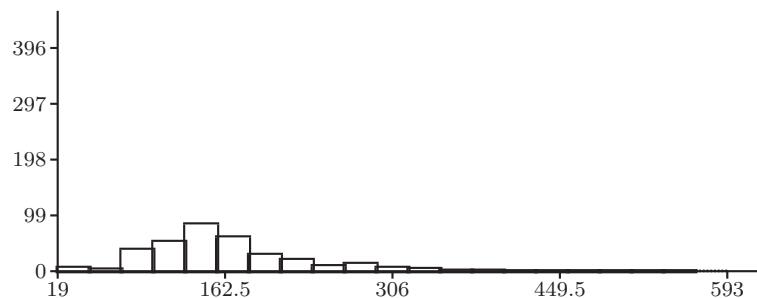


Figura C.7: Histograma de la variable SSV-E.

En el Histograma de la variable , que se muestra en la Figura C.7, se observa que la mayor parte de valores están concentrados en dos o tres intervalos, y su distribución es aplanada a la derecha, es decir, predominan los valores pequeños (debido a los saltos a valores elevados antes comentados) como en el caso anterior, aunque aquí no tan pronunciado.

Tanto la media como la variabilidad de los datos se mantiene constante, aunque aparecen datos de valor elevado que se dan repentinamente y no de forma gradual, aunque en menor medida que en el caso anterior. Se observa un outlier que predomina sobre los otros posibles; se da en el dato n° 210 (28-III-96) que corresponde al valor máximo, y coincide con el de la variable anterior, aunque era de esperar por el significado de las variables.

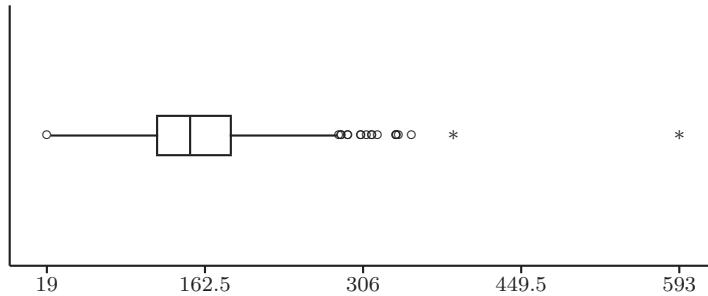


Figura C.8: Boxplot de la variable SSV-E.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	70
Número de observaciones útiles	326
Media	159.0368
Mediana	149
Primer quartil (Q1)	120
Tercer quartil (Q3)	185
Mínimo	19
Máximo	593
Variància	4,194.834
Desviación típica	64.7675
Quasi-Desviación típica	64.8671
Coeficiente de variación	0.4072

Variable DQO-E. Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua).

Tenemos una media de 442,23 mg oxígeno por litro de agua, con una desviación de 166,8, con unos valores que van de 27 a 1579 mg. Existen 8 missings para esta variable, lo que nos obliga a trabajar con 388 datos. Estos 8 missings corresponden a los datos n°: 63 (2-XI-95), 71 (10-XI-95), 93 (2-XII-95), 101 (10-XII-95), 122 (31-XII-95), 146 (24-I-96), 219 (6-IV-96) y 232 (19-IV-96).

Tanto la media como la variabilidad de los datos se mantiene constante, aunque observamos la presencia de 4 valores que dan los outliers que se observan claramente:

- 22-IX-95 (dato n° 22): Valor por encima de 1500.
- 29-IX-95 (dato n° 29): Valor por encima de 1200.
- 26-I-96 (dato n° 148): Valor por encima de 1000.
- 28-III-96 (dato n° 210): Valor por encima de 1000.

En el Histograma de la variable , que se muestra en la Figura C.9, se observan claramente los grupos de valores que quedan al margen del comportamiento normal, aunque el comportamiento de esta variable es bastante simétrico aunque predominan los valores pequeños de la variable.

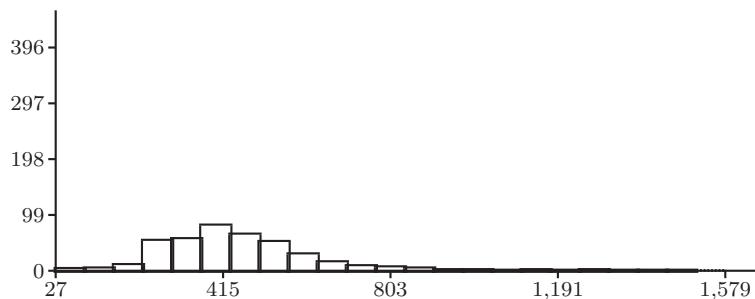


Figura C.9: Histograma de la variable DQO-E.

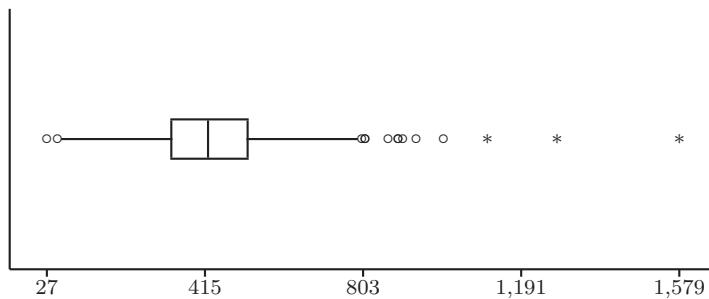


Figura C.10: Boxplot de la variable DQO-E.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	8
Número de observaciones útiles	388
Media	442.232
Mediana	422.5
Primer quartil (Q1)	335
Tercer quartil (Q3)	517
Mínimo	27
Màximo	1,579
Variància	27,749.5469
Desviación típica	166.582
Quasi-Desviación típica	166.7971
Coeficiente de variación	0.3767

Variable DBO-E. Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua).

Tenemos una media de 216,84 mg de oxígeno por litro de agua con una desviación de 102,28, con unos valores que van de 69 a 987. Existen 101 missings para esta variable distribuidos de forma uniforme, lo que nos obliga a trabajar con 295 datos solamente.

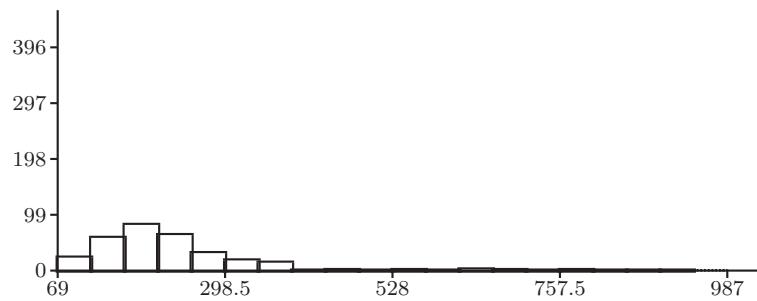


Figura C.11: Histograma de la variable DBO-E.

En el Histograma de la variable , que se muestra en la Figura C.11, se observan claramente los grupos de valores que quedan al margen del comportamiento normal, aunque éste es bastante simétrico aunque predominan los valores pequeños de la variable como en el caso anterior.

Tanto la media como la variabilidad de los datos se mantiene constante, aunque observamos la presencia de 7 valores que dan los outliers que se observan, aunque existen 3 que son bastante claros: 22-IX-95 (dato nº 22): Valor de 987 mg/l, 26-I-96 (dato nº 148): Valor de 768 mg/l, 28-III-96 (dato nº 210): Valor de 700 mg/l.

Que corresponden a tres de los cuatro anteriores.

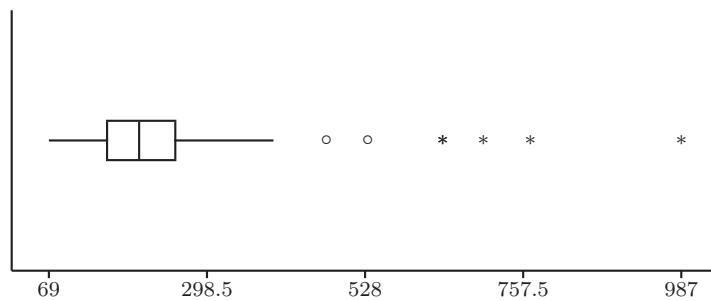


Figura C.12: Boxplot de la variable DBO-E.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	101
Número de observaciones útiles	295
Media	216.8373
Mediana	200
Primer quartil (Q1)	155
Tercer quartil (Q3)	251
Mínimo	69
Máximo	987
Variància	10,426.4424
Desviación típica	102.11
Quasi-Desviación típica	102.2835
Coeficiente de variación	0.4709

C.1.2 Variables después de la decantación

Variable PH-D. pH (unidades de pH)

La media de esta variable es de 7,56 con una desviación de 0,1461, con unos valores que van de 7,1 a 7,9, por lo que corresponde a una variable con muy poca variabilidad y un rango de variación también muy pequeño. Existen 71 missings para esta variable, lo que nos obliga a trabajar con 325 datos. Estos 71 missings coinciden con los de la variable PH-E.

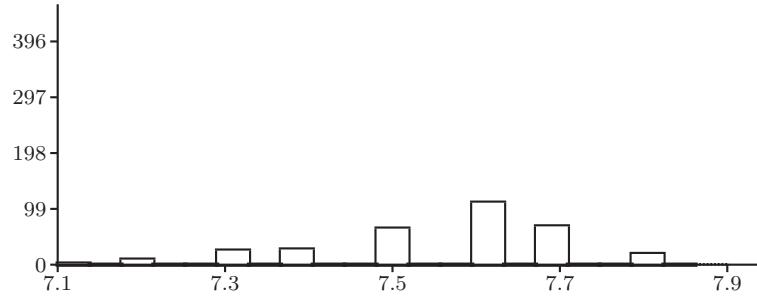


Figura C.13: Histograma de la variable PH-D.

En el Histograma de la variable , que se muestra en la Figura C.13, se observa que la mayor parte de valores están concentrados en un pequeño intervalo, tal y como esperábamos, y su distribución es bastante simétrica y próxima a una normal.

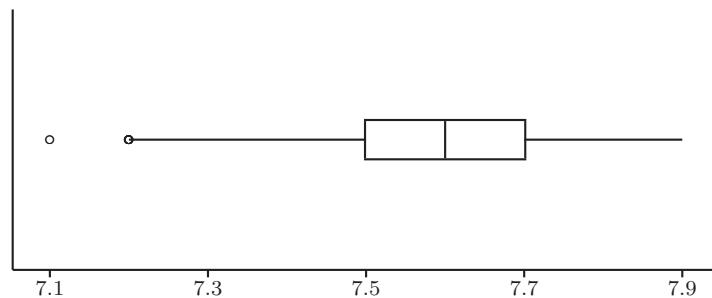


Figura C.14: Boxplot de la variable PH-D.

Se observa una cierta dependencia de los datos, ya que a partir de 1996 la tendencia de esta variable es creciente. Tanto la media como la variabilidad de los datos no se mantienen constantes y, en este caso, parece que esta última va disminuyendo a lo largo del tiempo. Por lo que respecta a la media y comparándola con la misma variable a la entrada de la Planta, observamos que en este caso los cambios de ésta son menores.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	71
Número de observaciones útiles	325
Media	7.56
Mediana	7.6
Primer quartil (Q1)	7.5
Tercer quartil (Q3)	7.7
Mínimo	7.1
Máximo	7.9
Variància	0.0212
Desviación típica	0.1457
Quasi-Desviación típica	0.1459
Coeficiente de variación	0.0193

Variable SS-D. Sólidos en suspensión (mg de sólidos por litro de agua).

La media de sólidos suspendidos después del decantador es de 89,62 mg con una desviación de 20.52, con unos valores que van de 40 a 192. Existen 70 missings para esta variable, lo que nos obliga a trabajar con 326 datos. Estos 70 missings coinciden con los de la misma variable medida a la entrada de la Planta (SS-E), lo cual parece lógico. Tanto la media como la variabilidad de los datos se mantiene constante, aunque aparecen datos de valor elevado que se dan repentinamente y no de forma gradual. Se observa gran número de éstos casos aunque el más pronunciado se da en la parte central-derecha de los datos; n° 288 (14-VI-96) que corresponde al valor máximo y no coincide con el de la misma variable medida a la entrada de la Planta, SS-E.

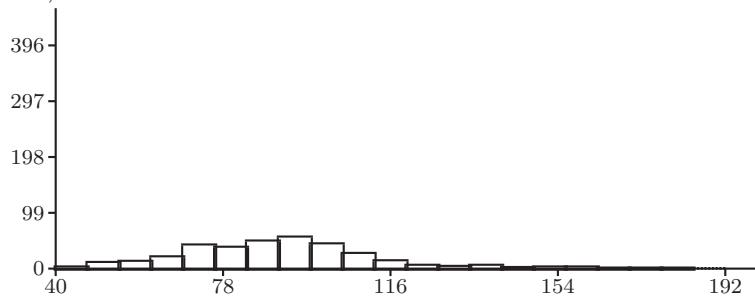


Figura C.15: Histograma de la variable SS-D.

En el Histograma de la variable , que se muestra en la Figura C.15, se observa que la mayor parte de valores están concentrados en dos intervalos, y su distribución es aplanada a la derecha, es decir, predominan los valores pequeños (debido a los saltos a valores elevados antes comentados).

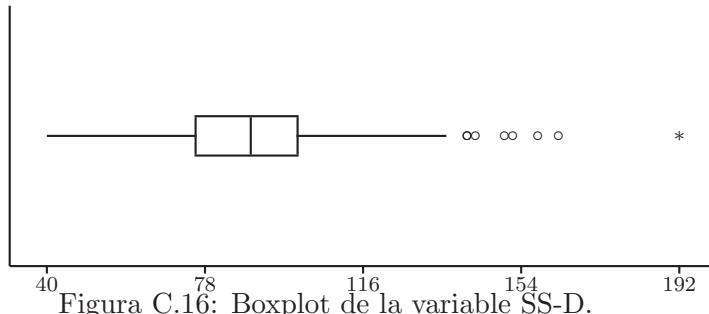


Figura C.16: Boxplot de la variable SS-D.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	70
Número de observaciones útiles	326
Media	89.6196
Mediana	89
Primer quartil (Q1)	76
Tercer quartil (Q3)	100
Mínimo	40
Máximo	192
Variància	419.6776
Desviación típica	20.486
Quasi-Desviación típica	20.5175
Coeficiente de variación	0.2286

Variable SSV-D. Sólidos volátiles en suspensión (mg de sólidos por litro de agua).

La media de sólidos volátiles suspendidos es de 65,417 mg con una desviación de 15,803, con unos valores que van de 13 a 134 mg. Existen 70 missings para esta variable, lo que nos obliga a trabajar con 326 datos. Estos 70 missings coinciden con los de la variable anterior. Tanto la media como la variabilidad de los datos se mantiene constante, aunque aparecen datos de valor elevado que se dan repentinamente y no de forma gradual, aunque en menor medida que en el caso anterior.

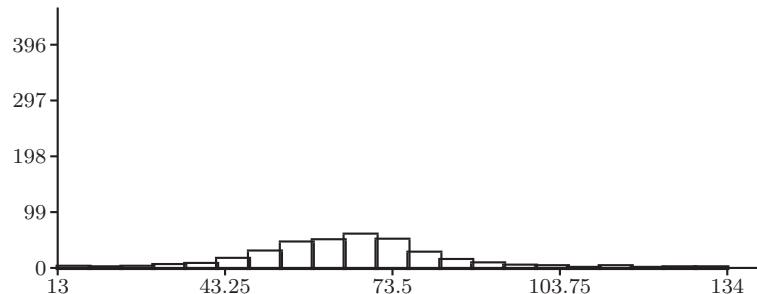


Figura C.17: Histograma de la variable SSV-D.

En el Histograma de la variable , que se muestra en la Figura C.17, se observa que la mayor parte de valores están concentrados en dos o tres intervalos, y su distribución es bastante centrada, es decir, no predominan ni los valores pequeños ni los elevados, sino los intermedios.

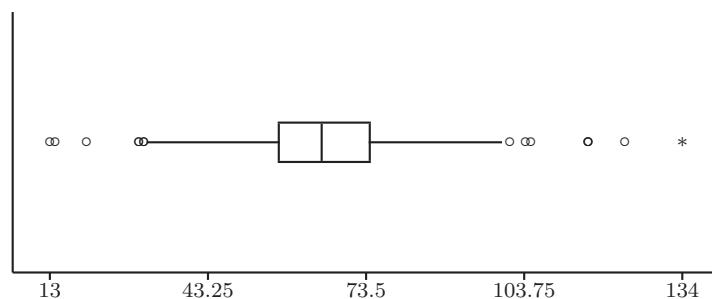


Figura C.18: Boxplot de la variable SSV-D.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	70
Número de observaciones útiles	326
Media	65.4172
Mediana	65
Primer quartil (Q1)	57
Tercer quartil (Q3)	74
Mínimo	13
Màximo	134
Variància	248.9735
Desviación típica	15.7789
Quasi-Desviación típica	15.8031
Coeficiente de variación	0.2412

Variable DQO-D. Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua).

La media de materia orgánica química oxidable es de 250,22 mg con una desviación de 60,52 con unos valores que van de 27 a 538 mg. Existen 9 missings para esta variable, lo que nos obliga a trabajar con 387 datos. Estos 9 missings corresponden a los datos nº: 63 (2-XI-95), 71 (10-XI-95), 86 (25-XI-95), 93 (2-XII-95), 101 (10-XII-95), 116 (25-XII-95), 122 (31-XII-95), 146 (24-I-96) y 232 (19-IV-96).

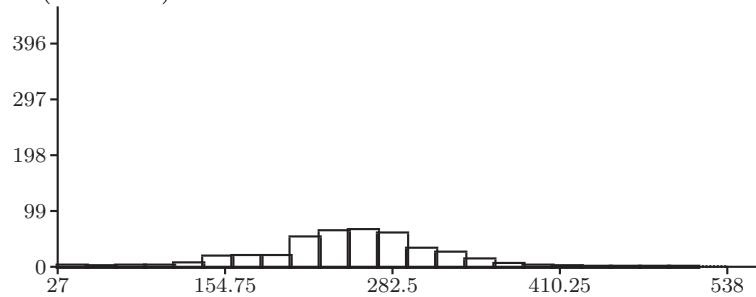


Figura C.19: Histograma de la variable DQO-D.

Tanto la media como la variabilidad de los datos se mantiene constante, aunque observamos la presencia de un valor especialmente elevado y otro muy bajo, pero este último se da de forma gradual.(30-I-96 (dato nº 152): Valor especialmente bajo, 5-VII-96 (dato nº 309): Valor especialmente elevado.)

En el Histograma de la variable , que se muestra en la Figura C.19, se observa claramente el valor especialmente elevado que queda en una categoría diferenciada, y su distribución es bastante centrada, es decir, no predominan ni los valores pequeños ni los elevados, sino los intermedios.

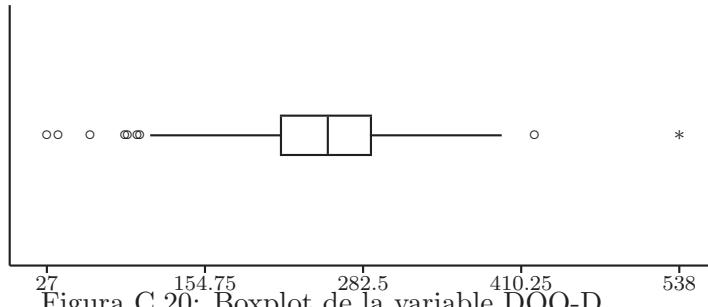


Figura C.20: Boxplot de la variable DQO-D.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	9
Número de observaciones útiles	387
Media	250.2196
Mediana	254
Primer quartil (Q1)	217
Tercer quartil (Q3)	288
Mínimo	27
Máximo	538
Variància	3,653.5015
Desviación típica	60.4442
Quasi-Desviación típica	60.5225
Coeficiente de variación	0.2416

Variable DBO-D. Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua).

La media de materia orgánica biodegradable después del decantador es de 121,86 mg con una desviación de 39,57 , con unos valores que van de 36 a 274 mg. Existen 96 missings para esta variable, lo que nos obliga a trabajar con 300 datos solamente.

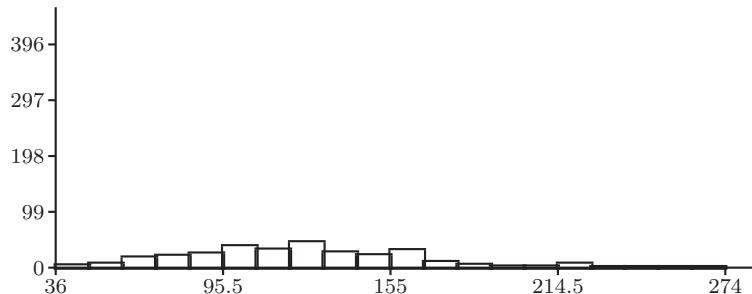


Figura C.21: Histograma de la variable DBO-D.

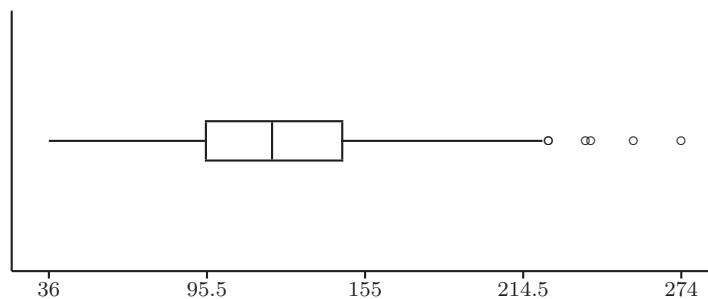


Figura C.22: Boxplot de la variable DBO-D.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	96
Número de observaciones útiles	300
Media	121.8633
Mediana	120
Primer quartil (Q1)	95.5
Tercer quartil (Q3)	146
Mínimo	36
Máximo	274
Variància	1,560.1772
Desviación típica	39.4991
Quasi-Desviación típica	39.5651
Coeficiente de variación	0.3241

Tanto la media como la variabilidad de los datos se mantienen constantes, aunque existe un pequeño número de datos que siguen un comportamiento especial; corresponden a los datos que van del 151 a 167 (del 29-I-96 al 14-II-96) que corresponde a valores pequeños y con poca variabilidad.

En el Histograma de la variable , que se muestra en la Figura C.21, se observa que no predomina ningún grupo de valores y existen muy pocos valores que se salgan del comportamiento normal.

C.1.3 Variables del tratamiento biológico

Variable V30-B. Análisis volumétrico 30; calidad de sedimentación de la mezcla (ml por litro de agua).

La media del índice V30-B es de 262,7 mililitros por litro con una desviación de 133,59, con unos valores que van de 77 a 770 mililitros por litro.

Existe 1 único missing para esta variable (dato nº 122: 31-XI-95), que contiene este mismo valor para todas y cada una de las variables.

Tanto la media como la variabilidad de los datos se mantiene constante, aunque aparecen datos de valor elevado que se dan repentinamente y no de forma gradual.

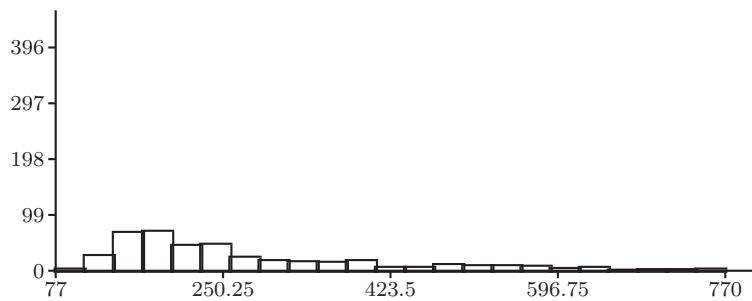


Figura C.23: Histograma de la variable V30-B.

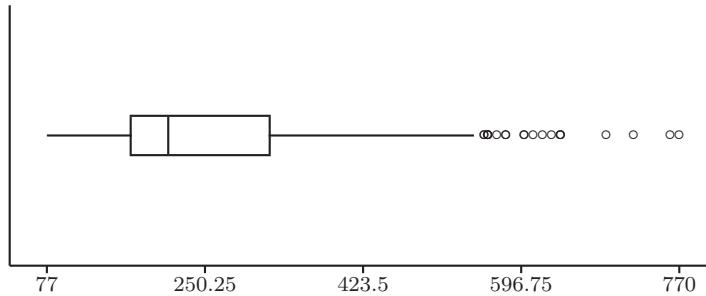


Figura C.24: Boxplot de la variable V30-B.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	1
Número de observaciones útiles	395
Media	262.6962
Mediana	210
Primer quartil (Q1)	170
Tercer quartil (Q3)	320
Mínimo	77
Máximo	770
Variància	17,801.8789
Desviación típica	133.4237
Quasi-Desviación típica	133.5929
Coeficiente de variación	0.5079

Variable MLSS-B. Sólidos en suspensión del licor mezcla (mg de sólidos por litro de licor mezcla).

La media de esta variable es de 1767,6 miligramos de sólidos por litro de agua con una desviación de 364,4, con unos valores que van de 754 a 3294 miligramos por litro.

Existen 7 missings para esta variable distribuidos de forma uniforme, lo que nos obliga a trabajar con 389 datos solamente.

Estos corresponden al 10-IX-95 (observación 10), 31-XII-95 (observación 122), 6-I-96 (observación 128), 7-IV-96 (observación 220) , 20-IV-96 (observación 233), 18-V-96 (observación 261) y 23-VI-96 (observación 297).

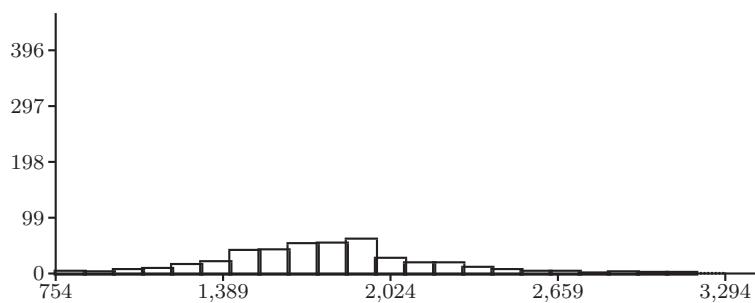


Figura C.25: Histograma de la variable MLSS-B.

En el Histograma de la variable , que se muestra en la Figura C.25, se observa que la mayor parte de valores están concentrados en dos o tres intervalos, y su distribución es simétrica, es decir, no predominan ni los valores pequeños ni los elevados, distribuyéndose de forma bastante similar.

Tanto la media como la variabilidad de los datos se mantiene constante, aunque aparecen datos de valor elevado y extremadamente bajos que se dan repentinamente y no de forma gradual.

Se observa gran número de éstos casos aunque los 5 más pronunciados se dan en los siguientes datos:

- 155-156 (2-II-96 a 3-II-96): Valores por encima de los 3000 mg/l.
- 213 (31-III-96): Valor por debajo de 1000 mg/l.
- 282 (8-VI-96): Valor por encima de 2000 mg/l.
- 313 (9-VII-96): Valor por debajo de 1000 mg/l.
- 381 (15-IX-96): Valor por debajo de 1000 mg/l.

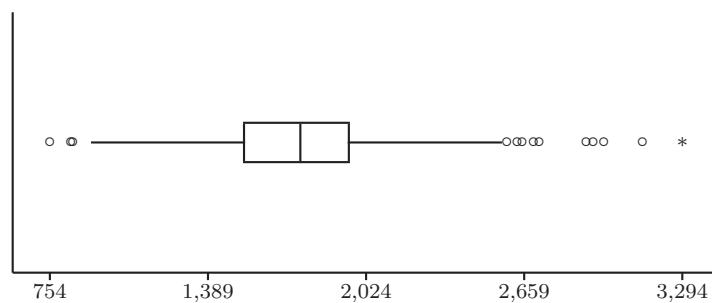


Figura C.26: Boxplot de la variable MLSS-B.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	7
Número de observaciones útiles	389
Media	1,767.6324
Mediana	1,760
Primer quartil (Q1)	1,538
Tercer quartil (Q3)	1,950.5
Mínimo	754
Màximo	3,294
Variància	132,482.0312
Desviación típica	363.9808
Quasi-Desviación típica	364.4497
Coeficiente de variación	0.2059

Variable MLVSS-B. Sólidos volátiles en suspensión del licor mezcla (mg de sólidos por litro de licor mezcla).

La media de esta variable es de 1344,7 miligramos por litro de agua con una desviación de 267,4, con unos valores que van de 185 a 2100 miligramos por litro. Existen 7 missings para esta variable lo que nos obliga a trabajar con 389 datos. Estos 7 missings coinciden con los de la variable anterior.

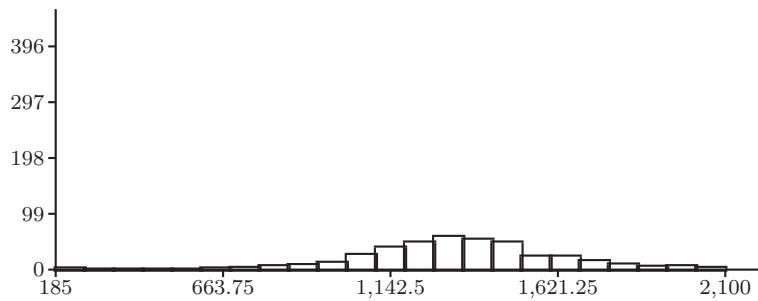


Figura C.27: Histograma de la variable MLVSS-B.

En el Histograma de la variable , que se muestra en la Figura C.27, se observa que la mayor parte de valores están concentrados en dos o tres intervalos, y su distribución es bastante simétrica, aunque se observa claramente como existe un grupo aislado que contiene los dos valores anteriormente comentados.

Tanto la media como la variabilidad de los datos se mantiene constante, aunque aparecen datos de valor elevado que se dan repentinamente y no de forma gradual, aunque en menor medida que en el caso anterior, más pronunciados en este. Se observan dos valores que predominan sobre el resto; se dan en los datos n° 76 (15-XI-95) y 106 (15-XII-95) que coinciden con el valor mínimo (185 mg/l).

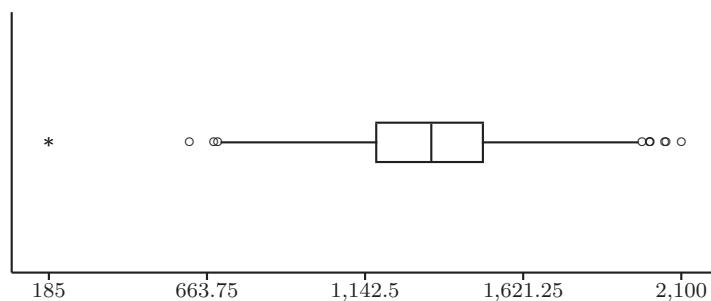


Figura C.28: Boxplot de la variable MLVSS-B.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	7
Número de observaciones útiles	389
Media	1,344.6813
Mediana	1,343
Primer quartil (Q1)	1,179.5
Tercer quartil (Q3)	1,496
Mínimo	185
Màximo	2,100
Variància	71,342.7734
Desviación típica	267.1007
Quasi-Desviación típica	267.4449
Coeficiente de variación	0.1986

Variable MCRT-B. Edad celular (días).

La media de esta variable es de 14,35 días con una desviación de 31,65 días, con unos valores que van de 1,78 a 341,99. Existen 9 missings para esta variable lo que nos obliga a trabajar con 387 datos. Los valores missing corresponden a los datos nº: 10 (10-IX-95), 80 (19-XI-95), 110 (19-XII-95), 122 (31-XII-95), 128 (6-I-96), 220 (7-IV-96), 233 (20-IV-96), 261 (18-V-96) y 297 (23-VI-96).

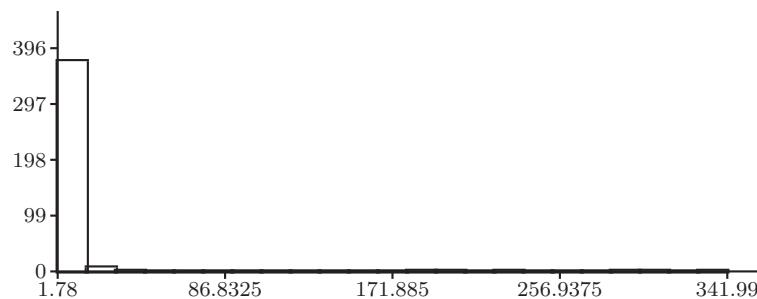


Figura C.29: Histograma de la variable MCRT-B.

En el Histograma de la variable , que se muestra en la Figura C.29, se observa que casi todos los valores están incluidos en los dos primeros intervalos, y su distribución es bastante difícil de intuir debido a la acumulación de todos los valores en uno o dos intervalos comentada anteriormente. Son tan pocos el número de valores extremos (6 valores) que casi no se observa la/s categoría/s que los incluyen.

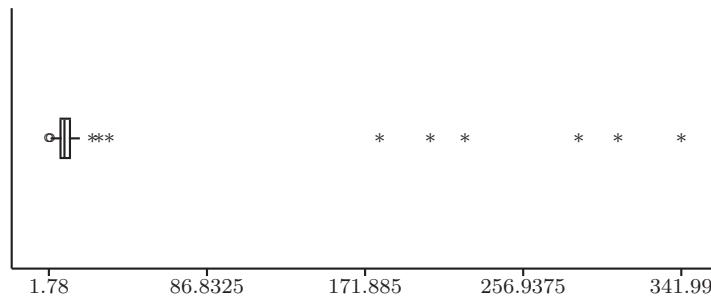


Figura C.30: Boxplot de la variable MCRT-B.

Tanto la media como la variabilidad de los datos se mantiene constante, aunque aparecen datos de valor muy diferenciado que se dan repentinamente y no de forma gradual. Se observan claramente un gran pico de valores que predomina sobre el resto; se dan en los datos nº 152 (30-I-96) a 157 (4-II-96) que corresponde a valores todos por encima de 150 días, cuando el resto no supera los 50 días. Al cerrar compuertas i tener valores de recirculación nulos, los bichos se reproducen y crecen. Esperan para hacer frente al agua de tormenta.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	9
Número de observaciones útiles	387
Media	14.3456
Mediana	10.16
Primer quartil (Q1)	8.61
Tercer quartil (Q3)	12.54
Mínimo	1.78
Máximo	341.99
Variància	999.0129
Desviación típica	31.6072
Quasi-Desviación típica	31.6481
Coeficiente de variación	2.2033

QB-B. Caudal del reactor biológico (m³/d) (Biological reactor flow).

En el Histograma de la variable que se muestra en la Figura C.31 se observa claramente el grupo de valores que quedan al margen del comportamiento habitual y otro hecho destacable es el gran número de datos que toman por valor la mediana (un 60 %).Hemos comprobado que 39.000 m³/d es la capacidad máxima del reactor biológico.

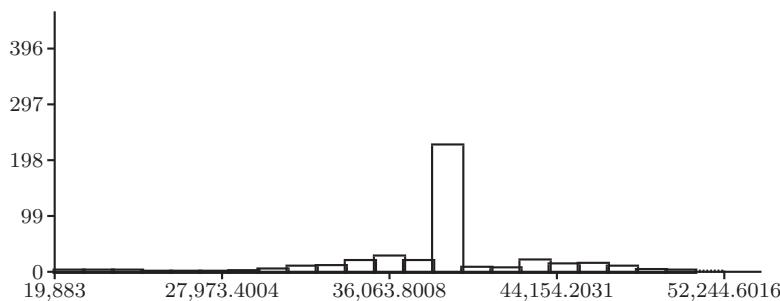


Figura C.31: Histograma de la variable QB-B.

Tenemos un caudal del reactor biológico de 38.908 metros cúbicos de caudal por día con una desviación de 4.088 metros cúbicos, con unos valores que van de 19.883 a 52.245

metros cúbicos. Existe un dato missing para esta variable (dato nº 122: 31-XI-95), aunque corresponde al del caso anteriormente comentado.

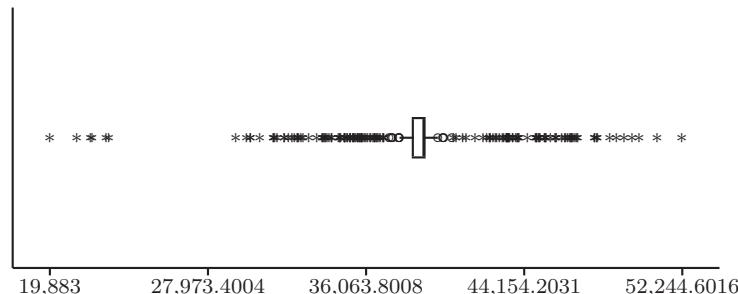


Figura C.32: Boxplot de la variable QB-B.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	1
Número de observaciones útiles	395
Media	38,907.6328
Mediana	39,000
Primer quartil (Q1)	38,510.1016
Tercer quartil (Q3)	39,000
Mínimo	19,883
Màximo	52,244.6016
Variància	16,666,253
Desviación típica	4,082.4321
Quasi-Desviación típica	4,087.6121
Coeficiente de variación	0.1049

Variable QR-G. Caudal de recirculación (metros cúbicos de agua por día).

El caudal de recirculación es de 40.964 m³/d en media con una desviación de 4.000 m³/d, con unos valores que van de 17.933 a 49.527. Existe un dato missing para esta variable (dato nº 122: 31-XI-95), aunque corresponde al caso anteriormente comentado.

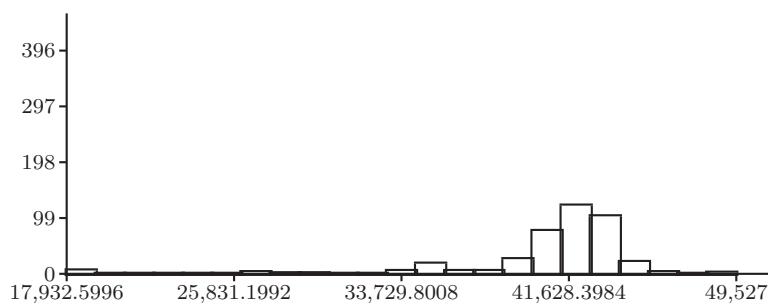


Figura C.33: Histograma de la variable QR-G.

A partir del Histograma de la variable que se muestra en la Figura C.33 se puede observar que desde los días 30-I-96 a 4-II-96 los niveles son de nuevo excepcionalmente bajos. No hay tendencias marcadas y la variabilidad es pequeña a excepción del período 25-IV-96 al 19-V-96 durante el cual el nivel del caudal de recirculación baja hasta 35.000 m³/d y se marca como un pico correspondiente a este período.

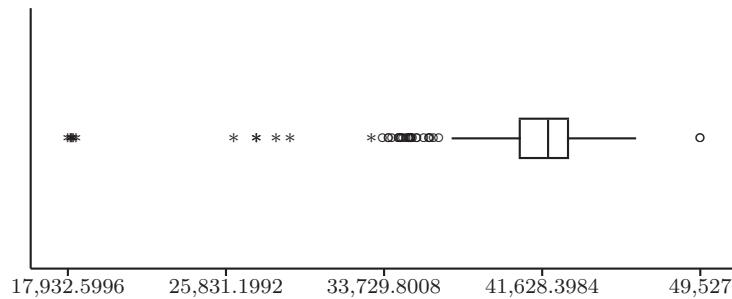


Figura C.34: Boxplot de la variable QR-G.

Estadísticos	
Número de observaciones	396
Número de observaciones missing	1
Número de observaciones útiles	395
Media	40,963.6719
Mediana	41,922.5
Primer quartil (Q1)	40,560.5
Tercer quartil (Q3)	42,864.8984
Mínimo	17,932.5996
Máximo	49,527
Variància	15,963,123
Desviación típica	3,995.3877
Quasi-Desviación típica	4,000.4316
Coeficiente de variación	0.0975

Variable QP-G. Caudal de la purga (metros cúbicos de agua por día).

El caudal de la purga es de 621.87 m³/d en media, con una desviación de 166.87 m³/d y unos valores que van de 0 a 1080. Existe un dato missing para esta variable (dato nº 122: 31-XI-95), aunque corresponde al caso inicialmente comentado.

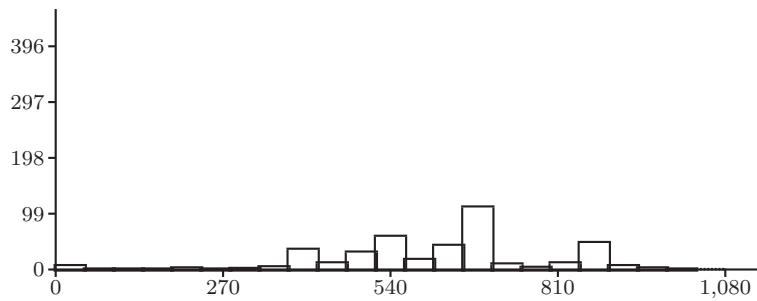


Figura C.35: Histograma de la variable QP-G.

La distribución que se muestra en la Figura C.35 presenta varios picos correspondientes a los distintos niveles de purga.

Se observa un comportamiento de los datos bastante cambiante, concretamente encontramos 5 cambios de media en los días: 18-X-95 (observación 48), 13-II-96 (observación 166), 12-VI-96 (observación 286), 27-VII-96 (observación 331) y 30-VIII-96 (observación 365). Este hecho nos podría hacer pensar que esta variable cambia según la estación del año en la que nos encontremos pero, aunque no encontramos los cambios de media próximos a los cambios

de estación (sino en la parte central de estos), si que observamos valores más elevados para verano y más bajos para invierno.

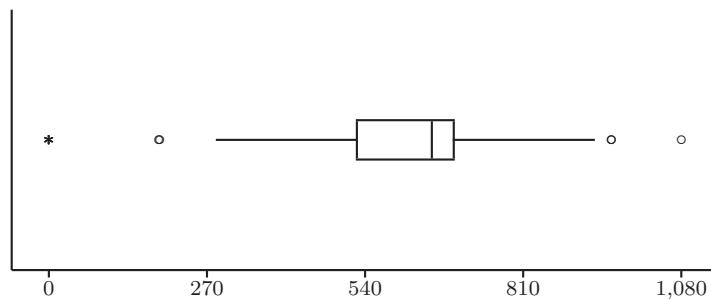


Figura C.36: Boxplot de la variable QP-G.

Estadísticos	
Número de observaciones	396
Número de observaciones missing	1
Número de observaciones útiles	395
Media	621.875
Mediana	654
Primer quartil (Q1)	527.8
Tercer quartil (Q3)	689.6
Mínimo	0
Màximo	1,080
Variància	27,776.7344
Desviación típica	166.6635
Quasi-Desviación típica	166.8749
Coeficiente de variación	0.268

La variabilidad de los datos es muy cambiante debido al patrón anteriormente comentado y aparece el mismo comportamiento anómalo entre 30-I-96 y 4-II-96, donde el caudal de recirculación es 0. Recordemos que en esta fecha el caudal de entrada era muy pequeño, de hecho responde a un cierre de compuertas. Bloquear la purga significa que los microorganismos se reproducen y aumentan considerablemente dentro del reactor biológico en estas fechas.

Variable QA-G. Afluencia de aire (metros cúbicos de aire por día).

La afluencia de aire es de 227.740 m³ de aire/d en media, con una desviación de 50.232 m³/d, con unos valores que van de 96.451 a 367.840. Existen dos datos missing para esta variable (datos nº 122: 31-XI-95 y 274: 31-V-96), aunque el primero corresponde al caso inicialmente comentado.

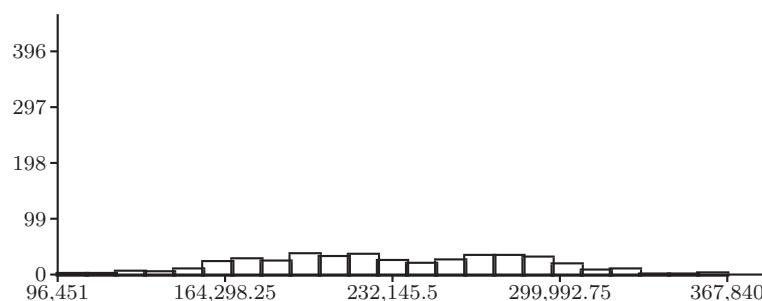


Figura C.37: Histograma de la variable QA-G.

En el Histograma de la variable que se muestra en la Figura C.37 se observa que la mayor parte de valores están concentrados en un pequeño intervalo y se observa un patrón bimodal, centrándose en 200.000 y 290.000 m³ de aire/día.

Se observa la independencia de los datos a excepción de un aparente cambio de tendencia al final que pondría en entredicho ésta. Tanto la media, a excepción del caso anterior, como la variabilidad de los datos se mantiene constante y, en este caso, no aparece el comportamiento anómalo en el grupo de datos centrales (del 152 al 157, del 30-I-96 al 4-II-96) o, no aparece tan marcado como en casos anteriores.

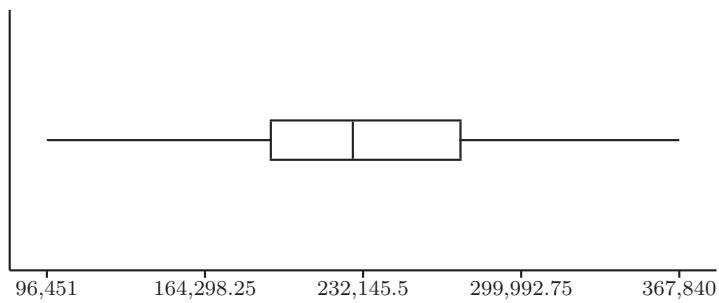


Figura C.38: Boxplot de la variable QA-G.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	2
Número de observaciones útiles	394
Media	231,663.9844
Mediana	227,740.5
Primer quartil (Q1)	193,010
Tercer quartil (Q3)	273,490
Mínimo	96,451
Máximo	367,840
Variància	2,516,859,648
Desviación típica	50,168.3125
Quasi-Desviación típica	50,232.1094
Coeficiente de variación	0.2166

C.1.4 Variables de salida

Variable PH-S. pH (unidades de pH).

La media de esta variable es de 7,5 con una desviación de 0,2027, con unos valores que van de 7 a 8, por lo que corresponde a una variable con muy poca variabilidad y un rango de variación también muy pequeño aunque superior al de las variables PH-E y PH-D. Existen 76 missings para esta variable, lo que nos obliga a trabajar con 320 datos.

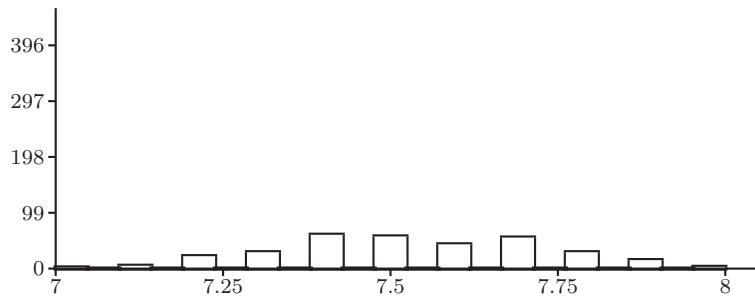


Figura C.39: Histograma de la variable PH-S.

Tanto la media como la variabilidad de los datos no se mantienen constantes y, en este caso, parece que esta última va disminuyendo a lo largo del tiempo aunque la disminución no es muy pronunciada, mientras la primera va aumentando con tres cambios de media. En el Histograma de la variable , que se muestra en la Figura C.39, se observa que los valores están distribuidos alrededor de varios picos, tal y como esperábamos debido a la tendencia creciente antes comentada.

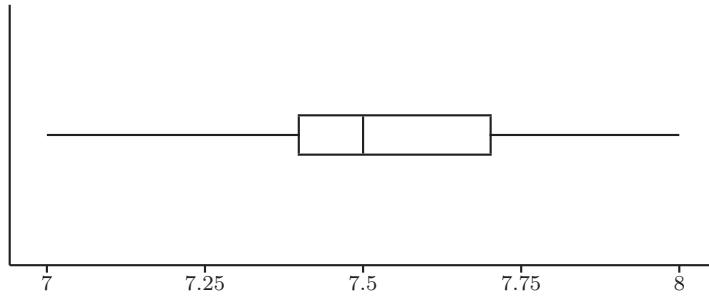


Figura C.40: Boxplot de la variable PH-S.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	76
Número de observaciones útiles	320
Media	7.5316
Mediana	7.5
Primer quartil (Q1)	7.4
Tercer quartil (Q3)	7.7
Mínimo	7
Màximo	8
Variància	0.0409
Desviación típica	0.2023
Quasi-Desviación típica	0.2026
Coeficiente de variación	0.0269

Variable SS-S. Sólidos en suspensión (mg de sólidos por litro de agua).

La media de sólidos suspendidos a la salida de la planta es de 16,885 mg con una desviación de 17,334, con unos valores que van de 2,8 a 175,8. Existen 76 missings para esta variable, lo que nos obliga a trabajar con 320 datos. Se observa que a medida que vamos pasando a fases posteriores se da una pérdida mayor de datos.

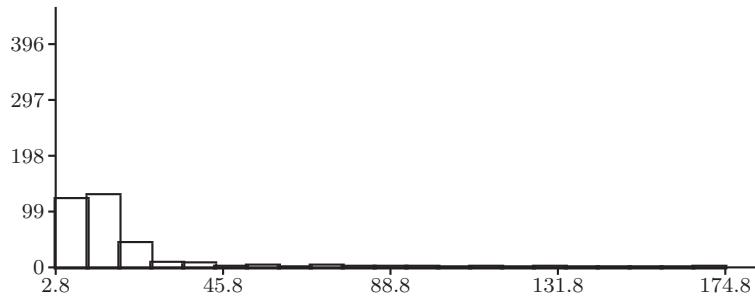


Figura C.41: Histograma de la variable SS-S.

En el Histograma de la variable , que se muestra en la Figura C.41, se observa que la mayor parte de valores están concentrados en dos o tres intervalos, y su distribución es aplanada a la derecha, es decir, predominan los valores pequeños (debido a los escasos saltos a valores elevados antes comentados).

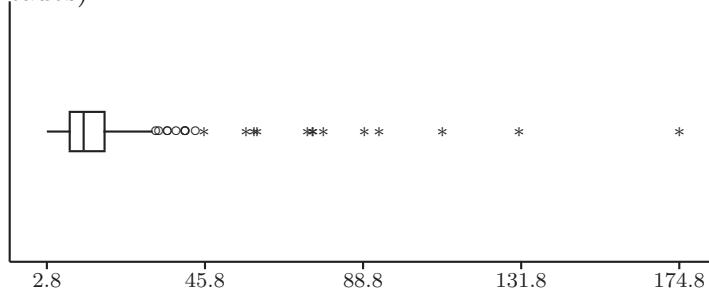


Figura C.42: Boxplot de la variable SS-S.

Tanto la media como la variabilidad de los datos se mantiene constante, aunque aparecen datos de valor elevado que se dan repentinamente y no de forma gradual. Se observa principalmente el salto que se da en el dato nº 378 (12-IX-96), aunque se dan otros de menor importancia también.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	76
Número de observaciones útiles	320
Media	16.8853
Mediana	12.8
Primer quartil (Q1)	9.3
Tercer quartil (Q3)	18.2
Mínimo	2.8
Máximo	174.8
Variància	299.536
Desviación típica	17.3071
Quasi-Desviación típica	17.3342
Coeficiente de variación	1.025

Variable SSV-S. Sólidos volátiles en suspensión (mg de sólidos por litro de agua).

La media de sólidos volátiles suspendidos es de 12,657 mg con una desviación de 13,447, con unos valores que van de 1,6 a 134,8 mg. Existen 76 missings para esta variable, lo que nos obliga a trabajar con 320 datos. Estos 76 missings coinciden con los de la variable anterior.

En el Histograma de la variable , que se muestra en la Figura C.43, se observa que la mayor parte de valores están concentrados en un intervalo, y su distribución es bastante simétrica a excepción de los valores extremadamente elevados.

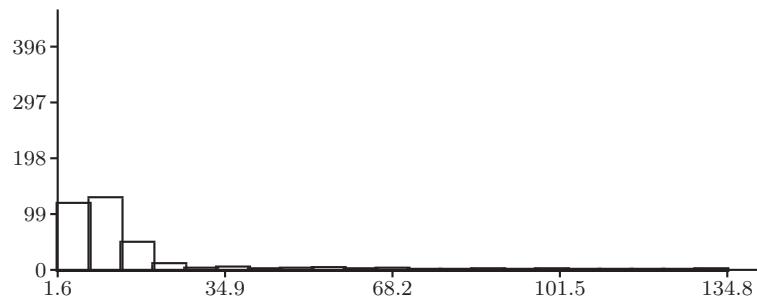


Figura C.43: Histograma de la variable SSV-S.

Tanto la media como la variabilidad de los datos se mantiene constante, aunque aparecen datos de valor elevado que se dan repentinamente y no de forma gradual, como en el caso anterior. Se observa un outlier que predomina sobre los otros posibles; se da en la el dato nº 378 (12-IX-96) que corresponde al valor máximo, y coincide con el de la variable anterior, aunque era de esperar por el significado de las variables.

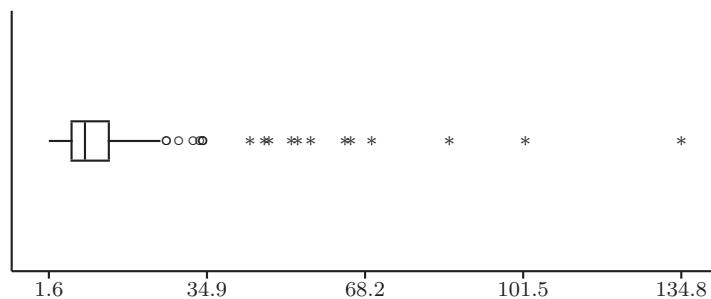


Figura C.44: Boxplot de la variable SSV-S.

Estadísticos	
Número de observaciones	396
Número de observaciones missings	76
Número de observaciones útiles	320
Media	12.6569
Mediana	9.2
Primer quartil (Q1)	6.6
Tercer quartil (Q3)	14
Mínimo	1.6
Màximo	134.8
Variància	180.2581
Desviación típica	13.426
Quasi-Desviación típica	13.4471
Coeficiente de variación	1.0608

Variable DQO-S. Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua).

La media de esta variable es de 51,25 mg de oxígeno por litro de agua con una desviación de 27.34, con unos valores que van de 9 a 163 mg. Existen 15 missings para esta variable, lo que nos obliga a trabajar con 381 datos.

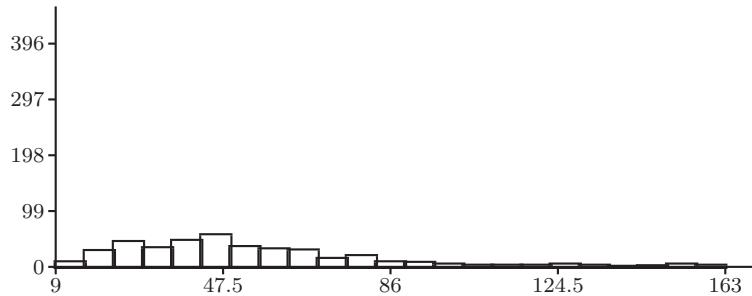


Figura C.45: Histograma de la variable DQO-S.

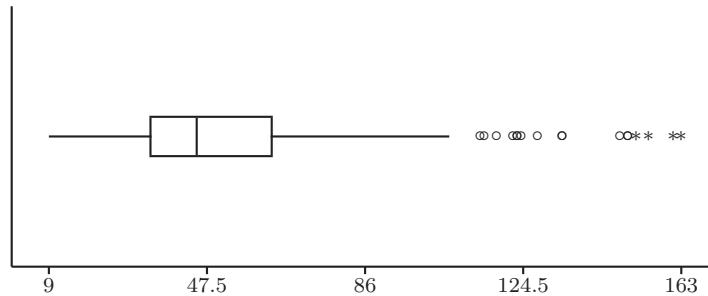


Figura C.46: Boxplot de la variable DQO-S.

Tanto la media como la variabilidad se mantienen constante, aunque observamos la presencia de al menos 4 picos o grupos de observaciones que se diferencian del resto (11-X-95 (dato nº 41) a 13-X-95 (dato nº 43); 21-III-96 (dato nº 203) a 23-III-96 (dato nº 205); 26-VII-96 (dato nº 330); 7-VIII-96 (dato nº 342)).

En el Histograma de la variable , que se muestra en la Figura C.45, se observan claramente los claramente los grupos que quedan al margen del comportamiento normal, aunque sigue un comportamiento bastante simétrico y parecido al de sus variables análogas (DQO-E y DQO-D).

Estadísticos	
Número de observaciones	396
Número de observaciones missings	15
Número de observaciones útiles	381
Media	51.2465
Mediana	45
Primer quartil (Q1)	34
Tercer quartil (Q3)	63
Mínimo	9
Máximo	163
Variància	745.2906
Desviación típica	27.3
Quasi-Desviación típica	27.3359
Coeficiente de variación	0.5327

Variable DBO-S. Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua).

La media de la materia orgánica biodegradable a la salida de la Planta es de 19,352 mg de oxígeno/l con una desviación de 11,109, con unos valores que van de 2 a 84. Existen 99 missings para esta variable distribuidos de forma uniforme, lo que nos obliga a trabajar con 297 datos solamente.

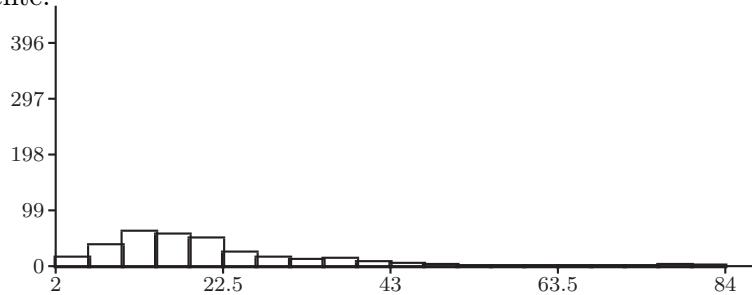


Figura C.47: Histograma de la variable DBO-S.

En el Histograma de la variable , que se muestra en la Figura C.47, se observan claramente los grupos de valores que quedan al margen del comportamiento normal, predominando los valores pequeños de la variable.

Tanto la media como la variabilidad de los datos se mantiene constante, aunque observamos la presencia de 2 valores que no siguen el comportamiento normal de la variable: 21-III-96 (dato nº 203), 27-III-96 (dato nº 209), 12-IX-96 (dato nº 378).

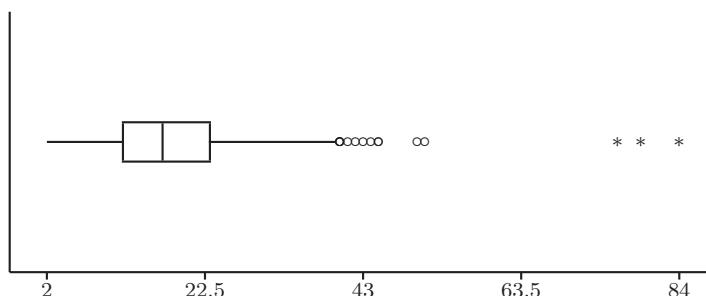


Figura C.48: Boxplot de la variable DBO-S

Estadísticos	
Número de observaciones	396
Número de observaciones missings	99
Número de observaciones útiles	297
Media	19.3515
Mediana	17
Primer quartil (Q1)	12
Tercer quartil (Q3)	23
Mínimo	2
Màximo	84
Variància	122.9995
Desviación típica	11.0905
Quasi-Desviación típica	11.1092
Coeficiente de variación	0.5731

C.2 Análisis Bivariante

C.2.1 Caudales

1. Caudal de entrada por el Caudal del reactor biológico.

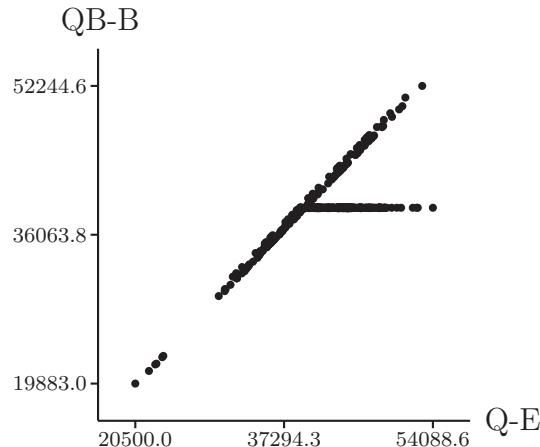


Figura C.49: Diagrama bivariante para las variables Q-E y QB-B.

Debido al gran parecido en las observaciones de estas dos primeras variables hemos realizado el gráfico se muestra en la Figura C.49 que corrobora que el caudal de entrada y el del reactor están muy relacionadas. Podríamos decir que existe una relación directa entre ellas a excepción de los casos donde la variable QB-B toma por valor 39.000 metros cúbicos (en estos casos la variable Q-E toma valores mayores o iguales a este valor).

En principio, todo el agua que entra en la Planta va al reactor biológico, sin embargo, una vez consultados los expertos, sabemos que 39.000 es la capacidad máxima del reactor en principio.

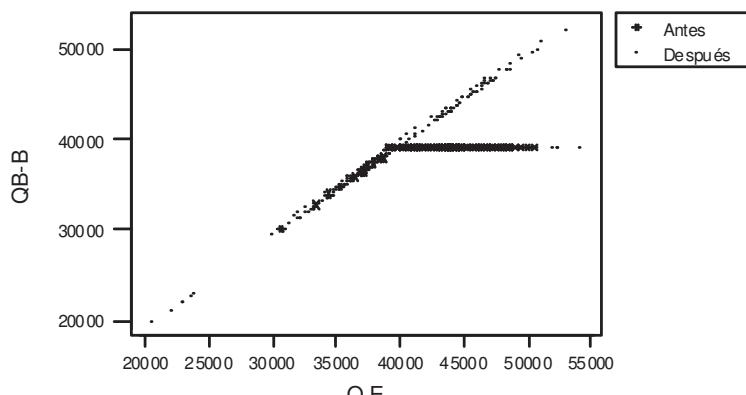


Figura C.50: Letter-plot del caudal de entrada por el caudal del reactor biológico.

Así, si el caudal de entrada superara esta cantidad, el excedente no iría al reactor y sólo 39.000 m³/d pasaría esta fase de la depuración. Esto explica la rama horizontal del gráfico. Observamos una serie de días en que el reactor biológico parece absorber todo el caudal de entrada, aun siendo éste superior a 39.000 m³/d. Esto nos hace pensar una posible ampliación del reactor a partir de cierta fecha. De hecho, el fenómeno antes comentado de saturación del reactor se produce siempre con anterioridad al 30-I-96.

Para corroborar que cuando el caudal de entrada es mayor a 39.000 m³ se bloquea el caudal del reactor biológico hasta el 30-I-96, hemos realizado un Letter Plot, ver figura C.50, que nos separa los días anteriores y posteriores a esta fecha. Después de ésta ya llega todo el

caudal de entrada al reactor biológico. Hemos conseguido una confirmación, por parte de la Planta, de que dicha ampliación tiene realmente lugar.

C.2.2 Sólidos en suspensión (SS)

2. Sólidos en suspensión (SS) en la salida y después del decantador.

La mayoría de las veces los Sólidos en suspensión a la salida de la planta están por debajo de los límites permitidos (ver sección §14.4), lo que verifica el correcto funcionamiento de la planta y el buen efecto del proceso de depuración. Los SS son eliminados sea cual sea su concentración inicial.

Con muy escasas excepciones en que entre el decantador y la salida puede aparecer un ligero aumento de SS debido, sin duda, a algún tipo de operación anómala.

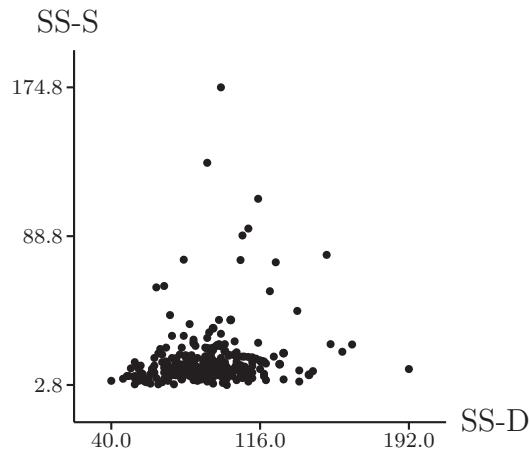


Figura C.51: Diagrama bivariante para las variables SS-D y SS-S .

3. Sólidos en suspensión (SS) en la salida y en la entrada.

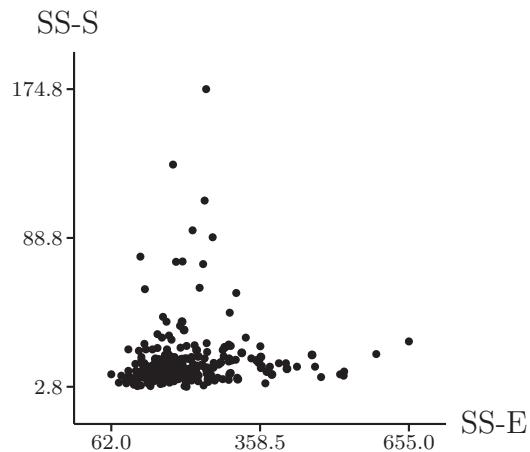


Figura C.52: Diagrama bivariante para las variables SS-E y SS-S .

En general, entre la entrada y la salida de la planta, siempre se reducen los Sólidos en Suspensión. La mayor variabilidad de los valores medios de los SS en la entrada.

C.2.3 Sólidos volátiles en suspensión (SSV).

1. Solidos volátiles en suspensión (SSV) a la entrada y después del decantador.

Aquí también se da una situación de heterocedasticidad, y aunque estamos por debajo de la bisectriz del primer cuadrante, esto indica que la decantación siempre reduce los sólidos volátiles en suspensión.

La variabilidad crece con los SSV-E. Ver figura C.53

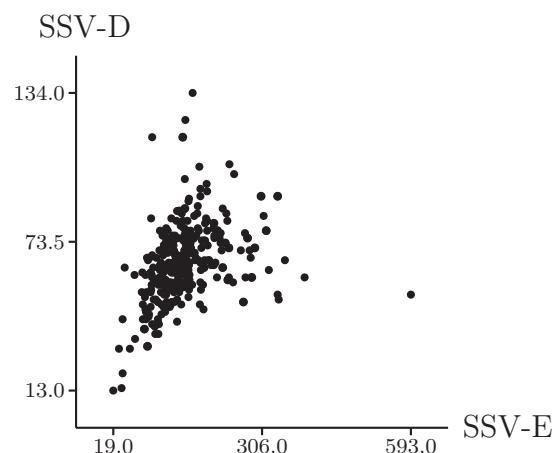


Figura C.53: Diagrama bivariante para las variables SSV-E y SSV-D.

2. Solidos volátiles en suspensión (SSV) a la salida y después del decantador.

Entre el decantador y la salida solo en contadas ocasiones pueden aumentar los Solidos volátiles en suspensión, siempre se da $SSV-D > SSV-S$.

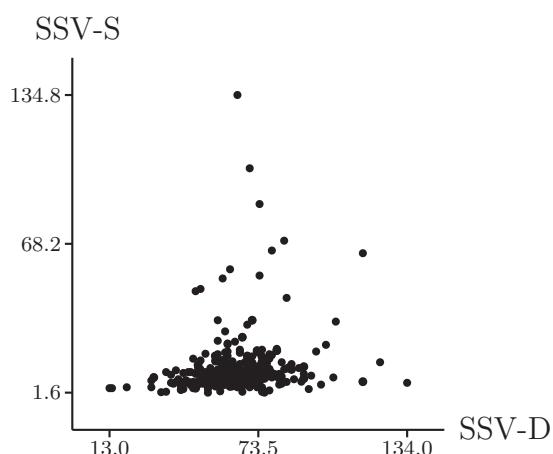


Figura C.54: Diagrama bivariante para las variables SSV-D y SSV-S.

3. Solidos volátiles en suspensión (SSV) a la salida y a la entrada.

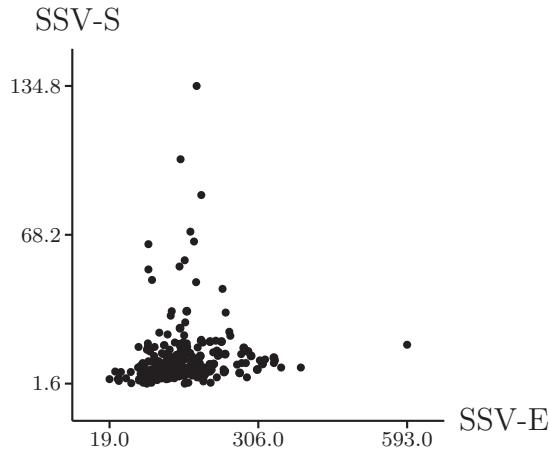


Figura C.55: Diagrama bivariante para las variables SSV-S y SSV-E.

C.2.4 Relación entre SS y SSV.

1. SS y SSV, en la entrada.

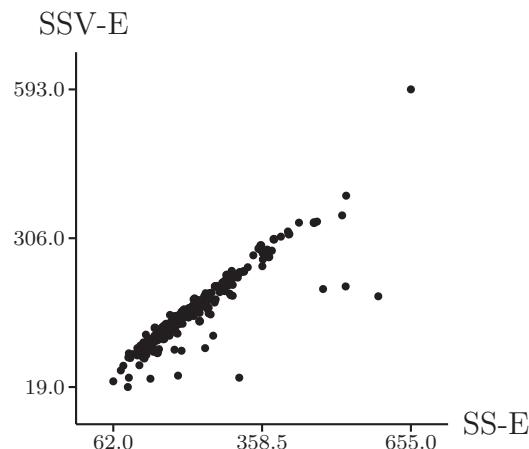


Figura C.56: Diagrama bivariante para las variables SS-E y SSV-E.

Debido al gran parecido en las observaciones de estas dos últimas variables hemos realizado este gráfico que corrobora nuestra suposición; estas dos variables están muy relacionadas y podríamos decir que existe una relación directa entre ellas (a mayor valor de los sólidos en suspensión a la entrada, mayor el valor de los sólidos volátiles en suspensión) a excepción de muy pocos valores; lo cual era de esperar por ser mediciones sobre mismas aguas y ser los sólidos volátiles en suspensión (SSV) parte de lo que contempla la variable SS. De hecho las observaciones se sitúan siempre por encima de la bisectriz del primer cuadrante.

Para corroborar este hecho numéricamente calculamos la correlación entre ambas variables. Tenemos una correlación (Pearson) de 0.91 lo que supone la existencia de una relación lineal intensa entre ambas variables.

C.2.5 Materia orgánica biodegradable (DBO)

1. DBO después del decantador y DBO en la entrada.

Debido a que se trata de la misma variable medida en instantes distintos observamos una cierta correlación (0.466) entre ambas. Siendo más pronunciada para valores bajos de las dos variables y distanciándose más para valores elevados. Ver figura C.57.

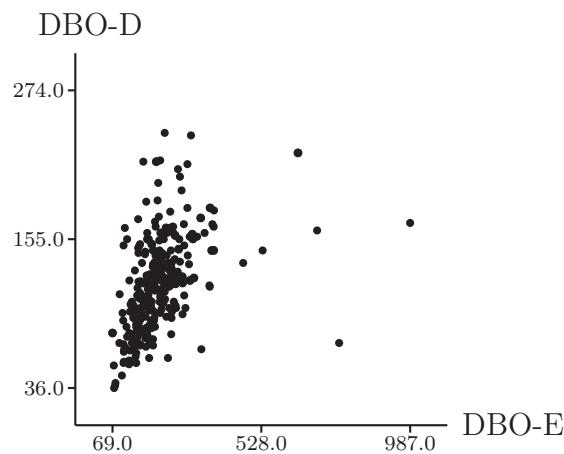


Figura C.57: Diagrama bivariante para las variables DBO-E y DBO-D.

C.2.6 Materia orgánica degradable (DQO)

1. DQO después del decantador y DQO en la entrada.

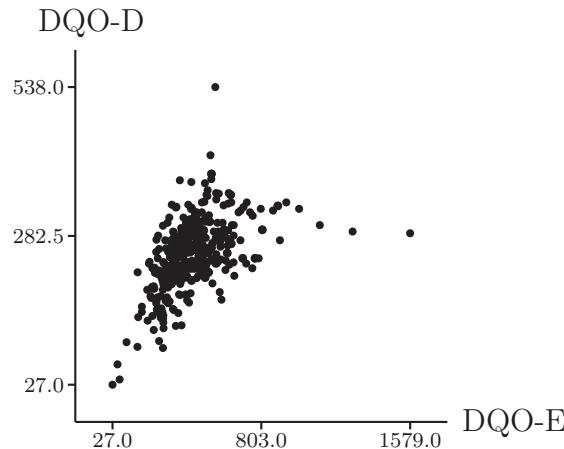


Figura C.58: Diagrama bivariante para las variables DQO-E y DQO-D.

Debido a que se trata de la misma variable medida en instantes distintos observamos una cierta correlación entre ambas. Siendo más pronunciada para valores bajos de las dos variables y distanciándose más para valores elevados. Ver figura C.58.

C.2.7 Materia orgánica biodegradable *vs* Materia orgánica degradable

1. DBO y DQO en la entrada.

Estas dos variables estan muy relacionadas y podríamos decir que existe una relación directa entre ellas (a mayor valor de una variable, mayor el valor de la otra) a excepción un solo valor; lo cual era de esperar por ser mediciones sobre mismas aguas. Como en el caso anterior y

para corroborar este hecho numéricamente calculamos la correlación entre ambas variables. Tenemos una correlación (Pearson) de 0.754, lo que supone la existencia de gran correlación entre ambas variables. Ver figura C.59.

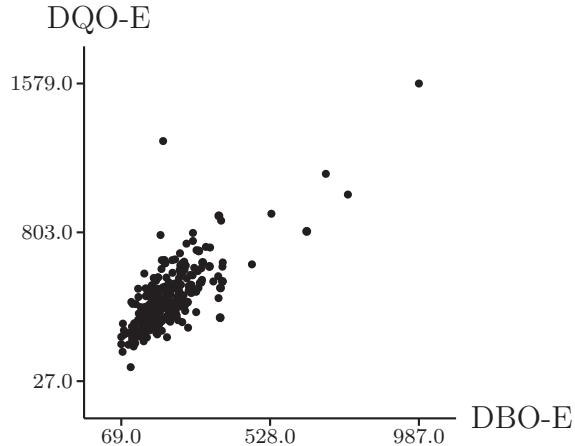


Figura C.59: Diagrama bivariante para las variables DQO-E y DBO-E.

2. DBO después del decantador y DQO después del decantador.

Debido al gran parecido en las observaciones de estas dos últimas variables hemos realizado este gráfico que corrobora nuestra suposición; estas dos variables están relacionadas (correlación de 0.575), aunque no en gran medida, y existe una relación directa entre ellas (a mayor valor de una variable, mayor el valor de la otra) a excepción de muy pocos valores; lo cual era de esperar por ser mediciones sobre datos en la misma fase de depuración. Ver figura C.60.

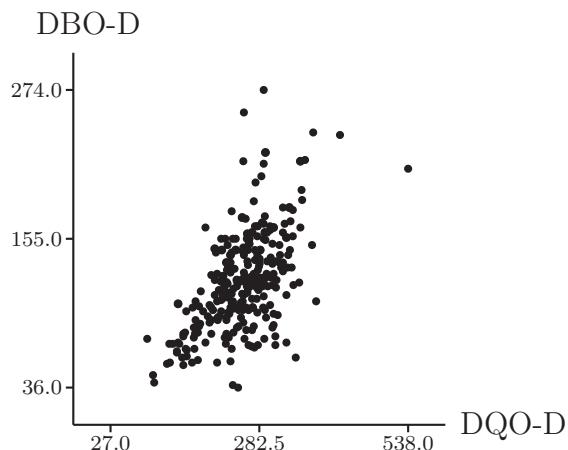


Figura C.60: Diagrama bivariante para las variables DBO-D y DQO-D.

C.2.8 Comportamiento del PH

Debido a una posible relación del PH en la salida de la planta con PH-E (0.57) y PH-D (0.659) realizamos estos los gráficos bivariantes Figuras C.61 y C.62 pero, en este caso no está tan clara esta relación, aunque si se observa una correspondencia de valores elevados con elevados y bajos con bajos (asociación positiva).

El PH de salida del agua residual podría venir influenciado por el caudal del reactor biológico (QB-B) y para comprobar esta suposición estudiamos la relación de ambas variables mediante la correlación (-0.001) que, en este caso, desmiente esta hipótesis. Ver gráfico bivariante en la Figura C.63.

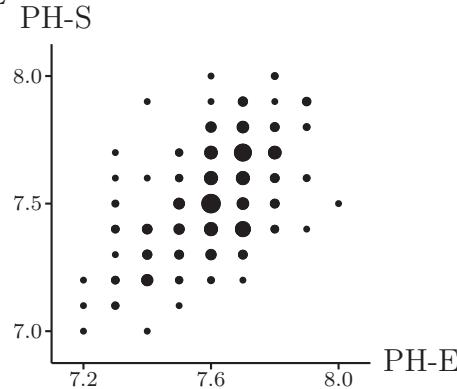


Figura C.61: Diagrama bivariante para las variables PH-E y PH-S.

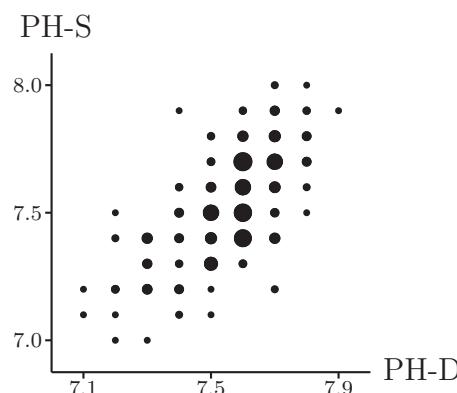


Figura C.62: Diagrama bivariante para las variables PH-D y PH-S.

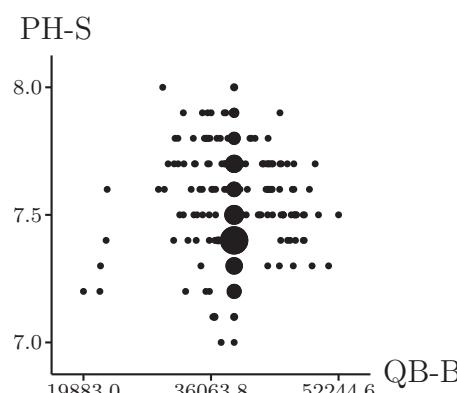


Figura C.63: Diagrama bivariante para las variables QB-B y PH-S.

C.2.9 Licor Mezcla. MLSS-B y MLVSS-B.

Debido al gran parecido en las observaciones de estas dos últimas variables hemos realizado este gráfico que corrobora nuestra suposición; estas dos variables estan muy relacionadas y podríamos decir que existe una relación directa entre ellas (a mayor valor de una variable, mayor el valor de la otra); lo cual era de esperar por ser mediciones sobre mismas aguas. En el caso de los valores elevados se produce un cambio de pendiente de la recta a ajustar. Naturalmente existen otras muchas relaciones de interés para estudiar, las que se analizaran mas adelante.

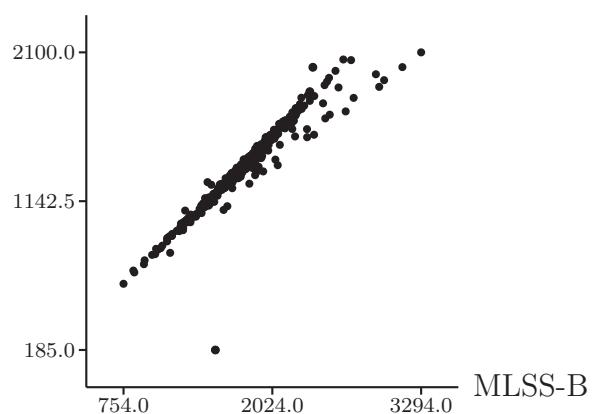


Figura C.64: Diagrama bivariante para las variables MLSS-B y MLVSS-B.

Anexo D

Análisis descriptivo por clases, planta catalana

D.1 Clustering Based on Rules planta catalunya

D.1.1 Class-variable: 2-classes $[P2_{Gi1,R1}^{En,G}]$

	CLASSE	classer393	classer392
VARIABLE	N = 396	$n_c = 390$	$n_c = 6$
Q-E	\bar{X}	42,112.9453	22,563.7988
	S	4,559.2437	1,168.8481
	min	29,920	20,500
	max	54,088.6016	23,662.9004
	N*	1	0
QB-B	\bar{X}	39,171.5156	21,799.3672
	S	3,514.9504	1,110.0852
	min	29,397.3008	19,883
	max	52,244.6016	22,891
	N*	1	0
QR-G	\bar{X}	41,315.8438	18,131.5156
	S	2,839.8806	138.9898
	min	26,218	17,932.5996
	max	49,527	18,343.5
	N*	1	0
QP-G	\bar{X}	631.4669	0
	S	149.0143	0
	min	188	0
	max	1,080	0
	N*	1	0
QA-G	\bar{X}	233,346.9844	122,829.3359
	S	48,712	15,340.5811
	min	124,120	96,451
	max	367,840	143,151
	N*	2	0
FE-E	\bar{X}	46.0835	2.6
	S	13.1221	5.8138
	min	0	0
	max	89.8	13
	N*	2	1
PH-E	\bar{X}	7.6191	7.5
	S	0.1459	0.1225
	min	7.2	7.3
	max	8	7.6
	N*	70	1
SS-E	\bar{X}	213.0935	142
	S	86.6327	98.1122
	min	62	77
	max	655	313
	N*	69	1

	CLASSE	classer393	classer392
VARIABLE	N = 396	$n_c = 390$	$n_c = 6$
SSV-E	\bar{X}	160.9564	35.8
	S	63.4892	11.3666
	min	30	19
	max	593	51
	N*	69	1
DQO-E	\bar{X}	447.4476	110.1667
	S	162.5791	69.6604
	min	100	27
	max	1.579	180
	N*	8	0
DBO-E	\bar{X}	217.3265	73
	S	102.1115	
	min	69	73
	max	987	73
	N*	96	5
PH-D	\bar{X}	7.5594	7.6
	S	0.1457	0.1732
	min	7.1	7.3
	max	7.9	7.7
	N*	70	1
SS-D	\bar{X}	90.081	60
	S	20.1673	23.5053
	min	48	40
	max	192	98
	N*	69	1
SSV-D	\bar{X}	66.0654	23.8
	S	14.9764	12.2147
	min	30	13
	max	134	42
	N*	69	1
DQO-D	\bar{X}	252.769	88.3333
	S	57.0698	57.5175
	min	90	27
	max	538	161
	N*	9	0
DBO-D	\bar{X}	122.0903	54
	S	39.4353	
	min	36	54
	max	274	54
	N*	91	5
PH-S	\bar{X}	7.5346	7.34
	S	0.2019	0.1674
	min	7	7.2
	max	8	7.6
	N*	75	1
SS-S	\bar{X}	17.0454	6.8
	S	17.4224	1.2
	min	2.8	5.2
	max	174.8	8
	N*	75	1
SSV-S	\bar{X}	12.8006	3.6
	S	13.5045	0.4899
	min	1.6	2.8
	max	134.8	4
	N*	75	1
DQO-S	\bar{X}	51.7171	21.8333
	S	27.1773	22.1035
	min	9	9
	max	163	66
	N*	15	0

	CLASSE	classer393	classer392
VARIABLE	N = 396	$n_c = 390$	$n_c = 6$
DBO-S	\bar{X}	19.4	5
	S	11.0965	
	min	2	5
	max	84	5
	N*	94	5
V30-B	\bar{X}	261.5141	339.3333
	S	134.2213	34.6275
	min	77	310
	max	770	383
	N*	1	0
MLSS-B	\bar{X}	1,749.4255	2,929.8333
	S	335.3853	259.174
	min	754	2,589
	max	2,978	3,294
	N*	7	0
MLVSS-B	\bar{X}	1,335.3995	1,937.1666
	S	258.6565	105.8563
	min	185	1,807
	max	2,054	2,100
	N*	7	0
MCRT-B	\bar{X}	10.5115	258.215
	S	3.2684	63.4712
	min	1.8	179.8
	max	34.4	341.99
	N*	9	0

D.1.2 Class variable: 3-classes $[P3_{G1}^{EnW,G}]$

	CLASSE	classer392	classer391	classer389
VARIABLE	$N = 396$	$n_c = 6$	$n_c = 320$	$n_c = 70$
Q-E	\bar{X}	22,563.7988	42,234.5156	41,558.9414
	S	1,168.8481	4,070.2441	6,336.9009
	min	20,500	29,920	30,592.1992
	max	23,662.9004	52,255.8008	54,088.6016
	N*	0	1	0
QB-B	\bar{X}	21,799.3672	39,200.4219	39,039.8125
	S	1,110.0852	3,019.0647	5,238.8672
	min	19,883	29,397.3008	29,936.8008
	max	22,891	49,695.8008	52,244.6016
	N*	0	1	0
QR-G	\bar{X}	18,131.5156	41,723.4023	39,458.6211
	S	138.9898	2,557.1418	3,307.1287
	min	17,932.5996	27,351	26,218
	max	18,343.5	49,527	43,298.1016
	N*	0	1	0
QP-G	\bar{X}	0	643.7445	575.5157
	S	0	146.2209	149.8517
	min	0	327.6	188
	max	0	1,080	866.7
	N*	0	1	0
QA-G	\bar{X}	122,829.3359	237,017.125	216,673.9375
	S	15,340.5811	49,651.9219	40,476.7305
	min	96,451	136,371	124,120
	max	143,151	367,840	324,470
	N*	0	2	0
FE-E	\bar{X}	2.6	46.2556	45.3014
	S	5.8138	13.4367	11.6417
	min	0	0	0
	max	13	89.8	65.6
	N*	1	2	0
PH-E	\bar{X}	7.5	7.6231	7.6031
	S	0.1225	0.1381	0.175
	min	7.3	7.2	7.2
	max	7.6	7.9	8
	N*	1	65	5
SS-E	\bar{X}	142	228.8078	152.3788
	S	98.1122	88.2386	42.5635
	min	77	82	62
	max	313	655	266
	N*	1	65	4
SSV-E	\bar{X}	35.8	174.1529	109.9697
	S	11.3666	62.9958	31.8419
	min	19	60	30
	max	51	593	193
	N*	1	65	4
DQO-E	\bar{X}	110.1667	473.4505	329.4928
	S	69.6604	164.4084	82.7448
	min	27	158	100
	max	180	1,579	595
	N*	0	7	1
DBO-E	\bar{X}	73	235.5107	147.8689
	S		105.0782	45.1108
	min	73	90	69
	max	73	987	258
	N*	5	87	9
PH-D	\bar{X}	7.6	7.5694	7.52
	S	0.1732	0.1343	0.1813
	min	7.3	7.2	7.1
	max	7.7	7.9	7.8
	N*	1	65	5

	CLASSE	classer392	classer391	classer389
VARIABLE	N = 396	$n_c = 6$	$n_c = 320$	$n_c = 70$
SS-D	\bar{X}	60	94.302	73.7727
	S	23.5053	18.3728	18.528
	min	40	63	48
	max	98	192	136
	N*	1	65	4
SSV-D	\bar{X}	23.8	69.9176	51.1818
	S	12.2147	13.1944	11.8645
	min	13	47	30
	max	42	134	99
	N*	1	65	4
DQO-D	\bar{X}	88.3333	266.4936	190.7101
	S	57.5175	50.6042	41.4364
	min	27	90	100
	max	161	538	269
	N*	0	8	1
DBO-D	\bar{X}	54	130.7773	88.1967
	S		36.7997	30.1241
	min	54	56	36
	max	54	274	171
	N*	5	82	9
PH-S	\bar{X}	7.34	7.5516	7.4692
	S	0.1674	0.1952	0.215
	min	7.2	7	7
	max	7.6	8	8
	N*	1	70	5
SS-S	\bar{X}	6.8	18.8402	10.2742
	S	1.2	19.0928	4.1061
	min	5.2	2.8	3.2
	max	8	174.8	20
	N*	1	71	4
SSV-S	\bar{X}	3.6	14.2325	7.3985
	S	0.4899	14.7749	3.2594
	min	2.8	1.6	1.6
	max	4	134.8	18
	N*	1	71	4
DQO-S	\bar{X}	21.8333	53.6696	43.058
	S	22.1035	28.2431	19.7811
	min	9	9	9
	max	66	163	94
	N*	0	14	1
DBO-S	\bar{X}	5	20.9677	13.3607
	S		11.4972	6.5702
	min	5	4	2
	max	5	84	26
	N*	5	85	9
V30-B	\bar{X}	339.3333	272.1348	213.1143
	S	34.6275	143.0995	63.4943
	min	310	77	115
	max	383	770	380
	N*	0	1	0
MLSS-B	\bar{X}	2,929.8333	1,745.2325	1,768.5072
	S	259.174	327.1617	372.5179
	min	2,589	754	846
	max	3,294	2,696	2,978
	N*	0	6	1
MLVSS-B	\bar{X}	1,937.1666	1,337.1625	1,327.3768
	S	105.8563	264.3528	232.577
	min	1,807	185	684
	max	2,100	2,054	1,921
	N*	0	6	1

	CLASSE	classer392	classer391	classer389
VARIABLE	N = 396	$n_c = 6$	$n_c = 320$	$n_c = 70$
MCRT-B	\bar{X}	258.215	10.2542	11.6754
	S	63.4712	3.0018	4.0998
	min	179.8	1.8	6.9
	max	341.99	28.8	34.4
	N*	0	8	1

D.1.3 Class variable: 4-classes [$P4_{G1}^{EnW,G}$]

	CLASSE	classer383	classer392	classer389	classer390
VARIABLE	N = 396	n _c = 34	n _c = 6	n _c = 70	n _c = 286
Q-E	\bar{X}	41,346.0195	22,563.7988	41,558.9414	42,340.5078
	S	3,681.1863	1,168.8481	6,336.9009	4,107.3608
	min	34,284.3984	20,500	30,592.1992	29,920
	max	50,500.5	23,662.9004	54,088.6016	52,255.8008
	N*	0	0	0	1
QB-B	\bar{X}	38,377.4492	21,799.3672	39,039.8125	39,298.6016
	S	1,459.3115	1,110.0852	5,238.8672	3,141.3909
	min	33,549.3984	19,883	29,936.8008	29,397.3008
	max	39,000	22,891	52,244.6016	49,695.8008
	N*	0	0	0	1
QR-G	\bar{X}	41,437.9883	18,131.5156	39,458.6211	41,757.4375
	S	3,135.0498	138.9898	3,307.1287	2,483.9214
	min	28,343.8008	17,932.5996	26,218	27,351
	max	44,568.6016	18,343.5	43,298.1016	49,527
	N*	0	0	0	1
QP-G	\bar{X}	575.5588	0	575.5157	651.879
	S	90.8449	0	149.8517	149.5262
	min	385.9	0	188	327.6
	max	831.1	0	866.7	1,080
	N*	0	0	0	1
QA-G	\bar{X}	250,470.3281	122,829.3359	216,673.9375	235,406.5156
	S	55,528.9688	15,340.5811	40,476.7305	48,760.4805
	min	156,320	96,451	124,120	136,371
	max	331,990	143,151	324,470	367,840
	N*	0	0	0	2
FE-E	\bar{X}	45.85	2.6	45.3014	46.3042
	S	12.2458	5.8138	11.6417	13.5915
	min	0	0	0	0
	max	63.3	13	65.6	89.8
	N*	0	1	0	2
PH-E	\bar{X}	7.5469	7.5	7.6031	7.6341
	S	0.1523	0.1225	0.175	0.1328
	min	7.3	7.3	7.2	7.2
	max	7.8	7.6	8	7.9
	N*	2	1	5	63
SS-E	\bar{X}	377.375	142	152.3788	207.4888
	S	109.5627	98.1122	42.5635	59.95
	min	114	77	62	82
	max	655	313	266	480
	N*	2	1	4	63
SSV-E	\bar{X}	276.7188	35.8	109.9697	159.435
	S	84.1095	11.3666	31.8419	42.646
	min	92	19	30	60
	max	593	51	193	336
	N*	2	1	4	63
DQO-E	\bar{X}	665.4688	110.1667	329.4928	451.5836
	S	227.651	69.6604	82.7448	140.3141
	min	414	27	100	158
	max	1,579	180	595	1,279
	N*	2	0	1	5
DBO-E	\bar{X}	394.2069	73	147.8689	212.951
	S	185.3937		45.1108	61.3901
	min	220	73	69	90
	max	987	73	258	382
	N*	5	5	9	82
PH-D	\bar{X}	7.5219	7.6	7.52	7.5762
	S	0.154	0.1732	0.1813	0.1301
	min	7.3	7.3	7.1	7.2
	max	7.8	7.7	7.8	7.9
	N*	2	1	5	63

	CLASSE	classer383	classer392	classer389	classer390
VARIABLE	N = 396	$n_c = 34$	$n_c = 6$	$n_c = 70$	$n_c = 286$
SS-D	\bar{X}	89.3438	60	73.7727	95.0135
	S	13.5137	23.5053	18.528	18.8855
	min	68	40	48	63
	max	112	98	136	192
	N*	2	1	4	63
SSV-D	\bar{X}	67.3438	23.8	51.1818	70.287
	S	11.6525	12.2147	11.8645	13.3841
	min	49	13	30	47
	max	92	42	99	134
	N*	2	1	4	63
DQO-D	\bar{X}	249.9062	88.3333	190.7101	268.3893
	S	34.3768	57.5175	41.4364	51.8465
	min	186	27	100	90
	max	329	161	269	538
	N*	2	0	1	6
DBO-D	\bar{X}	131.9655	54	88.1967	130.6124
	S	35.9349		30.1241	36.9996
	min	67	54	36	56
	max	224	54	171	274
	N*	5	5	9	77
PH-S	\bar{X}	7.5031	7.34	7.4692	7.5587
	S	0.1769	0.1674	0.215	0.1971
	min	7.2	7.2	7	7
	max	7.8	7.6	8	8
	N*	2	1	5	68
SS-S	\bar{X}	14.6531	6.8	10.2742	19.4576
	S	5.3777	1.2	4.1061	20.2833
	min	4.8	5.2	3.2	2.8
	max	29	8	20	174.8
	N*	2	1	4	69
SSV-S	\bar{X}	11.2875	3.6	7.3985	14.6668
	S	4.0185	0.4899	3.2594	15.7113
	min	4.4	2.8	1.6	1.6
	max	19	4	18	134.8
	N*	2	1	4	69
DQO-S	\bar{X}	54.0594	21.8333	43.058	53.6241
	S	18.3737	22.1035	19.7811	29.2031
	min	20	9	9	9
	max	95	66	94	163
	N*	2	0	1	12
DBO-S	\bar{X}	17.3345	5	13.3607	21.4791
	S	8.4495		6.5702	11.79
	min	6	5	2	4
	max	35	5	26	84
	N*	5	5	9	80
V30-B	\bar{X}	262.2353	339.3333	213.1143	273.3158
	S	130.3929	34.6275	63.4943	144.7074
	min	140	310	115	77
	max	760	383	380	770
	N*	0	0	0	1
MLSS-B	\bar{X}	1,691.1471	2,929.8333	1,768.5072	1,751.8
	S	263.5027	259.174	372.5179	333.8648
	min	1,046	2,589	846	754
	max	2,248	3,294	2,978	2,696
	N*	0	0	1	6
MLVSS-B	\bar{X}	1,251.1765	1,937.1666	1,327.3768	1,347.6035
	S	334.1502	105.8563	232.577	253.3404
	min	185	1,807	684	611
	max	1,726	2,100	1,921	2,054
	N*	0	0	1	6

	CLASSE	classer383	classer392	classer389	classer390
VARIABLE	N = 396	$n_c = 34$	$n_c = 6$	$n_c = 70$	$n_c = 286$
MCRT-B	\bar{X}	10.3438	258.215	11.6754	10.2439
	S	2.1918	63.4712	4.0998	3.0837
	min	6.2	179.8	6.9	1.8
	max	16	341.99	34.4	28.8
	N*	2	0	1	6

Anexo E

BbD, BbIR y $\mathcal{R}(\mathcal{P}_\xi^*)$ planta catalana

E.1 BbD para $\mathcal{P}_2^* \equiv \mathcal{P}2_{Gi1,R1}^{EnW,G}$

Patrón: centro abierto para la variable Q-E de la partición \mathcal{P}_2^*

$$I_1^{Q-E,2} = [20500.0, 23662.9]$$

$$I_2^{Q-E,2} = (23662.9, 29920.0)$$

$$I_3^{Q-E,2} = [29920.0, 54088.6]$$

Patrón: centro abierto para la variable QB-B de la partición \mathcal{P}_2^*

$$I_1^{QB-B,2} = [19883.0, 22891.0]$$

$$I_2^{QB-B,2} = (22891.0, 29397.3)$$

$$I_3^{QB-B,2} = [29397.3, 52244.6]$$

Patrón: centro abierto para la variable QR-G de la partición \mathcal{P}_2^*

$$I_1^{QR-G,2} = [17932.6, 18343.5]$$

$$I_2^{QR-G,2} = (18343.5, 26218.0)$$

$$I_3^{QR-G,2} = [26218.0, 49527.0]$$

Patrón: centro abierto para la variable QP-G de la partición \mathcal{P}_2^*

$$I_1^{QP-G,2} = [0.0, 0.0]$$

$$I_2^{QP-G,2} = (0.0, 188.0)$$

$$I_3^{QP-G,2} = [188.0, 1080.0]$$

Patrón: centro cerrado para la variable QA-G de la partición en \mathcal{P}_2^*

$$I_1^{QA-G,2} = [96451.0, 124120.0)$$

$$I_2^{QA-G,2} = [124120.0, 143151.0]$$

$$I_3^{QA-G,2} = (143151.0, 367840.0]$$

Patrón: centro cerrado para la variable FE-E de la partición en \mathcal{P}_2^*

$$I_1^{FE-E,2} = [0.0, 0.0)$$

$$I_2^{FE-E,2} = [0.0, 13.0]$$

$$I_3^{FE-E,2} = (13.0, 89.8]$$

Patrón: centro cerrado para la variable PH-E de la partición en \mathcal{P}_2^*

$$I_1^{PH-E,2} = [7.2, 7.3)$$

$$\begin{aligned}I_2^{PH-E,2} &= [7.3, 7.6] \\I_3^{PH-E,2} &= (7.6, 8.0]\end{aligned}$$

Patrón: centro cerrado para la variable SS-E de la partición en \mathcal{P}_2^*

$$\begin{aligned}I_1^{SS-E,2} &= [62.0, 77.0) \\I_2^{SS-E,2} &= [77.0, 313.0] \\I_3^{SS-E,2} &= (313.0, 655.0]\end{aligned}$$

Patrón: centro cerrado para la variable SSV-E de la partición en \mathcal{P}_2^*

$$\begin{aligned}I_1^{SSV-E,2} &= [19.0, 30.0) \\I_2^{SSV-E,2} &= [30.0, 51.0] \\I_3^{SSV-E,2} &= (51.0, 593.0]\end{aligned}$$

Patrón: centro cerrado para la variable DQO-E de la partición en \mathcal{P}_2^*

$$\begin{aligned}I_1^{DQO-E,2} &= [27.0, 100.0) \\I_2^{DQO-E,2} &= [100.0, 180.0] \\I_3^{DQO-E,2} &= (180.0, 1579.0]\end{aligned}$$

Patrón: centro cerrado para la variable DBO-E de la partición en \mathcal{P}_2^*

$$\begin{aligned}I_1^{DBO-E,2} &= [69.0, 73.0) \\I_2^{DBO-E,2} &= [73.0, 73.0] \\I_3^{DBO-E,2} &= (73.0, 987.0]\end{aligned}$$

Patrón: centro cerrado para la variable PH-D de la partición en \mathcal{P}_2^*

$$\begin{aligned}I_1^{PH-D,2} &= [7.1, 7.3) \\I_2^{PH-D,2} &= [7.3, 7.7] \\I_3^{PH-D,2} &= (7.7, 7.9]\end{aligned}$$

Patrón: centro cerrado para la variable SS-D de la partición en \mathcal{P}_2^*

$$\begin{aligned}I_1^{SS-D,2} &= [40.0, 48.0) \\I_2^{SS-D,2} &= [48.0, 98.0] \\I_3^{SS-D,2} &= (98.0, 192.0]\end{aligned}$$

Patrón: centro cerrado para la variable SSV-D de la partición en \mathcal{P}_2^*

$$\begin{aligned}I_1^{SSV-D,2} &= [13.0, 30.0) \\I_2^{SSV-D,2} &= [30.0, 42.0] \\I_3^{SSV-D,2} &= (42.0, 134.0]\end{aligned}$$

Patrón: centro cerrado para la variable DQO-D de la partición en \mathcal{P}_2^*

$$\begin{aligned}I_1^{DQO-D,2} &= [27.0, 90.0) \\I_2^{DQO-D,2} &= [90.0, 161.0] \\I_3^{DQO-D,2} &= (161.0, 538.0]\end{aligned}$$

Patrón: centro cerrado para la variable DBO-D de la partición en \mathcal{P}_2^*

$$\begin{aligned}I_1^{DBO-D,2} &= [36.0, 54.0) \\I_2^{DBO-D,2} &= [54.0, 54.0] \\I_3^{DBO-D,2} &= (54.0, 274.0]\end{aligned}$$

Patrón: centro cerrado para la variable PH-S de la partición \mathcal{P}_2^*

$$I_1^{PH-S,2} = [7.0, 7.2)$$

$$I_2^{PH-S,2} = [7.2, 7.6]$$

$$I_3^{PH-S,2} = (7.6, 8.0]$$

Patrón: centro cerrado para la variable SS-S de la partición \mathcal{P}_2^*

$$I_1^{SS-S,2} = [2.8, 5.2)$$

$$I_2^{SS-S,2} = [5.2, 8.0]$$

$$I_3^{SS-S,2} = (8.0, 174.8]$$

Patrón: centro cerrado para la variable SSV-S de la partición \mathcal{P}_2^*

$$I_1^{SSV-S,2} = [1.6, 2.8)$$

$$I_2^{SSV-S,2} = [2.8, 4.0]$$

$$I_3^{SSV-S,2} = (4.0, 134.8]$$

Patrón: centro cerrado para la variable DQO-S de la partición \mathcal{P}_2^*

$$I_1^{DQO-S,2} = [9.0, 66.0]$$

$$I_3^{DQO-S,2} = (66.0, 163.0]$$

Patrón: centro cerrado para la variable DBO-S de la partición \mathcal{P}_2^*

$$I_1^{DBO-S,2} = [2.0, 5.0)$$

$$I_2^{DBO-S,2} = [5.0, 5.0]$$

$$I_3^{DBO-S,2} = (5.0, 84.0]$$

Patrón: centro cerrado para la variable V30-B de la partición \mathcal{P}_2^*

$$I_1^{V30-B,2} = [77.0, 310.0)$$

$$I_2^{V30-B,2} = [310.0, 383.0]$$

$$I_3^{V30-B,2} = (383.0, 770.0]$$

Patrón: centro cerrado para la variable MLSS-B de la partición \mathcal{P}_2^*

$$I_1^{MLSS-B,2} = [754.0, 2589.0)$$

$$I_2^{MLSS-B,2} = [2589.0, 2978.0]$$

$$I_3^{MLSS-B,2} = (2978.0, 3294.0]$$

Patrón: centro cerrado para la variable MLVSS-B de la partición \mathcal{P}_2^*

$$I_1^{MLVSS-B,2} = [185.0, 1807.0)$$

$$I_2^{MLVSS-B,2} = [1807.0, 2054.0]$$

$$I_3^{MLVSS-B,2} = (2054.0, 2100.0]$$

Patrón: centro abierto para la variable MCRT-B de la partición \mathcal{P}_2^*

$$I_1^{MCRT-B,2} = [1.8, 34.4]$$

$$I_2^{MCRT-B,2} = (34.4, 179.8)$$

$$I_3^{MCRT-B,2} = [179.8, 341.99]$$

E.2 $\mathcal{R}(\mathcal{P}_2^*)$

$$\mathcal{R}(\mathcal{P}_2^*) = \{ \begin{array}{l} r_{1, classser392}^{Q-E} : x_{Q-E,i} \in [20500.0, 23662.9] \xrightarrow{1.0} classer392, \\ r_{3, classser392}^{Q-E} : x_{Q-E,i} \in [29920.0, 54088.6] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{QB-B} : x_{QB-B,i} \in [19883.0, 22891.0] \xrightarrow{1.0} classer392, \\ r_{3, classser392}^{QB-B} : x_{QB-B,i} \in [29397.3, 52244.6] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{QR-G} : x_{QR-G,i} \in [17932.6, 18343.5] \xrightarrow{1.0} classer392, \\ r_{3, classser392}^{QR-G} : x_{QR-G,i} \in [26218.0, 49527.0] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{QP-G} : x_{QP-G,i} \in [0.0, 0.0] \xrightarrow{1.0} classer392, \\ r_{3, classser392}^{QP-G} : x_{QP-G,i} \in [188.0, 1080.0] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{QA-G} : x_{QA-G,i} \in [96451.0, 124120.0] \xrightarrow{1.0} classer392, \\ r_{2, classser392}^{QA-G} : x_{QA-G,i} \in [124120.0, 143151.0] \xrightarrow{0.4286} classer392, \\ r_{3, classser392}^{QA-G} : x_{QA-G,i} \in (143151.0, 367840.0] \xrightarrow{0.0} classer392, \\ r_{2, classser392}^{FE-E} : x_{FE-E,i} \in [0.0, 13.0] \xrightarrow{0.2381} classer392, \\ r_{3, classser392}^{FE-E} : x_{FE-E,i} \in (13.0, 89.8] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{PH-E} : x_{PH-E,i} \in [7.2, 7.3] \xrightarrow{0.0} classer392, \\ r_{2, classser392}^{PH-E} : x_{PH-E,i} \in [7.3, 7.6] \xrightarrow{0.0287} classer392, \\ r_{3, classser392}^{PH-E} : x_{PH-E,i} \in (7.6, 8.0] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{SS-E} : x_{SS-E,i} \in [62.0, 77.0] \xrightarrow{0.0} classer392, \\ r_{2, classser392}^{SS-E} : x_{SS-E,i} \in [77.0, 313.0] \xrightarrow{0.0172} classer392, \\ r_{3, classser392}^{SS-E} : x_{SS-E,i} \in (313.0, 655.0] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{SSV-E} : x_{SSV-E,i} \in [19.0, 30.0] \xrightarrow{1.0} classer392, \\ r_{2, classser392}^{SSV-E} : x_{SSV-E,i} \in [30.0, 51.0] \xrightarrow{0.6667} classer392, \\ r_{3, classser392}^{SSV-E} : x_{SSV-E,i} \in (51.0, 593.0] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{DQO-E} : x_{DQO-E,i} \in [27.0, 100.0] \xrightarrow{1.0} classer392, \\ r_{2, classser392}^{DQO-E} : x_{DQO-E,i} \in [100.0, 180.0] \xrightarrow{0.5} classer392, \\ r_{3, classser392}^{DQO-E} : x_{DQO-E,i} \in (180.0, 1579.0] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{DBO-E} : x_{DBO-E,i} \in [69.0, 73.0] \xrightarrow{0.0} classer392, \\ r_{2, classser392}^{DBO-E} : x_{DBO-E,i} \in [73.0, 73.0] \xrightarrow{1.0} classer392, \\ r_{3, classser392}^{DBO-E} : x_{DBO-E,i} \in (73.0, 987.0] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{PH-D} : x_{PH-D,i} \in [7.1, 7.3] \xrightarrow{0.0} classer392, \\ r_{2, classser392}^{PH-D} : x_{PH-D,i} \in [7.3, 7.7] \xrightarrow{0.017} classer392, \\ r_{3, classser392}^{PH-D} : x_{PH-D,i} \in (7.7, 7.9] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{SS-D} : x_{SS-D,i} \in [40.0, 48.0] \xrightarrow{1.0} classer392, \\ r_{2, classser392}^{SS-D} : x_{SS-D,i} \in [48.0, 98.0] \xrightarrow{0.0127} classer392, \\ r_{3, classser392}^{SS-D} : x_{SS-D,i} \in (98.0, 192.0] \xrightarrow{0.0} classer392, \\ r_{1, classser392}^{SSV-D} : x_{SSV-D,i} \in [13.0, 30.0] \xrightarrow{1.0} classer392, \\ r_{2, classser392}^{SSV-D} : x_{SSV-D,i} \in [30.0, 42.0] \xrightarrow{0.1111} classer392, \\ r_{3, classser392}^{SSV-D} : x_{SSV-D,i} \in (42.0, 134.0] \xrightarrow{0.0} classer392, \end{array}$$

$$\begin{aligned}
r_{1,\text{classer392}}^{DQO-D} : & x_{DQO-D,i} \in [27.0, 90.0] \xrightarrow{1.0} \text{classer392}, \\
r_{2,\text{classer392}}^{DQO-D} : & x_{DQO-D,i} \in [90.0, 161.0] \xrightarrow{0.1034} \text{classer392}, \\
r_{3,\text{classer392}}^{DQO-D} : & x_{DQO-D,i} \in (161.0, 538.0] \xrightarrow{0.0} \text{classer392}, \\
r_{1,\text{classer392}}^{DBO-D} : & x_{DBO-D,i} \in [36.0, 54.0] \xrightarrow{0.0} \text{classer392}, \\
r_{2,\text{classer392}}^{DBO-D} : & x_{DBO-D,i} \in [54.0, 54.0] \xrightarrow{1.0} \text{classer392}, \\
r_{3,\text{classer392}}^{DBO-D} : & x_{DBO-D,i} \in (54.0, 274.0] \xrightarrow{0.0} \text{classer392}, \\
r_{1,\text{classer392}}^{PH-S} : & x_{PH-S,i} \in [7.0, 7.2) \xrightarrow{0.0} \text{classer392}, \\
r_{2,\text{classer392}}^{PH-S} : & x_{PH-S,i} \in [7.2, 7.6] \xrightarrow{0.0237} \text{classer392}, \\
r_{3,\text{classer392}}^{PH-S} : & x_{PH-S,i} \in (7.6, 8.0] \xrightarrow{0.0} \text{classer392}, \\
r_{1,\text{classer392}}^{SS-S} : & x_{SS-S,i} \in [2.8, 5.2) \xrightarrow{0.0} \text{classer392}, \\
r_{2,\text{classer392}}^{SS-S} : & x_{SS-S,i} \in [5.2, 8.0] \xrightarrow{0.1042} \text{classer392}, \\
r_{3,\text{classer392}}^{SS-S} : & x_{SS-S,i} \in (8.0, 174.8] \xrightarrow{0.0} \text{classer392}, \\
r_{1,\text{classer392}}^{SSV-S} : & x_{SSV-S,i} \in [1.6, 2.8) \xrightarrow{0.0} \text{classer392}, \\
r_{2,\text{classer392}}^{SSV-S} : & x_{SSV-S,i} \in [2.8, 4.0] \xrightarrow{0.2941} \text{classer392}, \\
r_{3,\text{classer392}}^{SSV-S} : & x_{SSV-S,i} \in (4.0, 134.8] \xrightarrow{0.0} \text{classer392}, \\
r_{2,\text{classer392}}^{DQO-S} : & x_{DQO-S,i} \in [9.0, 66.0] \xrightarrow{0.0197} \text{classer392}, \\
r_{3,\text{classer392}}^{DQO-S} : & x_{DQO-S,i} \in (66.0, 163.0] \xrightarrow{0.0} \text{classer392}, \\
r_{1,\text{classer392}}^{DBO-S} : & x_{DBO-S,i} \in [2.0, 5.0) \xrightarrow{0.0} \text{classer392}, \\
r_{2,\text{classer392}}^{DBO-S} : & x_{DBO-S,i} \in [5.0, 5.0] \xrightarrow{0.2} \text{classer392}, \\
r_{3,\text{classer392}}^{DBO-S} : & x_{DBO-S,i} \in (5.0, 84.0] \xrightarrow{0.0} \text{classer392}, \\
r_{1,\text{classer392}}^{V30-B} : & x_{V30-B,i} \in [77.0, 310.0) \xrightarrow{0.0035} \text{classer392}, \\
r_{2,\text{classer392}}^{V30-B} : & x_{V30-B,i} \in [310.0, 383.0] \xrightarrow{0.1111} \text{classer392}, \\
r_{3,\text{classer392}}^{V30-B} : & x_{V30-B,i} \in (383.0, 770.0] \xrightarrow{0.0} \text{classer392}, \\
r_{1,\text{classer392}}^{MLSS-B} : & x_{MLSS-B,i} \in [754.0, 2589.0) \xrightarrow{0.0} \text{classer392}, \\
r_{2,\text{classer392}}^{MLSS-B} : & x_{MLSS-B,i} \in [2589.0, 2978.0] \xrightarrow{0.5} \text{classer392}, \\
r_{3,\text{classer392}}^{MLSS-B} : & x_{MLSS-B,i} \in (2978.0, 3294.0] \xrightarrow{1.0} \text{classer392}, \\
r_{1,\text{classer392}}^{MLVSS-B} : & x_{MLVSS-B,i} \in [185.0, 1807.0) \xrightarrow{0.0} \text{classer392}, \\
r_{2,\text{classer392}}^{MLVSS-B} : & x_{MLVSS-B,i} \in [1807.0, 2054.0] \xrightarrow{0.2632} \text{classer392}, \\
r_{3,\text{classer392}}^{MLVSS-B} : & x_{MLVSS-B,i} \in (2054.0, 2100.0] \xrightarrow{1.0} \text{classer392}, \\
r_{1,\text{classer392}}^{MCRT-B} : & x_{MCRT-B,i} \in [1.8, 34.4] \xrightarrow{0.0} \text{classer392}, \\
r_{3,\text{classer392}}^{MCRT-B} : & x_{MCRT-B,i} \in [179.8, 341.99] \xrightarrow{1.0} \text{classer392}, \\
r_{1,\text{classer393}}^{Q-E} : & x_{Q-E,i} \in [20500.0, 23662.9] \xrightarrow{0.0} \text{classer393}, \\
r_{3,\text{classer393}}^{Q-E} : & x_{Q-E,i} \in [29920.0, 54088.6] \xrightarrow{1.0} \text{classer393}, \\
r_{1,\text{classer393}}^{QB-B} : & x_{QB-B,i} \in [19883.0, 22891.0] \xrightarrow{0.0} \text{classer393}, \\
r_{3,\text{classer393}}^{QB-B} : & x_{QB-B,i} \in [29397.3, 52244.6] \xrightarrow{1.0} \text{classer393}, \\
r_{1,\text{classer393}}^{QR-G} : & x_{QR-G,i} \in [17932.6, 18343.5] \xrightarrow{0.0} \text{classer393}, \\
r_{3,\text{classer393}}^{QR-G} : & x_{QR-G,i} \in [26218.0, 49527.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,\text{classer393}}^{QP-G} : & x_{QP-G,i} \in [0.0, 0.0] \xrightarrow{0.0} \text{classer393}, \\
r_{3,\text{classer393}}^{QP-G} : & x_{QP-G,i} \in [188.0, 1080.0] \xrightarrow{1.0} \text{classer393},
\end{aligned}$$

$$\begin{aligned}
r_{1, \text{classer393}}^{QA-G} : x_{QA-G,i} \in [96451.0, 124120.0] &\xrightarrow{0.0} \text{classer393}, \\
r_{2, \text{classer393}}^{QA-G} : x_{QA-G,i} \in [124120.0, 143151.0] &\xrightarrow{0.5714} \text{classer393}, \\
r_{3, \text{classer393}}^{QA-G} : x_{QA-G,i} \in (143151.0, 367840.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{FE-E} : x_{FE-E,i} \in [0.0, 13.0] &\xrightarrow{0.7619} \text{classer393}, \\
r_{3, \text{classer393}}^{FE-E} : x_{FE-E,i} \in (13.0, 89.8] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{PH-E} : x_{PH-E,i} \in [7.2, 7.3] &\xrightarrow{1.0} \text{classer393}, \\
r_{2, \text{classer393}}^{PH-E} : x_{PH-E,i} \in [7.3, 7.6] &\xrightarrow{0.9713} \text{classer393}, \\
r_{3, \text{classer393}}^{PH-E} : x_{PH-E,i} \in (7.6, 8.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{SS-E} : x_{SS-E,i} \in [62.0, 77.0) &\xrightarrow{1.0} \text{classer393}, \\
r_{2, \text{classer393}}^{SS-E} : x_{SS-E,i} \in [77.0, 313.0] &\xrightarrow{0.9828} \text{classer393}, \\
r_{3, \text{classer393}}^{SS-E} : x_{SS-E,i} \in (313.0, 655.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{SSV-E} : x_{SSV-E,i} \in [19.0, 30.0] &\xrightarrow{0.0} \text{classer393}, \\
r_{2, \text{classer393}}^{SSV-E} : x_{SSV-E,i} \in [30.0, 51.0] &\xrightarrow{0.3333} \text{classer393}, \\
r_{3, \text{classer393}}^{SSV-E} : x_{SSV-E,i} \in (51.0, 593.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{DQO-E} : x_{DQO-E,i} \in [27.0, 100.0) &\xrightarrow{0.0} \text{classer393}, \\
r_{2, \text{classer393}}^{DQO-E} : x_{DQO-E,i} \in [100.0, 180.0] &\xrightarrow{0.5} \text{classer393}, \\
r_{3, \text{classer393}}^{DQO-E} : x_{DQO-E,i} \in (180.0, 1579.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{DBO-E} : x_{DBO-E,i} \in [69.0, 73.0) &\xrightarrow{1.0} \text{classer393}, \\
r_{2, \text{classer393}}^{DBO-E} : x_{DBO-E,i} \in [73.0, 73.0] &\xrightarrow{0.0} \text{classer393}, \\
r_{3, \text{classer393}}^{DBO-E} : x_{DBO-E,i} \in (73.0, 987.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{PH-D} : x_{PH-D,i} \in [7.1, 7.3] &\xrightarrow{1.0} \text{classer393}, \\
r_{2, \text{classer393}}^{PH-D} : x_{PH-D,i} \in [7.3, 7.7] &\xrightarrow{0.983} \text{classer393}, \\
r_{3, \text{classer393}}^{PH-D} : x_{PH-D,i} \in (7.7, 7.9] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{SS-D} : x_{SS-D,i} \in [40.0, 48.0) &\xrightarrow{0.0} \text{classer393}, \\
r_{2, \text{classer393}}^{SS-D} : x_{SS-D,i} \in [48.0, 98.0] &\xrightarrow{0.9873} \text{classer393}, \\
r_{3, \text{classer393}}^{SS-D} : x_{SS-D,i} \in (98.0, 192.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{SSV-D} : x_{SSV-D,i} \in [13.0, 30.0] &\xrightarrow{0.0} \text{classer393}, \\
r_{2, \text{classer393}}^{SSV-D} : x_{SSV-D,i} \in [30.0, 42.0] &\xrightarrow{0.8889} \text{classer393}, \\
r_{3, \text{classer393}}^{SSV-D} : x_{SSV-D,i} \in (42.0, 134.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{DQO-D} : x_{DQO-D,i} \in [27.0, 90.0) &\xrightarrow{0.0} \text{classer393}, \\
r_{2, \text{classer393}}^{DQO-D} : x_{DQO-D,i} \in [90.0, 161.0] &\xrightarrow{0.8966} \text{classer393}, \\
r_{3, \text{classer393}}^{DQO-D} : x_{DQO-D,i} \in (161.0, 538.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{DBO-D} : x_{DBO-D,i} \in [36.0, 54.0) &\xrightarrow{1.0} \text{classer393}, \\
r_{2, \text{classer393}}^{DBO-D} : x_{DBO-D,i} \in [54.0, 54.0] &\xrightarrow{0.0} \text{classer393}, \\
r_{3, \text{classer393}}^{DBO-D} : x_{DBO-D,i} \in (54.0, 274.0] &\xrightarrow{1.0} \text{classer393}, \\
r_{1, \text{classer393}}^{PH-S} : x_{PH-S,i} \in [7.0, 7.2) &\xrightarrow{1.0} \text{classer393}, \\
r_{2, \text{classer393}}^{PH-S} : x_{PH-S,i} \in [7.2, 7.6] &\xrightarrow{0.9763} \text{classer393}, \\
r_{3, \text{classer393}}^{PH-S} : x_{PH-S,i} \in (7.6, 8.0] &\xrightarrow{1.0} \text{classer393},
\end{aligned}$$

$$\begin{aligned}
r_{1,\text{classer393}}^{SS-S} &: x_{SS-S,i} \in [2.8, 5.2] \xrightarrow{1.0} \text{classer393}, \\
r_{2,\text{classer393}}^{SS-S} &: x_{SS-S,i} \in [5.2, 8.0] \xrightarrow{0.8958} \text{classer393}, \\
r_{3,\text{classer393}}^{SS-S} &: x_{SS-S,i} \in (8.0, 174.8] \xrightarrow{1.0} \text{classer393}, \\
r_{1,\text{classer393}}^{SSV-S} &: x_{SSV-S,i} \in [1.6, 2.8) \xrightarrow{1.0} \text{classer393}, \\
r_{2,\text{classer393}}^{SSV-S} &: x_{SSV-S,i} \in [2.8, 4.0] \xrightarrow{0.7059} \text{classer393}, \\
r_{3,\text{classer393}}^{SSV-S} &: x_{SSV-S,i} \in (4.0, 134.8] \xrightarrow{1.0} \text{classer393}, \\
r_{2,\text{classer393}}^{DQO-S} &: x_{DQO-S,i} \in [9.0, 66.0] \xrightarrow{0.9803} \text{classer393}, \\
r_{3,\text{classer393}}^{DQO-S} &: x_{DQO-S,i} \in (66.0, 163.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,\text{classer393}}^{DBO-S} &: x_{DBO-S,i} \in [2.0, 5.0) \xrightarrow{1.0} \text{classer393}, \\
r_{2,\text{classer393}}^{DBO-S} &: x_{DBO-S,i} \in [5.0, 5.0] \xrightarrow{0.8} \text{classer393}, \\
r_{3,\text{classer393}}^{DBO-S} &: x_{DBO-S,i} \in (5.0, 84.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,\text{classer393}}^{V30-B} &: x_{V30-B,i} \in [77.0, 310.0) \xrightarrow{0.9965} \text{classer393}, \\
r_{2,\text{classer393}}^{V30-B} &: x_{V30-B,i} \in [310.0, 383.0] \xrightarrow{0.8889} \text{classer393}, \\
r_{3,\text{classer393}}^{V30-B} &: x_{V30-B,i} \in (383.0, 770.0] \xrightarrow{1.0} \text{classer393}, \\
r_{1,\text{classer393}}^{MLSS-B} &: x_{MLSS-B,i} \in [754.0, 2589.0) \xrightarrow{1.0} \text{classer393}, \\
r_{2,\text{classer393}}^{MLSS-B} &: x_{MLSS-B,i} \in [2589.0, 2978.0] \xrightarrow{0.5} \text{classer393}, \\
r_{3,\text{classer393}}^{MLSS-B} &: x_{MLSS-B,i} \in (2978.0, 3294.0] \xrightarrow{0.0} \text{classer393}, \\
r_{1,\text{classer393}}^{MLVSS-B} &: x_{MLVSS-B,i} \in [185.0, 1807.0) \xrightarrow{1.0} \text{classer393}, \\
r_{2,\text{classer393}}^{MLVSS-B} &: x_{MLVSS-B,i} \in [1807.0, 2054.0] \xrightarrow{0.7368} \text{classer393}, \\
r_{3,\text{classer393}}^{MLVSS-B} &: x_{MLVSS-B,i} \in (2054.0, 2100.0] \xrightarrow{0.0} \text{classer393}, \\
r_{1,\text{classer393}}^{MCRT-B} &: x_{MCRT-B,i} \in [1.8, 34.4] \xrightarrow{1.0} \text{classer393}, \\
r_{3,\text{classer393}}^{MCRT-B} &: x_{MCRT-B,i} \in [179.8, 341.99] \xrightarrow{0.0} \text{classer393} \quad \}
\end{aligned}$$

E.3 BbD para $\mathcal{P}_3^* \subseteq \mathcal{P}_{Gi1,R1}^{EnW,G}$

Patrón: centro cerrado para la variable Q-E de la partición en \mathcal{P}_3^*

$$I_1^{Q-E,3} = [29920.0, 30592.2)$$

$$I_2^{Q-E,3} = [30592.2, 52255.8]$$

$$I_3^{Q-E,3} = (52255.8, 54088.6]$$

Patrón: centro cerrado para la variable QB-B de la partición en \mathcal{P}_3^*

$$I_1^{QB-B,3} = [29397.3, 29936.8)$$

$$I_2^{QB-B,3} = [29936.8, 49695.8]$$

$$I_3^{QB-B,3} = (49695.8, 52244.6]$$

Patrón: centro cerrado para la variable QR-G de la partición en \mathcal{P}_3^*

$$I_1^{QR-G,3} = [26218.0, 27351.0)$$

$$I_2^{QR-G,3} = [27351.0, 43298.1]$$

$$I_3^{QR-G,3} = (43298.1, 49527.0]$$

Patrón: centro cerrado para la variable QP-G de la partición en \mathcal{P}_3^*

$$I_1^{QP-G,3} = [188.0, 327.6)$$

$$I_2^{QP-G,3} = [327.6, 866.7]$$

$$I_3^{QP-G,3} = (866.7, 1080.0]$$

Patrón: centro cerrado para la variable QA-G de la partición en \mathcal{P}_3^*

$$I_1^{QA-G,3} = [124120.0, 136371.0)$$

$$I_2^{QA-G,3} = [136371.0, 324470.0]$$

$$I_3^{QA-G,3} = (324470.0, 367840.0]$$

Patrón: centro cerrado para la variable FE-E de la partición en \mathcal{P}_3^*

$$I_1^{FE-E,3} = [0.0, 0.0)$$

$$I_2^{FE-E,3} = [0.0, 65.6]$$

$$I_3^{FE-E,3} = (65.6, 89.8]$$

Patrón: centro cerrado para la variable PH-E de la partición en \mathcal{P}_3^*

$$I_1^{PH-E,3} = [7.2, 7.2)$$

$$I_2^{PH-E,3} = [7.2, 7.9]$$

$$I_3^{PH-E,3} = (7.9, 8.0]$$

Patrón: centro cerrado para la variable SS-E de la partición en \mathcal{P}_3^*

$$I_1^{SS-E,3} = [62.0, 82.0)$$

$$I_2^{SS-E,3} = [82.0, 266.0]$$

$$I_3^{SS-E,3} = (266.0, 655.0]$$

Patrón: centro cerrado para la variable SSV-E de la partición en \mathcal{P}_3^*

$$I_1^{SSV-E,3} = [30.0, 60.0)$$

$$I_2^{SSV-E,3} = [60.0, 193.0]$$

$$I_3^{SSV-E,3} = (193.0, 593.0]$$

Patrón: centro cerrado para la variable DQO-E de la partición en \mathcal{P}_3^*

$$I_1^{DQO-E,3} = [100.0, 158.0)$$

$$I_2^{DQO-E,3} = [158.0, 595.0]$$

$$I_3^{DQO-E,3} = (595.0, 1579.0]$$

Patrón: centro cerrado para la variable DBO-E de la partición en \mathcal{P}_3^*

$$I_1^{DBO-E,3} = [69.0, 90.0)$$

$$I_2^{DBO-E,3} = [90.0, 258.0]$$

$$I_3^{DBO-E,3} = (258.0, 987.0]$$

Patrón: centro cerrado para la variable PH-D de la partición en \mathcal{P}_3^*

$$I_1^{PH-D,3} = [7.1, 7.2)$$

$$I_2^{PH-D,3} = [7.2, 7.8]$$

$$I_3^{PH-D,3} = (7.8, 7.9]$$

Patrón: centro cerrado para la variable SS-D de la partición en \mathcal{P}_3^*

$$I_1^{SS-D,3} = [48.0, 63.0)$$

$$I_2^{SS-D,3} = [63.0, 136.0]$$

$$I_3^{SS-D,3} = (136.0, 192.0]$$

Patrón: centro cerrado para la variable SSV-D de la partición en \mathcal{P}_3^*

$$I_1^{SSV-D,3} = [30.0, 47.0)$$

$$I_2^{SSV-D,3} = [47.0, 99.0]$$

$$I_3^{SSV-D,3} = (99.0, 134.0]$$

Patrón: centro cerrado para la variable DQO-D de la partición en \mathcal{P}_3^*

$$I_1^{DQO-D,3} = [90.0, 100.0)$$

$$I_2^{DQO-D,3} = [100.0, 269.0]$$

$$I_3^{DQO-D,3} = (269.0, 538.0]$$

Patrón: centro cerrado para la variable DBO-D de la partición en \mathcal{P}_3^*

$$I_1^{DBO-D,3} = [36.0, 56.0)$$

$$I_2^{DBO-D,3} = [56.0, 171.0]$$

$$I_3^{DBO-D,3} = (171.0, 274.0]$$

Patrón: centro cerrado para la variable PH-S de la partición en \mathcal{P}_3^*

$$I_1^{PH-S,3} = [7.0, 7.0)$$

$$I_2^{PH-S,3} = [7.0, 8.0]$$

$$I_3^{PH-S,3} = (8.0, 8.0]$$

Patrón: centro cerrado para la variable SS-S de la partición en \mathcal{P}_3^*

$$I_1^{SS-S,3} = [2.8, 3.2)$$

$$I_2^{SS-S,3} = [3.2, 20.0]$$

$$I_3^{SS-S,3} = (20.0, 174.8]$$

Patrón: centro cerrado para la variable SSV-S de la partición en \mathcal{P}_3^*

$$I_1^{SSV-S,3} = [1.6, 1.6)$$

$$\begin{aligned} I_2^{SSV-S,3} &= [1.6, 18.0] \\ I_3^{SSV-S,3} &= (18.0, 134.8] \end{aligned}$$

Patrón: centro cerrado para la variable DQO-S de la partición en \mathcal{P}_3^*

$$\begin{aligned} I_1^{DQO-S,3} &= [9.0, 9.0] \\ I_2^{DQO-S,3} &= [9.0, 94.0] \\ I_3^{DQO-S,3} &= (94.0, 163.0] \end{aligned}$$

Patrón: centro cerrado para la variable DBO-S de la partición en \mathcal{P}_3^*

$$\begin{aligned} I_1^{DBO-S,3} &= [2.0, 4.0] \\ I_2^{DBO-S,3} &= [4.0, 26.0] \\ I_3^{DBO-S,3} &= (26.0, 84.0] \end{aligned}$$

Patrón: centro cerrado para la variable V30-B de la partición en \mathcal{P}_3^*

$$\begin{aligned} I_1^{V30-B,3} &= [77.0, 115.0] \\ I_2^{V30-B,3} &= [115.0, 380.0] \\ I_3^{V30-B,3} &= (380.0, 770.0] \end{aligned}$$

Patrón: centro cerrado para la variable MLSS-B de la partición en \mathcal{P}_3^*

$$\begin{aligned} I_1^{MLSS-B,3} &= [754.0, 846.0] \\ I_2^{MLSS-B,3} &= [846.0, 2696.0] \\ I_3^{MLSS-B,3} &= (2696.0, 2978.0] \end{aligned}$$

Patrón: centro cerrado para la variable MLVSS-B de la partición en \mathcal{P}_3^*

$$\begin{aligned} I_1^{MLVSS-B,3} &= [185.0, 684.0] \\ I_2^{MLVSS-B,3} &= [684.0, 1921.0] \\ I_3^{MLVSS-B,3} &= (1921.0, 2054.0] \end{aligned}$$

Patrón: centro cerrado para la variable MCRT-B de la partición en \mathcal{P}_3^*

$$\begin{aligned} I_1^{MCRT-B,3} &= [1.8, 6.9] \\ I_2^{MCRT-B,3} &= [6.9, 28.8] \\ I_3^{MCRT-B,3} &= (28.8, 34.4] \end{aligned}$$

E.4 $\mathcal{R}(\mathcal{P}_3^*)$

$$\mathcal{R}(\mathcal{P}_3^*) = \{ \begin{array}{l} r_{1,\text{classer389}}^{Q-E} : x_{Q-E,i} \in [29920.0, 30592.2] \xrightarrow{0.0} \text{classer389}, \\ r_{2,\text{classer389}}^{Q-E} : x_{Q-E,i} \in [30592.2, 52255.8] \xrightarrow{0.174} \text{classer389}, \\ r_{3,\text{classer389}}^{Q-E} : x_{Q-E,i} \in (52255.8, 54088.6] \xrightarrow{1.0} \text{classer389}, \\ r_{1,\text{classer389}}^{QB-B} : x_{QB-B,i} \in [29397.3, 29936.8] \xrightarrow{0.0} \text{classer389}, \\ r_{2,\text{classer389}}^{QB-B} : x_{QB-B,i} \in [29936.8, 49695.8] \xrightarrow{0.174} \text{classer389}, \\ r_{3,\text{classer389}}^{QB-B} : x_{QB-B,i} \in (49695.8, 52244.6] \xrightarrow{1.0} \text{classer389}, \\ r_{1,\text{classer389}}^{QR-G} : x_{QR-G,i} \in [26218.0, 27351.0) \xrightarrow{1.0} \text{classer389}, \\ r_{2,\text{classer389}}^{QR-G} : x_{QR-G,i} \in [27351.0, 43298.1] \xrightarrow{0.2105} \text{classer389}, \\ r_{3,\text{classer389}}^{QR-G} : x_{QR-G,i} \in (43298.1, 49527.0] \xrightarrow{0.0} \text{classer389}, \\ r_{1,\text{classer389}}^{QP-G} : x_{QP-G,i} \in [188.0, 327.6) \xrightarrow{1.0} \text{classer389}, \\ r_{2,\text{classer389}}^{QP-G} : x_{QP-G,i} \in [327.6, 866.7] \xrightarrow{0.1937} \text{classer389}, \\ r_{3,\text{classer389}}^{QP-G} : x_{QP-G,i} \in (866.7, 1080.0] \xrightarrow{0.0} \text{classer389}, \\ r_{1,\text{classer389}}^{QA-G} : x_{QA-G,i} \in [124120.0, 136371.0) \xrightarrow{1.0} \text{classer389}, \\ r_{2,\text{classer389}}^{QA-G} : x_{QA-G,i} \in [136371.0, 324470.0] \xrightarrow{0.1799} \text{classer389}, \\ r_{3,\text{classer389}}^{QA-G} : x_{QA-G,i} \in (324470.0, 367840.0] \xrightarrow{0.0} \text{classer389}, \\ r_{2,\text{classer389}}^{FE-E} : x_{FE-E,i} \in [0.0, 65.6] \xrightarrow{0.1823} \text{classer389}, \\ r_{3,\text{classer389}}^{FE-E} : x_{FE-E,i} \in (65.6, 89.8] \xrightarrow{0.0} \text{classer389}, \\ r_{2,\text{classer389}}^{PH-E} : x_{PH-E,i} \in [7.2, 7.9] \xrightarrow{0.2006} \text{classer389}, \\ r_{3,\text{classer389}}^{PH-E} : x_{PH-E,i} \in (7.9, 8.0] \xrightarrow{1.0} \text{classer389}, \\ r_{1,\text{classer389}}^{SS-E} : x_{SS-E,i} \in [62.0, 82.0) \xrightarrow{1.0} \text{classer389}, \\ r_{2,\text{classer389}}^{SS-E} : x_{SS-E,i} \in [82.0, 266.0] \xrightarrow{0.25} \text{classer389}, \\ r_{3,\text{classer389}}^{SS-E} : x_{SS-E,i} \in (266.0, 655.0] \xrightarrow{0.0} \text{classer389}, \\ r_{1,\text{classer389}}^{SSV-E} : x_{SSV-E,i} \in [30.0, 60.0) \xrightarrow{1.0} \text{classer389}, \\ r_{2,\text{classer389}}^{SSV-E} : x_{SSV-E,i} \in [60.0, 193.0] \xrightarrow{0.257} \text{classer389}, \\ r_{3,\text{classer389}}^{SSV-E} : x_{SSV-E,i} \in (193.0, 593.0] \xrightarrow{0.0} \text{classer389}, \\ r_{1,\text{classer389}}^{DQO-E} : x_{DQO-E,i} \in [100.0, 158.0) \xrightarrow{1.0} \text{classer389}, \\ r_{2,\text{classer389}}^{DQO-E} : x_{DQO-E,i} \in [158.0, 595.0] \xrightarrow{0.2067} \text{classer389}, \\ r_{3,\text{classer389}}^{DQO-E} : x_{DQO-E,i} \in (595.0, 1579.0] \xrightarrow{0.0} \text{classer389}, \\ r_{1,\text{classer389}}^{DBO-E} : x_{DBO-E,i} \in [69.0, 90.0) \xrightarrow{1.0} \text{classer389}, \\ r_{2,\text{classer389}}^{DBO-E} : x_{DBO-E,i} \in [90.0, 258.0] \xrightarrow{0.2489} \text{classer389}, \\ r_{3,\text{classer389}}^{DBO-E} : x_{DBO-E,i} \in (258.0, 987.0] \xrightarrow{0.0} \text{classer389}, \\ r_{1,\text{classer389}}^{PH-D} : x_{PH-D,i} \in [7.1, 7.2] \xrightarrow{0.5} \text{classer389}, \\ r_{2,\text{classer389}}^{PH-D} : x_{PH-D,i} \in [7.2, 7.8] \xrightarrow{0.1044} \text{classer389}, \\ r_{3,\text{classer389}}^{PH-D} : x_{PH-D,i} \in (7.8, 7.9] \xrightarrow{0.0} \text{classer389}, \\ r_{1,\text{classer389}}^{SS-D} : x_{SS-D,i} \in [48.0, 63.0) \xrightarrow{1.0} \text{classer389}, \\ r_{2,\text{classer389}}^{SS-D} : x_{SS-D,i} \in [63.0, 136.0] \xrightarrow{0.1453} \text{classer389}, \\ r_{3,\text{classer389}}^{SS-D} : x_{SS-D,i} \in (136.0, 192.0] \xrightarrow{0.0} \text{classer389}, \end{array} \}$$

$$\begin{aligned}
r_{1, \text{classer389}}^{SSV-D} : x_{SSV-D,i} \in [30.0, 47.0] &\xrightarrow{1.0} \text{classer389}, \\
r_{2, \text{classer389}}^{SSV-D} : x_{SSV-D,i} \in [47.0, 99.0] &\xrightarrow{0.1424} \text{classer389}, \\
r_{3, \text{classer389}}^{SSV-D} : x_{SSV-D,i} \in (99.0, 134.0] &\xrightarrow{0.0} \text{classer389}, \\
r_{1, \text{classer389}}^{DQO-D} : x_{DQO-D,i} \in [90.0, 100.0] &\xrightarrow{0.0} \text{classer389}, \\
r_{2, \text{classer389}}^{DQO-D} : x_{DQO-D,i} \in [100.0, 269.0] &\xrightarrow{0.2949} \text{classer389}, \\
r_{3, \text{classer389}}^{DQO-D} : x_{DQO-D,i} \in (269.0, 538.0] &\xrightarrow{0.0} \text{classer389}, \\
r_{1, \text{classer389}}^{DBO-D} : x_{DBO-D,i} \in [36.0, 56.0) &\xrightarrow{1.0} \text{classer389}, \\
r_{2, \text{classer389}}^{DBO-D} : x_{DBO-D,i} \in [56.0, 171.0] &\xrightarrow{0.2082} \text{classer389}, \\
r_{3, \text{classer389}}^{DBO-D} : x_{DBO-D,i} \in (171.0, 274.0] &\xrightarrow{0.0} \text{classer389}, \\
r_{1, \text{classer389}}^{PH-S} : x_{PH-S} \in [7.0, 8.0] &\xrightarrow{0.2063} \text{classer389}, \\
r_{1, \text{classer389}}^{SS-S} : x_{SS-S,i} \in [2.8, 3.2) &\xrightarrow{0.0} \text{classer389}, \\
r_{2, \text{classer389}}^{SS-S} : x_{SS-S,i} \in [3.2, 20.0] &\xrightarrow{0.2588} \text{classer389}, \\
r_{3, \text{classer389}}^{SS-S} : x_{SS-S,i} \in (20.0, 174.8] &\xrightarrow{0.0} \text{classer389}, \\
r_{2, \text{classer389}}^{SSV-S} : x_{SSV-S,i} \in [1.6, 18.0] &\xrightarrow{0.2418} \text{classer389}, \\
r_{3, \text{classer389}}^{SSV-S} : x_{SSV-S,i} \in (18.0, 134.8] &\xrightarrow{0.0} \text{classer389}, \\
r_{2, \text{classer389}}^{DQO-S} : x_{DQO-S,i} \in [9.0, 94.0] &\xrightarrow{0.1977} \text{classer389}, \\
r_{3, \text{classer389}}^{DQO-S} : x_{DQO-S,i} \in (94.0, 163.0] &\xrightarrow{0.0} \text{classer389}, \\
r_{1, \text{classer389}}^{DBO-S} : x_{DBO-S,i} \in [2.0, 4.0) &\xrightarrow{1.0} \text{classer389}, \\
r_{2, \text{classer389}}^{DBO-S} : x_{DBO-S,i} \in [4.0, 26.0] &\xrightarrow{0.2469} \text{classer389}, \\
r_{3, \text{classer389}}^{DBO-S} : x_{DBO-S,i} \in (26.0, 84.0] &\xrightarrow{0.0} \text{classer389}, \\
r_{1, \text{classer389}}^{V30-B} : x_{V30-B,i} \in [77.0, 115.0) &\xrightarrow{0.0} \text{classer389}, \\
r_{2, \text{classer389}}^{V30-B} : x_{V30-B,i} \in [115.0, 380.0] &\xrightarrow{0.2194} \text{classer389}, \\
r_{3, \text{classer389}}^{V30-B} : x_{V30-B,i} \in (380.0, 770.0] &\xrightarrow{0.0} \text{classer389}, \\
r_{1, \text{classer389}}^{MLSS-B} : x_{MLSS-B,i} \in [754.0, 846.0) &\xrightarrow{0.0} \text{classer389}, \\
r_{2, \text{classer389}}^{MLSS-B} : x_{MLSS-B,i} \in [846.0, 2696.0] &\xrightarrow{0.1789} \text{classer389}, \\
r_{3, \text{classer389}}^{MLSS-B} : x_{MLSS-B,i} \in (2696.0, 2978.0] &\xrightarrow{1.0} \text{classer389}, \\
r_{1, \text{classer389}}^{MLVSS-B} : x_{MLVSS-B,i} \in [185.0, 684.0) &\xrightarrow{0.0} \text{classer389}, \\
r_{2, \text{classer389}}^{MLVSS-B} : x_{MLVSS-B,i} \in [684.0, 1921.0] &\xrightarrow{0.1845} \text{classer389}, \\
r_{3, \text{classer389}}^{MLVSS-B} : x_{MLVSS-B,i} \in (1921.0, 2054.0] &\xrightarrow{0.0} \text{classer389}, \\
r_{1, \text{classer389}}^{MCRT-B} : x_{MCRT-B,i} \in [1.8, 6.9) &\xrightarrow{0.0} \text{classer389}, \\
r_{2, \text{classer389}}^{MCRT-B} : x_{MCRT-B,i} \in [6.9, 28.8] &\xrightarrow{0.1926} \text{classer389}, \\
r_{3, \text{classer389}}^{MCRT-B} : x_{MCRT-B,i} \in (28.8, 34.4] &\xrightarrow{1.0} \text{classer389}, \\
r_{1, \text{classer391}}^{Q-E} : x_{Q-E,i} \in [29920.0, 30592.2) &\xrightarrow{1.0} \text{classer391}, \\
r_{2, \text{classer391}}^{Q-E} : x_{Q-E,i} \in [30592.2, 52255.8] &\xrightarrow{0.826} \text{classer391}, \\
r_{3, \text{classer391}}^{Q-E} : x_{Q-E,i} \in (52255.8, 54088.6] &\xrightarrow{0.0} \text{classer391}, \\
r_{1, \text{classer391}}^{QB-B} : x_{QB-B,i} \in [29397.3, 29936.8] &\xrightarrow{1.0} \text{classer391}, \\
r_{2, \text{classer391}}^{QB-B} : x_{QB-B,i} \in [29936.8, 49695.8] &\xrightarrow{0.826} \text{classer391}, \\
r_{3, \text{classer391}}^{QB-B} : x_{QB-B,i} \in (49695.8, 52244.6] &\xrightarrow{0.0} \text{classer391},
\end{aligned}$$

$$\begin{aligned}
r_{1,\text{classer391}}^{QR-G} : x_{QR-G,i} \in [26218.0, 27351.0] &\xrightarrow{0.0} \text{classer391}, \\
r_{2,\text{classer391}}^{QR-G} : x_{QR-G,i} \in [27351.0, 43298.1] &\xrightarrow{0.7895} \text{classer391}, \\
r_{3,\text{classer391}}^{QR-G} : x_{QR-G,i} \in (43298.1, 49527.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{QP-G} : x_{QP-G,i} \in [188.0, 327.6) &\xrightarrow{0.0} \text{classer391}, \\
r_{2,\text{classer391}}^{QP-G} : x_{QP-G,i} \in [327.6, 866.7] &\xrightarrow{0.8063} \text{classer391}, \\
r_{3,\text{classer391}}^{QP-G} : x_{QP-G,i} \in (866.7, 1080.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{QA-G} : x_{QA-G,i} \in [124120.0, 136371.0) &\xrightarrow{0.0} \text{classer391}, \\
r_{2,\text{classer391}}^{QA-G} : x_{QA-G,i} \in [136371.0, 324470.0] &\xrightarrow{0.8201} \text{classer391}, \\
r_{3,\text{classer391}}^{QA-G} : x_{QA-G,i} \in (324470.0, 367840.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{2,\text{classer391}}^{FE-E} : x_{FE-E,i} \in [0.0, 65.6] &\xrightarrow{0.8177} \text{classer391}, \\
r_{3,\text{classer391}}^{FE-E} : x_{FE-E,i} \in (65.6, 89.8] &\xrightarrow{1.0} \text{classer391}, \\
r_{2,\text{classer391}}^{PH-E} : x_{PH-E,i} \in [7.2, 7.9] &\xrightarrow{0.7994} \text{classer391}, \\
r_{3,\text{classer391}}^{PH-E} : x_{PH-E,i} \in (7.9, 8.0] &\xrightarrow{0.0} \text{classer391}, \\
r_{1,\text{classer391}}^{SS-E} : x_{SS-E,i} \in [62.0, 82.0) &\xrightarrow{0.0} \text{classer391}, \\
r_{2,\text{classer391}}^{SS-E} : x_{SS-E,i} \in [82.0, 266.0] &\xrightarrow{0.75} \text{classer391}, \\
r_{3,\text{classer391}}^{SS-E} : x_{SS-E,i} \in (266.0, 655.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{SSV-E} : x_{SSV-E,i} \in [30.0, 60.0) &\xrightarrow{0.0} \text{classer391}, \\
r_{2,\text{classer391}}^{SSV-E} : x_{SSV-E,i} \in [60.0, 193.0] &\xrightarrow{0.743} \text{classer391}, \\
r_{3,\text{classer391}}^{SSV-E} : x_{SSV-E,i} \in (193.0, 593.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{DQO-E} : x_{DQO-E,i} \in [100.0, 158.0) &\xrightarrow{0.0} \text{classer391}, \\
r_{2,\text{classer391}}^{DQO-E} : x_{DQO-E,i} \in [158.0, 595.0] &\xrightarrow{0.7933} \text{classer391}, \\
r_{3,\text{classer391}}^{DQO-E} : x_{DQO-E,i} \in (595.0, 1579.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{DBO-E} : x_{DBO-E,i} \in [69.0, 90.0) &\xrightarrow{0.0} \text{classer391}, \\
r_{2,\text{classer391}}^{DBO-E} : x_{DBO-E,i} \in [90.0, 258.0] &\xrightarrow{0.7511} \text{classer391}, \\
r_{3,\text{classer391}}^{DBO-E} : x_{DBO-E,i} \in (258.0, 987.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{PH-D} : x_{PH-D,i} \in [7.1, 7.2) &\xrightarrow{0.5} \text{classer391}, \\
r_{2,\text{classer391}}^{PH-D} : x_{PH-D,i} \in [7.2, 7.8] &\xrightarrow{0.8956} \text{classer391}, \\
r_{3,\text{classer391}}^{PH-D} : x_{PH-D,i} \in (7.8, 7.9] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{SS-D} : x_{SS-D,i} \in [48.0, 63.0) &\xrightarrow{0.0} \text{classer391}, \\
r_{2,\text{classer391}}^{SS-D} : x_{SS-D,i} \in [63.0, 136.0] &\xrightarrow{0.8547} \text{classer391}, \\
r_{3,\text{classer391}}^{SS-D} : x_{SS-D,i} \in (136.0, 192.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{SSV-D} : x_{SSV-D,i} \in [30.0, 47.0) &\xrightarrow{0.0} \text{classer391}, \\
r_{2,\text{classer391}}^{SSV-D} : x_{SSV-D,i} \in [47.0, 99.0] &\xrightarrow{0.8576} \text{classer391}, \\
r_{3,\text{classer391}}^{SSV-D} : x_{SSV-D,i} \in (99.0, 134.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{DQO-D} : x_{DQO-D,i} \in [90.0, 100.0) &\xrightarrow{1.0} \text{classer391}, \\
r_{2,\text{classer391}}^{DQO-D} : x_{DQO-D,i} \in [100.0, 269.0] &\xrightarrow{0.7051} \text{classer391}, \\
r_{3,\text{classer391}}^{DQO-D} : x_{DQO-D,i} \in (269.0, 538.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{DBO-D} : x_{DBO-D,i} \in [36.0, 56.0) &\xrightarrow{0.0} \text{classer391}, \\
r_{2,\text{classer391}}^{DBO-D} : x_{DBO-D,i} \in [56.0, 171.0] &\xrightarrow{0.7918} \text{classer391},
\end{aligned}$$

$$\begin{aligned}
r_{3,\text{classer391}}^{DBO-D} : x_{DBO-D,i} \in (171.0, 274.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer389}}^{PH-S} : x_{PH-S} \in [7.0, 8.0] &\xrightarrow{0.7937} \text{classer391}, \\
r_{1,\text{classer391}}^{SS-S} : x_{SS-S,i} \in [2.8, 3.2) &\xrightarrow{1.0} \text{classer391}, \\
r_{2,\text{classer391}}^{SS-S} : x_{SS-S,i} \in [3.2, 20.0] &\xrightarrow{0.7412} \text{classer391}, \\
r_{3,\text{classer391}}^{SS-S} : x_{SS-S,i} \in (20.0, 174.8] &\xrightarrow{1.0} \text{classer391}, \\
r_{2,\text{classer391}}^{SSV-S} : x_{SSV-S,i} \in [1.6, 18.0] &\xrightarrow{0.7582} \text{classer391}, \\
r_{3,\text{classer391}}^{SSV-S} : x_{SSV-S,i} \in (18.0, 134.8] &\xrightarrow{1.0} \text{classer391}, \\
r_{2,\text{classer391}}^{DQO-S} : x_{DQO-S,i} \in [9.0, 94.0] &\xrightarrow{0.8023} \text{classer391}, \\
r_{3,\text{classer391}}^{DQO-S} : x_{DQO-S,i} \in (94.0, 163.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{DBO-S} : x_{DBO-S,i} \in [2.0, 4.0) &\xrightarrow{0.0} \text{classer391}, \\
r_{2,\text{classer391}}^{DBO-S} : x_{DBO-S,i} \in [4.0, 26.0] &\xrightarrow{0.7531} \text{classer391}, \\
r_{3,\text{classer391}}^{DBO-S} : x_{DBO-S,i} \in (26.0, 84.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{V30-B} : x_{V30-B,i} \in [77.0, 115.0) &\xrightarrow{1.0} \text{classer391}, \\
r_{2,\text{classer391}}^{V30-B} : x_{V30-B,i} \in [115.0, 380.0] &\xrightarrow{0.7806} \text{classer391}, \\
r_{3,\text{classer391}}^{V30-B} : x_{V30-B,i} \in (380.0, 770.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{MLSS-B} : x_{MLSS-B,i} \in [754.0, 846.0) &\xrightarrow{1.0} \text{classer391}, \\
r_{2,\text{classer391}}^{MLSS-B} : x_{MLSS-B,i} \in [846.0, 2696.0] &\xrightarrow{0.8211} \text{classer391}, \\
r_{3,\text{classer391}}^{MLSS-B} : x_{MLSS-B,i} \in (2696.0, 2978.0] &\xrightarrow{0.0} \text{classer391}, \\
r_{1,\text{classer391}}^{MLVSS-B} : x_{MLVSS-B,i} \in [185.0, 684.0) &\xrightarrow{1.0} \text{classer391}, \\
r_{2,\text{classer391}}^{MLVSS-B} : x_{MLVSS-B,i} \in [684.0, 1921.0] &\xrightarrow{0.8155} \text{classer391}, \\
r_{3,\text{classer391}}^{MLVSS-B} : x_{MLVSS-B,i} \in (1921.0, 2054.0] &\xrightarrow{1.0} \text{classer391}, \\
r_{1,\text{classer391}}^{MCRT-B} : x_{MCRT-B,i} \in [1.8, 6.9) &\xrightarrow{1.0} \text{classer391}, \\
r_{2,\text{classer391}}^{MCRT-B} : x_{MCRT-B,i} \in [6.9, 28.8] &\xrightarrow{0.8074} \text{classer391}, \\
r_{3,\text{classer391}}^{MCRT-B} : x_{MCRT-B,i} \in (28.8, 34.4] &\xrightarrow{0.0} \text{classer391} \quad \}
\end{aligned}$$

E.5 BbD para $\mathcal{P}_4^* \subseteq \mathcal{P}^{EnW,G}_{Gi1,R1}$

Patrón: centro cerrado para la variable Q-E de la partición en \mathcal{P}_4^*

$$I_1^{Q-E,4} = [29920.0, 34284.4)$$

$$I_2^{Q-E,4} = [34284.4, 50500.5]$$

$$I_3^{Q-E,4} = (50500.5, 52255.8]$$

Patrón: centro cerrado para la variable QB-B de la partición en \mathcal{P}_4^*

$$I_1^{QB-B,4} = [29397.3, 33549.4)$$

$$I_2^{QB-B,4} = [33549.4, 39000.0]$$

$$I_3^{QB-B,4} = (39000.0, 49695.8]$$

Patrón: centro cerrado para la variable QR-G de la partición en \mathcal{P}_4^*

$$I_1^{QR-G,4} = [27351.0, 28343.8)$$

$$I_2^{QR-G,4} = [28343.8, 44568.6]$$

$$I_3^{QR-G,4} = (44568.6, 49527.0]$$

Patrón: centro cerrado para la variable QP-G de la partición en \mathcal{P}_4^*

$$I_1^{QP-G,4} = [327.6, 385.9)$$

$$I_2^{QP-G,4} = [385.9, 831.1]$$

$$I_3^{QP-G,4} = (831.1, 1080.0]$$

Patrón: centro cerrado para la variable QA-G de la partición en \mathcal{P}_4^*

$$I_1^{QA-G,4} = [136371.0, 156320.0)$$

$$I_2^{QA-G,4} = [156320.0, 331990.0]$$

$$I_3^{QA-G,4} = (331990.0, 367840.0]$$

Patrón: centro cerrado para la variable FE-E de la partición en \mathcal{P}_4^*

$$I_1^{FE-E,4} = [0.0, 0.0)$$

$$I_2^{FE-E,4} = [0.0, 63.3]$$

$$I_3^{FE-E,4} = (63.3, 89.8]$$

Patrón: centro cerrado para la variable PH-E de la partición en \mathcal{P}_4^*

$$I_1^{PH-E,4} = [7.2, 7.3)$$

$$I_2^{PH-E,4} = [7.3, 7.8]$$

$$I_3^{PH-E,4} = (7.8, 7.9]$$

Patrón: centro cerrado para la variable SS-E de la partición en \mathcal{P}_4^*

$$I_1^{SS-E,4} = [82.0, 114.0)$$

$$I_2^{SS-E,4} = [114.0, 480.0]$$

$$I_3^{SS-E,4} = (480.0, 655.0]$$

Patrón: centro cerrado para la variable SSV-E de la partición en \mathcal{P}_4^*

$$I_1^{SSV-E,4} = [60.0, 92.0)$$

$$I_2^{SSV-E,4} = [92.0, 336.0]$$

$$I_3^{SSV-E,4} = (336.0, 593.0]$$

Patrón: centro cerrado para la variable DQO-E de la partición en \mathcal{P}_4^*

$$I_1^{DQO-E,4} = [158.0, 414.0)$$

$$I_2^{DQO-E,4} = [414.0, 1279.0]$$

$$I_3^{DQO-E,4} = (1279.0, 1579.0]$$

Patrón: centro cerrado para la variable DBO-E de la partición en \mathcal{P}_4^*

$$I_1^{DBO-E,4} = [90.0, 220.0)$$

$$I_2^{DBO-E,4} = [220.0, 382.0]$$

$$I_3^{DBO-E,4} = (382.0, 987.0]$$

Patrón: centro cerrado para la variable PH-D de la partición en \mathcal{P}_4^*

$$I_1^{PH-D,4} = [7.2, 7.3)$$

$$I_2^{PH-D,4} = [7.3, 7.8]$$

$$I_3^{PH-D,4} = (7.8, 7.9]$$

Patrón: centro cerrado para la variable SS-D de la partición en \mathcal{P}_4^*

$$I_1^{SS-D,4} = [63.0, 68.0)$$

$$I_2^{SS-D,4} = [68.0, 112.0]$$

$$I_3^{SS-D,4} = (112.0, 192.0]$$

Patrón: centro cerrado para la variable SSV-D de la partición en \mathcal{P}_4^*

$$I_1^{SSV-D,4} = [47.0, 49.0)$$

$$I_2^{SSV-D,4} = [49.0, 92.0]$$

$$I_3^{SSV-D,4} = (92.0, 134.0]$$

Patrón: centro cerrado para la variable DQO-D de la partición en \mathcal{P}_4^*

$$I_1^{DQO-D,4} = [90.0, 186.0)$$

$$I_2^{DQO-D,4} = [186.0, 329.0]$$

$$I_3^{DQO-D,4} = (329.0, 538.0]$$

Patrón: centro cerrado para la variable DBO-D de la partición en \mathcal{P}_4^*

$$I_1^{DBO-D,4} = [56.0, 67.0)$$

$$I_2^{DBO-D,4} = [67.0, 224.0]$$

$$I_3^{DBO-D,4} = (224.0, 274.0]$$

Patrón: centro cerrado para la variable PH-S de la partición en \mathcal{P}_4^*

$$I_1^{PH-S,4} = [7.0, 7.2)$$

$$I_2^{PH-S,4} = [7.2, 7.8]$$

$$I_3^{PH-S,4} = (7.8, 8.0]$$

Patrón: centro cerrado para la variable SS-S de la partición en \mathcal{P}_4^*

$$I_1^{SS-S,4} = [2.8, 4.8)$$

$$I_2^{SS-S,4} = [4.8, 29.0]$$

$$I_3^{SS-S,4} = (29.0, 174.8]$$

Patrón: centro cerrado para la variable SSV-S de la partición en \mathcal{P}_4^*

$$I_1^{SSV-S,4} = [1.6, 4.4)$$

$$I_2^{SSV-S,4} = [4.4, 19.0]$$

$$I_3^{SSV-S,4} = (19.0, 134.8]$$

Patrón: centro cerrado para la variable DQO-S de la partición en \mathcal{P}_4^*

$$I_1^{DQO-S,4} = [9.0, 20.0)$$

$$I_2^{DQO-S,4} = [20.0, 95.0]$$

$$I_3^{DQO-S,4} = (95.0, 163.0]$$

Patrón: centro cerrado para la variable DBO-S de la partición en \mathcal{P}_4^*

$$I_1^{DBO-S,4} = [4.0, 6.0)$$

$$I_2^{DBO-S,4} = [6.0, 35.0]$$

$$I_3^{DBO-S,4} = (35.0, 84.0]$$

Patrón: centro cerrado para la variable V30-B de la partición en \mathcal{P}_4^*

$$I_1^{V30-B,4} = [77.0, 140.0)$$

$$I_2^{V30-B,4} = [140.0, 760.0]$$

$$I_3^{V30-B,4} = (760.0, 770.0]$$

Patrón: centro cerrado para la variable MLSS-B de la partición en \mathcal{P}_4^*

$$I_1^{MLSS-B,4} = [754.0, 1046.0)$$

$$I_2^{MLSS-B,4} = [1046.0, 2248.0]$$

$$I_3^{MLSS-B,4} = (2248.0, 2696.0]$$

Patrón: centro cerrado para la variable MLVSS-B de la partición en \mathcal{P}_4^*

$$I_1^{MLVSS-B,4} = [185.0, 611.0)$$

$$I_2^{MLVSS-B,4} = [611.0, 1726.0]$$

$$I_3^{MLVSS-B,4} = (1726.0, 2054.0]$$

Patrón: centro cerrado para la variable MCRT-B de la partición en \mathcal{P}_4^*

$$I_1^{MCRT-B,4} = [1.8, 6.2)$$

$$I_2^{MCRT-B,4} = [6.2, 16.0]$$

$$I_3^{MCRT-B,4} = (16.0, 28.8]$$

E.6 $\mathcal{R}(\mathcal{P}_4^*)$

$$\mathcal{R}(\mathcal{P}_4^*) = \{ \begin{array}{l} r_{1, classer383}^{Q-E} : x_{Q-E,i} \in [29920.0, 34284.4] \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{Q-E} : x_{Q-E,i} \in [34284.4, 50500.5] \xrightarrow{0.11} classer383, \\ r_{3, classer383}^{Q-E} : x_{Q-E,i} \in (50500.5, 52255.8] \xrightarrow{0.0} classer383, \\ r_{1, classer383}^{QB-B} : x_{QB-B,i} \in [29397.3, 33549.4) \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{QB-B} : x_{QB-B,i} \in [33549.4, 39000.0] \xrightarrow{0.1355} classer383, \\ r_{3, classer383}^{QB-B} : x_{QB-B,i} \in (39000.0, 49695.8] \xrightarrow{0.0} classer383, \\ r_{1, classer383}^{QR-G} : x_{QR-G,i} \in [27351.0, 28343.8) \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{QR-G} : x_{QR-G,i} \in [28343.8, 44568.6] \xrightarrow{0.1115} classer383, \\ r_{3, classer383}^{QR-G} : x_{QR-G,i} \in (44568.6, 49527.0] \xrightarrow{0.0} classer383, \\ r_{1, classer383}^{QP-G} : x_{QP-G,i} \in [327.6, 385.9) \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{QP-G} : x_{QP-G,i} \in [385.9, 831.1] \xrightarrow{0.1328} classer383, \\ r_{3, classer383}^{QP-G} : x_{QP-G,i} \in (831.1, 1080.0] \xrightarrow{0.0} classer383, \\ r_{1, classer383}^{QA-G} : x_{QA-G,i} \in [136371.0, 156320.0) \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{QA-G} : x_{QA-G,i} \in [156320.0, 331990.0] \xrightarrow{0.11} classer383, \\ r_{3, classer383}^{QA-G} : x_{QA-G,i} \in (331990.0, 367840.0] \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{FE-E} : x_{FE-E,i} \in [0.0, 63.3] \xrightarrow{0.1093} classer383, \\ r_{3, classer383}^{FE-E} : x_{FE-E,i} \in (63.3, 89.8] \xrightarrow{0.0} classer383, \\ r_{1, classer383}^{PH-E} : x_{PH-E,i} \in [7.2, 7.3) \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{PH-E} : x_{PH-E,i} \in [7.3, 7.8] \xrightarrow{0.1301} classer383, \\ r_{3, classer383}^{PH-E} : x_{PH-E,i} \in (7.8, 7.9] \xrightarrow{0.0} classer383, \\ r_{1, classer383}^{SS-E} : x_{SS-E,i} \in [82.0, 114.0) \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{SS-E} : x_{SS-E,i} \in [114.0, 480.0] \xrightarrow{0.1098} classer383, \\ r_{3, classer383}^{SS-E} : x_{SS-E,i} \in (480.0, 655.0] \xrightarrow{1.0} classer383, \\ r_{1, classer383}^{SSV-E} : x_{SSV-E,i} \in [60.0, 92.0) \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{SSV-E} : x_{SSV-E,i} \in [92.0, 336.0] \xrightarrow{0.1148} classer383, \\ r_{3, classer383}^{SSV-E} : x_{SSV-E,i} \in (336.0, 593.0] \xrightarrow{1.0} classer383, \\ r_{1, classer383}^{DQO-E} : x_{DQO-E,i} \in [158.0, 414.0) \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{DQO-E} : x_{DQO-E,i} \in [414.0, 1279.0] \xrightarrow{0.1558} classer383, \\ r_{3, classer383}^{DQO-E} : x_{DQO-E,i} \in (1279.0, 1579.0] \xrightarrow{1.0} classer383, \\ r_{1, classer383}^{DBO-E} : x_{DBO-E,i} \in [90.0, 220.0) \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{DBO-E} : x_{DBO-E,i} \in [220.0, 382.0] \xrightarrow{0.22} classer383, \\ r_{3, classer383}^{DBO-E} : x_{DBO-E,i} \in (382.0, 987.0] \xrightarrow{1.0} classer383, \\ r_{1, classer383}^{PH-D} : x_{PH-D,i} \in [7.2, 7.3) \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{PH-D} : x_{PH-D,i} \in [7.3, 7.8] \xrightarrow{0.1275} classer383, \\ r_{3, classer383}^{PH-D} : x_{PH-D,i} \in (7.8, 7.9] \xrightarrow{0.0} classer383, \\ r_{1, classer383}^{SS-D} : x_{SS-D,i} \in [63.0, 68.0) \xrightarrow{0.0} classer383, \\ r_{2, classer383}^{SS-D} : x_{SS-D,i} \in [68.0, 112.0] \xrightarrow{0.1468} classer383, \\ r_{3, classer383}^{SS-D} : x_{SS-D,i} \in (112.0, 192.0] \xrightarrow{0.0} classer383, \end{array} \}$$

$$\begin{aligned}
r_{1, \text{classer383}}^{SSV-D} : x_{SSV-D,i} \in [47.0, 49.0] &\xrightarrow{0.0} \text{classer383}, \\
r_{2, \text{classer383}}^{SSV-D} : x_{SSV-D,i} \in [49.0, 92.0] &\xrightarrow{0.1333} \text{classer383}, \\
r_{3, \text{classer383}}^{SSV-D} : x_{SSV-D,i} \in (92.0, 134.0] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer383}}^{DQO-D} : x_{DQO-D,i} \in [90.0, 186.0) &\xrightarrow{0.0} \text{classer383}, \\
r_{2, \text{classer383}}^{DQO-D} : x_{DQO-D,i} \in [186.0, 329.0] &\xrightarrow{0.1199} \text{classer383}, \\
r_{3, \text{classer383}}^{DQO-D} : x_{DQO-D,i} \in (329.0, 538.0] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer383}}^{DBO-D} : x_{DBO-D,i} \in [56.0, 67.0) &\xrightarrow{0.0} \text{classer383}, \\
r_{2, \text{classer383}}^{DBO-D} : x_{DBO-D,i} \in [67.0, 224.0] &\xrightarrow{0.1255} \text{classer383}, \\
r_{3, \text{classer383}}^{DBO-D} : x_{DBO-D,i} \in (224.0, 274.0] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer383}}^{PH-S} : x_{PH-S,i} \in [7.0, 7.2) &\xrightarrow{0.0} \text{classer383}, \\
r_{2, \text{classer383}}^{PH-S} : x_{PH-S,i} \in [7.2, 7.8] &\xrightarrow{0.1391} \text{classer383}, \\
r_{3, \text{classer383}}^{PH-S} : x_{PH-S,i} \in (7.8, 8.0] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer383}}^{SS-S} : x_{SS-S,i} \in [2.8, 4.8) &\xrightarrow{0.0} \text{classer383}, \\
r_{2, \text{classer383}}^{SS-S} : x_{SS-S,i} \in [4.8, 29.0] &\xrightarrow{0.1455} \text{classer383}, \\
r_{3, \text{classer383}}^{SS-S} : x_{SS-S,i} \in (29.0, 174.8] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer383}}^{SSV-S} : x_{SSV-S,i} \in [1.6, 4.4) &\xrightarrow{0.0} \text{classer383}, \\
r_{2, \text{classer383}}^{SSV-S} : x_{SSV-S,i} \in [4.4, 19.0] &\xrightarrow{0.1608} \text{classer383}, \\
r_{3, \text{classer383}}^{SSV-S} : x_{SSV-S,i} \in (19.0, 134.8] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer383}}^{DQO-S} : x_{DQO-S,i} \in [9.0, 20.0) &\xrightarrow{0.0} \text{classer383}, \\
r_{2, \text{classer383}}^{DQO-S} : x_{DQO-S,i} \in [20.0, 95.0] &\xrightarrow{0.1185} \text{classer383}, \\
r_{3, \text{classer383}}^{DQO-S} : x_{DQO-S,i} \in (95.0, 163.0] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer383}}^{DBO-S} : x_{DBO-S,i} \in [4.0, 6.0) &\xrightarrow{0.0} \text{classer383}, \\
r_{2, \text{classer383}}^{DBO-S} : x_{DBO-S,i} \in [6.0, 35.0] &\xrightarrow{0.1368} \text{classer383}, \\
r_{3, \text{classer383}}^{DBO-S} : x_{DBO-S,i} \in (35.0, 84.0] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer383}}^{V30-B} : x_{V30-B,i} \in [77.0, 140.0) &\xrightarrow{0.0} \text{classer383}, \\
r_{2, \text{classer383}}^{V30-B} : x_{V30-B,i} \in [140.0, 760.0] &\xrightarrow{0.1164} \text{classer383}, \\
r_{3, \text{classer383}}^{V30-B} : x_{V30-B,i} \in (760.0, 770.0] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer383}}^{MLSS-B} : x_{MLSS-B,i} \in [754.0, 1046.0) &\xrightarrow{0.0} \text{classer383}, \\
r_{2, \text{classer383}}^{MLSS-B} : x_{MLSS-B,i} \in [1046.0, 2248.0] &\xrightarrow{0.1172} \text{classer383}, \\
r_{3, \text{classer383}}^{MLSS-B} : x_{MLSS-B,i} \in (2248.0, 2696.0] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer383}}^{MLVSS-B} : x_{MLVSS-B,i} \in [185.0, 611.0) &\xrightarrow{1.0} \text{classer383}, \\
r_{2, \text{classer383}}^{MLVSS-B} : x_{MLVSS-B,i} \in [611.0, 1726.0] &\xrightarrow{0.1088} \text{classer383}, \\
r_{3, \text{classer383}}^{MLVSS-B} : x_{MLVSS-B,i} \in (1726.0, 2054.0] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer383}}^{MCRT-B} : x_{MCRT-B,i} \in [1.8, 6.2) &\xrightarrow{1.0} \text{classer383}, \\
r_{2, \text{classer383}}^{MCRT-B} : x_{MCRT-B,i} \in [6.2, 16.0] &\xrightarrow{0.1088} \text{classer383}, \\
r_{3, \text{classer383}}^{MCRT-B} : x_{MCRT-B,i} \in (16.0, 28.8] &\xrightarrow{0.0} \text{classer383}, \\
r_{1, \text{classer390}}^{Q-E} : x_{Q-E,i} \in [29920.0, 34284.4) &\xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{Q-E} : x_{Q-E,i} \in [34284.4, 50500.5] &\xrightarrow{0.89} \text{classer390},
\end{aligned}$$

$$\begin{aligned}
r_{3, \text{classer390}}^{Q-E} : & x_{Q-E,i} \in (50500.5, 52255.8] \xrightarrow{1.0} \text{classer390}, \\
r_{1, \text{classer390}}^{QB-B} : & x_{QB-B,i} \in [29397.3, 33549.4) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{QB-B} : & x_{QB-B,i} \in [33549.4, 39000.0] \xrightarrow{0.8645} \text{classer390}, \\
r_{3, \text{classer390}}^{QB-B} : & x_{QB-B,i} \in (39000.0, 49695.8] \xrightarrow{1.0} \text{classer390}, \\
r_{1, \text{classer390}}^{QR-G} : & x_{QR-G,i} \in [27351.0, 28343.8) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{QR-G} : & x_{QR-G,i} \in [28343.8, 44568.6] \xrightarrow{0.8885} \text{classer390}, \\
r_{3, \text{classer390}}^{QR-G} : & x_{QR-G,i} \in (44568.6, 49527.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1, \text{classer390}}^{QP-G} : & x_{QP-G,i} \in [327.6, 385.9) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{QP-G} : & x_{QP-G,i} \in [385.9, 831.1] \xrightarrow{0.8672} \text{classer390}, \\
r_{3, \text{classer390}}^{QP-G} : & x_{QP-G,i} \in (831.1, 1080.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1, \text{classer390}}^{QA-G} : & x_{QA-G,i} \in [136371.0, 156320.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{QA-G} : & x_{QA-G,i} \in [156320.0, 331990.0] \xrightarrow{0.89} \text{classer390}, \\
r_{3, \text{classer390}}^{QA-G} : & x_{QA-G,i} \in (331990.0, 367840.0] \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{FE-E} : & x_{FE-E,i} \in [0.0, 63.3] \xrightarrow{0.8907} \text{classer390}, \\
r_{3, \text{classer390}}^{FE-E} : & x_{FE-E,i} \in (63.3, 89.8] \xrightarrow{1.0} \text{classer390}, \\
r_{1, \text{classer390}}^{PH-E} : & x_{PH-E,i} \in [7.2, 7.3) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{PH-E} : & x_{PH-E,i} \in [7.3, 7.8] \xrightarrow{0.8699} \text{classer390}, \\
r_{3, \text{classer390}}^{PH-E} : & x_{PH-E,i} \in (7.8, 7.9] \xrightarrow{1.0} \text{classer390}, \\
r_{1, \text{classer390}}^{SS-E} : & x_{SS-E,i} \in [82.0, 114.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{SS-E} : & x_{SS-E,i} \in [114.0, 480.0] \xrightarrow{0.8902} \text{classer390}, \\
r_{3, \text{classer390}}^{SS-E} : & x_{SS-E,i} \in (480.0, 655.0] \xrightarrow{0.0} \text{classer390}, \\
r_{1, \text{classer390}}^{SSV-E} : & x_{SSV-E,i} \in [60.0, 92.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{SSV-E} : & x_{SSV-E,i} \in [92.0, 336.0] \xrightarrow{0.8852} \text{classer390}, \\
r_{3, \text{classer390}}^{SSV-E} : & x_{SSV-E,i} \in (336.0, 593.0] \xrightarrow{0.0} \text{classer390}, \\
r_{1, \text{classer390}}^{DQO-E} : & x_{DQO-E,i} \in [158.0, 414.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{DQO-E} : & x_{DQO-E,i} \in [414.0, 1279.0] \xrightarrow{0.8442} \text{classer390}, \\
r_{3, \text{classer390}}^{DQO-E} : & x_{DQO-E,i} \in (1279.0, 1579.0] \xrightarrow{0.0} \text{classer390}, \\
r_{1, \text{classer390}}^{DBO-E} : & x_{DBO-E,i} \in [90.0, 220.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{DBO-E} : & x_{DBO-E,i} \in [220.0, 382.0] \xrightarrow{0.78} \text{classer390}, \\
r_{3, \text{classer390}}^{DBO-E} : & x_{DBO-E,i} \in (382.0, 987.0] \xrightarrow{0.0} \text{classer390}, \\
r_{1, \text{classer390}}^{PH-D} : & x_{PH-D,i} \in [7.2, 7.3) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{PH-D} : & x_{PH-D,i} \in [7.3, 7.8] \xrightarrow{0.8725} \text{classer390}, \\
r_{3, \text{classer390}}^{PH-D} : & x_{PH-D,i} \in (7.8, 7.9] \xrightarrow{1.0} \text{classer390}, \\
r_{1, \text{classer390}}^{SS-D} : & x_{SS-D,i} \in [63.0, 68.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{SS-D} : & x_{SS-D,i} \in [68.0, 112.0] \xrightarrow{0.8532} \text{classer390}, \\
r_{3, \text{classer390}}^{SS-D} : & x_{SS-D,i} \in (112.0, 192.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1, \text{classer390}}^{SSV-D} : & x_{SSV-D,i} \in [47.0, 49.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2, \text{classer390}}^{SSV-D} : & x_{SSV-D,i} \in [49.0, 92.0] \xrightarrow{0.8667} \text{classer390},
\end{aligned}$$

$$\begin{aligned}
r_{3,\text{classer390}}^{SSV-D} : & x_{SSV-D,i} \in (92.0, 134.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1,\text{classer390}}^{DQO-D} : & x_{DQO-D,i} \in [90.0, 186.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2,\text{classer390}}^{DQO-D} : & x_{DQO-D,i} \in [186.0, 329.0] \xrightarrow{0.8801} \text{classer390}, \\
r_{3,\text{classer390}}^{DQO-D} : & x_{DQO-D,i} \in (329.0, 538.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1,\text{classer390}}^{DBO-D} : & x_{DBO-D,i} \in [56.0, 67.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2,\text{classer390}}^{DBO-D} : & x_{DBO-D,i} \in [67.0, 224.0] \xrightarrow{0.8745} \text{classer390}, \\
r_{3,\text{classer390}}^{DBO-D} : & x_{DBO-D,i} \in (224.0, 274.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1,\text{classer390}}^{PH-S} : & x_{PH-S,i} \in [7.0, 7.2) \xrightarrow{1.0} \text{classer390}, \\
r_{2,\text{classer390}}^{PH-S} : & x_{PH-S,i} \in [7.2, 7.8] \xrightarrow{0.8609} \text{classer390}, \\
r_{3,\text{classer390}}^{PH-S} : & x_{PH-S,i} \in (7.8, 8.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1,\text{classer390}}^{SS-S} : & x_{SS-S,i} \in [2.8, 4.8) \xrightarrow{1.0} \text{classer390}, \\
r_{2,\text{classer390}}^{SS-S} : & x_{SS-S,i} \in [4.8, 29.0] \xrightarrow{0.8545} \text{classer390}, \\
r_{3,\text{classer390}}^{SS-S} : & x_{SS-S,i} \in (29.0, 174.8] \xrightarrow{1.0} \text{classer390}, \\
r_{1,\text{classer390}}^{SSV-S} : & x_{SSV-S,i} \in [1.6, 4.4) \xrightarrow{1.0} \text{classer390}, \\
r_{2,\text{classer390}}^{SSV-S} : & x_{SSV-S,i} \in [4.4, 19.0] \xrightarrow{0.8392} \text{classer390}, \\
r_{3,\text{classer390}}^{SSV-S} : & x_{SSV-S,i} \in (19.0, 134.8] \xrightarrow{1.0} \text{classer390}, \\
r_{1,\text{classer390}}^{DQO-S} : & x_{DQO-S,i} \in [9.0, 20.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2,\text{classer390}}^{DQO-S} : & x_{DQO-S,i} \in [20.0, 95.0] \xrightarrow{0.8815} \text{classer390}, \\
r_{3,\text{classer390}}^{DQO-S} : & x_{DQO-S,i} \in (95.0, 163.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1,\text{classer390}}^{DBO-S} : & x_{DBO-S,i} \in [4.0, 6.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2,\text{classer390}}^{DBO-S} : & x_{DBO-S,i} \in [6.0, 35.0] \xrightarrow{0.8632} \text{classer390}, \\
r_{3,\text{classer390}}^{DBO-S} : & x_{DBO-S,i} \in (35.0, 84.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1,\text{classer390}}^{V30-B} : & x_{V30-B,i} \in [77.0, 140.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2,\text{classer390}}^{V30-B} : & x_{V30-B,i} \in [140.0, 760.0] \xrightarrow{0.8836} \text{classer390}, \\
r_{3,\text{classer390}}^{V30-B} : & x_{V30-B,i} \in (760.0, 770.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1,\text{classer390}}^{MLSS-B} : & x_{MLSS-B,i} \in [754.0, 1046.0) \xrightarrow{1.0} \text{classer390}, \\
r_{2,\text{classer390}}^{MLSS-B} : & x_{MLSS-B,i} \in [1046.0, 2248.0] \xrightarrow{0.8828} \text{classer390}, \\
r_{3,\text{classer390}}^{MLSS-B} : & x_{MLSS-B,i} \in (2248.0, 2696.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1,\text{classer390}}^{MLVSS-B} : & x_{MLVSS-B,i} \in [185.0, 611.0) \xrightarrow{0.0} \text{classer390}, \\
r_{2,\text{classer390}}^{MLVSS-B} : & x_{MLVSS-B,i} \in [611.0, 1726.0] \xrightarrow{0.8912} \text{classer390}, \\
r_{3,\text{classer390}}^{MLVSS-B} : & x_{MLVSS-B,i} \in (1726.0, 2054.0] \xrightarrow{1.0} \text{classer390}, \\
r_{1,\text{classer390}}^{MCRT-B} : & x_{MCRT-B,i} \in [1.8, 6.2) \xrightarrow{0.0} \text{classer390}, \\
r_{2,\text{classer390}}^{MCRT-B} : & x_{MCRT-B,i} \in [6.2, 16.0] \xrightarrow{0.8912} \text{classer390}, \\
r_{3,\text{classer390}}^{MCRT-B} : & x_{MCRT-B,i} \in (16.0, 28.8] \xrightarrow{1.0} \text{classer390} \quad \}
\end{aligned}$$

Anexo F

Análisis descriptivo de los datos, planta eslovena

F.1 Análisis univariante

F.1.1 Variable NH4-influent

En el Histograma que se muestra en la Figura F.1 se observa claramente cómo este grupo de valores queda al margen del comportamiento normal y según lo anteriormente comentado convierte la distribución en asimétrica aplanada a la izquierda y muy dispersa.

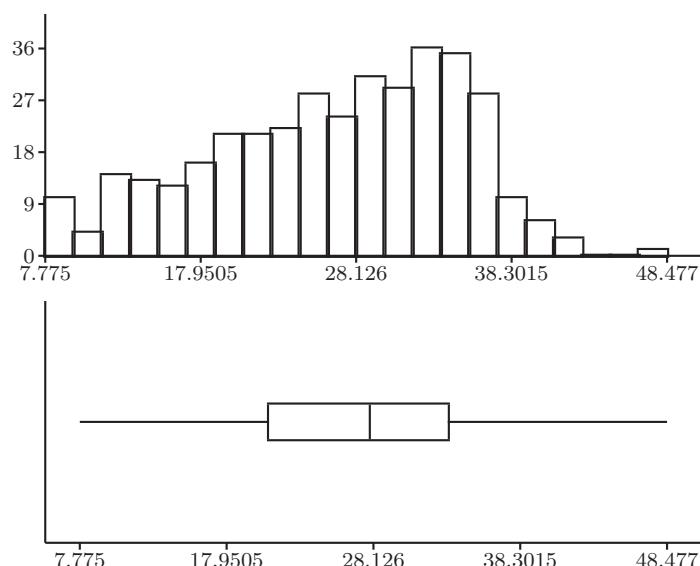


Figura F.1: Histograma y Boxplot de la variable NH4-influent.

Tenemos una concentración de amonio diaria en la entrada de la planta piloto de 26.6923 miligramos por litro y por día en media con una desviación de 8.151 miligramos por litro y con unos valores que van de 7.775 a 48.477 miligramos por litro. Existe un dato missing para esta variable correspondiente al 08-02-06, registro 253 de la base de datos.

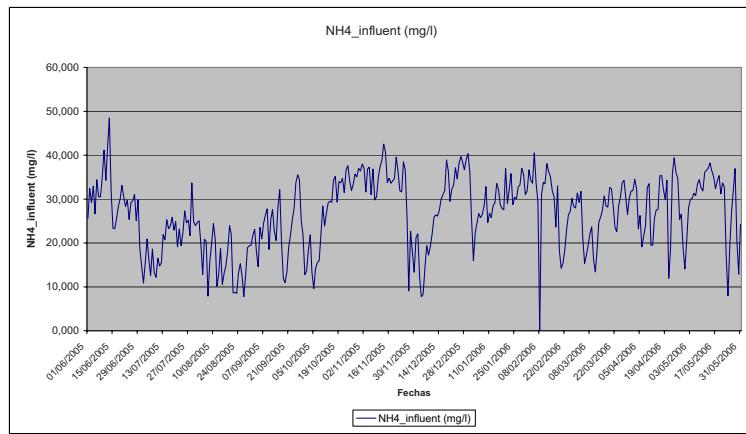


Figura F.2: Serie temporal para variable NH4-influent.

Estadísticos	
Número de observaciones	365
Número de missings	1
Número de observaciones útiles	364
Mitjana	26.6923
Mediana	27.875
Primer quartil (Q1)	20.8925
Tercer quartil (Q3)	33.279
Mínimo	7.775
Màximo	48.477
Quasi-desviación típica	8.151
Coeficiente de variación	0.3049

Tabla de frecuencias	
Modalidades	Freq. absolut.
7.775 - 9.6251	10
9.6251 - 11.4752	4
11.4752 - 13.3253	14
13.3253 - 15.1754	13
15.1754 - 17.0255	12
17.0255 - 18.8755	16
18.8755 - 20.7256	21
20.7256 - 22.5757	21
22.5757 - 24.4258	22
24.4258 - 26.2759	28
26.2759 - 28.126	24
28.126 - 29.9761	31
29.9761 - 31.8262	29
31.8262 - 33.6763	36
33.6763 - 35.5264	35
35.5264 - 37.3764	28
37.3764 - 39.2265	10
39.2265 - 41.0766	6
41.0766 - 42.9267	3
42.9267 - 44.7768	0
44.7768 - 46.6269	0
46.6269 - 48.477	1
Missings	1

F.1.2 Variable NH4-2aerobic

La media de esta variable es de 3.0864 miligramos por litro con una desviación de 4.3451 miligramos por litro, con unos valores que van de 0 a 20.282 miligramos por litro.

En el Histograma, que se muestra en la Figura F.4, se observa que casi todos los valores

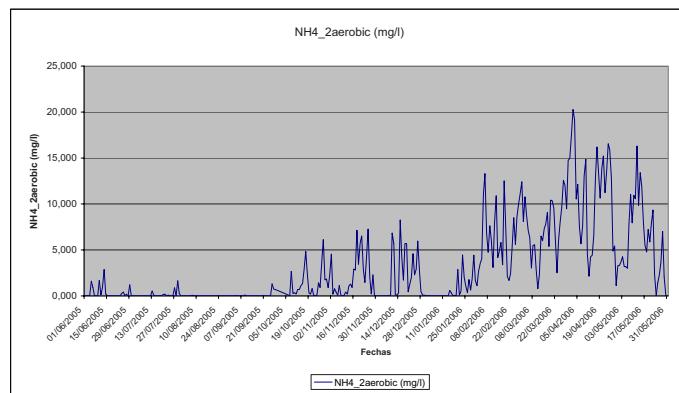


Figura F.3: Serie temporal para variable NH4-2aerobic.

están incluidos en el primer intervalo (191 observaciones). Son tan pocos el número de valores extremos (12 valores) que casi no se observa la/s categoría/s que los incluyen. Es una distribución fuertemente asimétrica aplanada a la derecha con alta concentración de valores bajos y cola muy larga a la derecha.

Tabla de frecuencias	
Modalidades	Freq. absolut.
0 - 0.9219	191
0.9219 - 1.8438	24
1.8438 - 2.7657	19
2.7657 - 3.6876	20
3.6876 - 4.6095	14
4.6095 - 5.5315	10
5.5315 - 6.4534	16
6.4534 - 7.3753	12
7.3753 - 8.2972	12
8.2972 - 9.2191	4
9.2191 - 10.141	6
10.141 - 11.0629	10
11.0629 - 11.9848	3
11.9848 - 12.9067	7
12.9067 - 13.8286	5
13.8286 - 14.7505	2
14.7505 - 15.6725	3
15.6725 - 16.5944	4
16.5944 - 17.5163	1
17.5163 - 18.4382	0
18.4382 - 19.3601	1
19.3601 - 20.282	1
Missings	0

Estadísticos	
Número de observaciones	365
Número de missings	0
Número de observaciones útiles	365
Media	3.0864
Mediana	0.7
Primer quartil (Q1)	0.004
Tercer quartil (Q3)	5.4115
Mínimo	0
Máximo	20.282
Quasi-desviación típica	4.3451
Coeficiente de variación	1.4059

La media aparece inclinada hacia el lado izquierdo de la caja, que es donde se concentran la mayor cantidad de observaciones, aparecen datos de valor muy diferenciado que se dan

repentinamente y no de forma gradual, esta variable presenta una alta variabilidad (mas del 140%). Se observan claramente un gran pico de valores que predomina sobre el resto; existen 193 observaciones que corresponden a valores todos por debajo de 1 miligramo por litro, cuando sólo 92 observaciones superan los 5 miligramos por litro y 38 observaciones los 10 miligramos por litro. Por debajo de la media se encuentran 245 observaciones, lo que corresponde al 67% del total de observaciones. Existe un claro dato atípico (20.282 mg/l) en la observación nº:306 el día 2-abr-2006, ver Figura F.3.

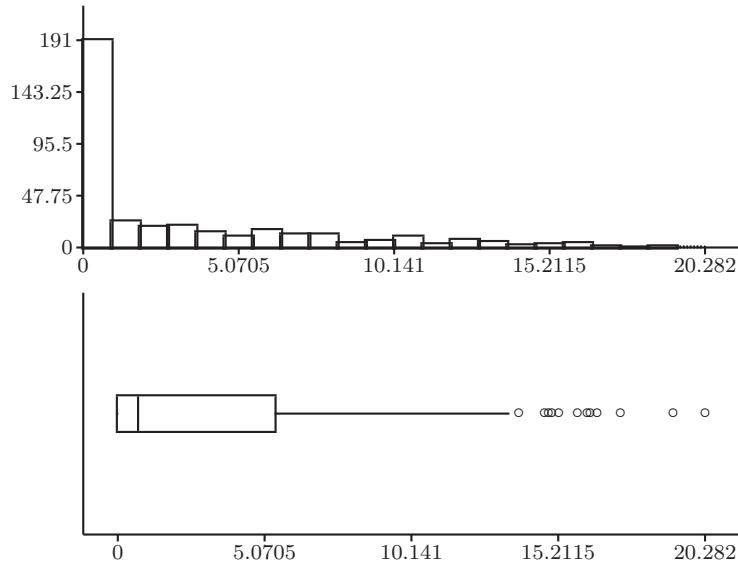


Figura F.4: Histograma y Boxplot de la variable NH4-2aerobic.

F.1.3 Variable O2-1aerobic

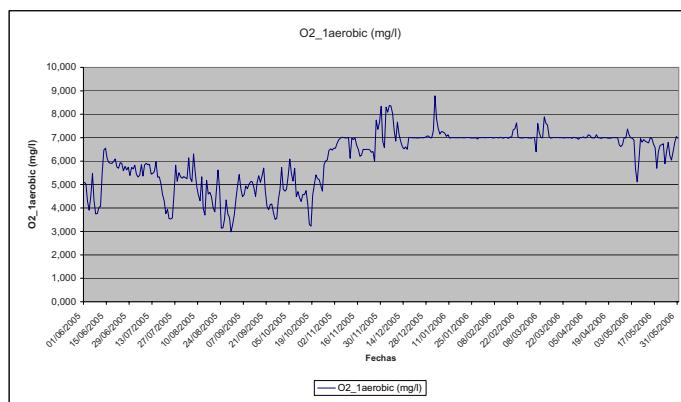


Figura F.5: Serie temporal para variable O2-1aerobic.

En el Histograma que se muestra en la Figura F.6 se observa claramente que un gran grupo de observaciones se concentra en un solo intervalo que representa el 36% de las observaciones (132 valores) con valores que se encuentran entre los 6.938 y los 7.2018 miligramos por litro de la concentración de oxígeno disuelto en el primer tanque aerobic de la planta piloto. Hemos comprobado que 8.785 miligramos por litro es la capacidad máxima de oxígeno disuelto en el primer tanque.

Tenemos una media de 6.1212 miligramos por litro de la concentración de oxígeno disuelto en el primer tanque aerobic por día con una desviación de 1.1795 miligramos por litro, con unos valores que van de 2.98 a 8.785 miligramos por litro. En esta variable no se observan ni dato atípico ni missing, la variabilidad de los datos no se mantiene constante, presentando un índice de variabilidad que se considera alto, 19%.

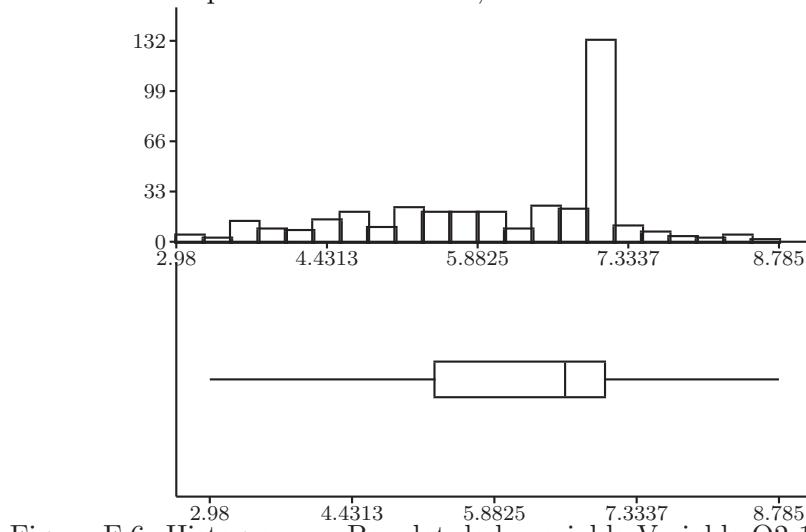


Figura F.6: Histograma y Boxplot de la variable Variable O2-1aerobic.

Tabla de frecuencias	
Modalidades	Freq. absolut.
2.98 - 3.2439	4
3.2439 - 3.5077	2
3.5077 - 3.7716	13
3.7716 - 4.0355	8
4.0355 - 4.2993	7
4.2993 - 4.5632	14
4.5632 - 4.827	19
4.827 - 5.0909	9
5.0909 - 5.3548	22
5.3548 - 5.6186	19
5.6186 - 5.8825	19
5.8825 - 6.1464	19
6.1464 - 6.4102	8
6.4102 - 6.6741	23
6.6741 - 6.938	21
6.938 - 7.2018	132
7.2018 - 7.4657	10
7.4657 - 7.7295	6
7.7295 - 7.9934	3
7.9934 - 8.2573	2
8.2573 - 8.5211	4
8.5211 - 8.785	1
Missings	0

F.1.4 Variable O2-2aerobic

En el Histograma que se muestra en la Figura F.8 se observa claramente como un grupo de valores queda al margen del comportamiento normal (observaciones entre 5.8427 y 6.0843 miligramos por litro de concentración de oxígeno disuelto, 63 observaciones, correspondiente al 17% del total) y cómo el periodo anteriormente comentado convierte la distribución en asimétrica aplanada a la izquierda con un pico que concentra el número de observaciones anteriormente comentado.

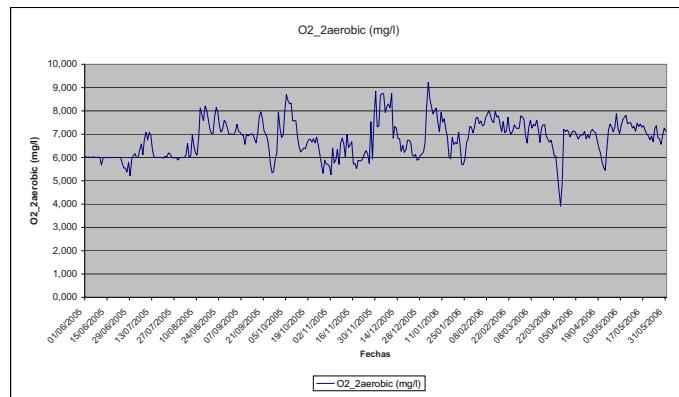


Figura F.7: Serie temporal para variable O2-2aerobic.

Una media de 6.876 miligramos por litro de la concentración de oxígeno disuelto en el segundo tanque aerobico, en la planta piloto, por día con una desviación de 0.7949 miligramos por litro, con unos valores que van de 3.91 a 9.225 miligramos por litro.

En esta variable no se observa ningún missing, la variabilidad de los datos se mantiene constante con un coeficiente de variación bajo, 11%. Existe un claro dato atípico en la observación nº:299 el día 26-mar-2006.

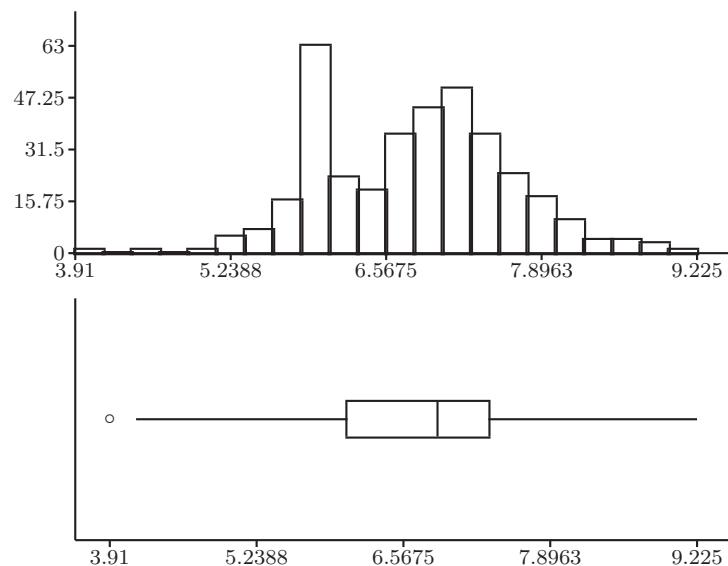


Figura F.8: Histograma y Boxplot de la variable Variable O2-2aerobic.

Tabla de frecuencias	
Modalidades	Freq. absol.
3.91 - 4.1516	1
4.1516 - 4.3932	0
4.3932 - 4.6348	1
4.6348 - 4.8764	0
4.8764 - 5.118	1
5.118 - 5.3595	5
5.3595 - 5.6011	7
5.6011 - 5.8427	16
5.8427 - 6.0843	63
6.0843 - 6.3259	23
6.3259 - 6.5675	19
6.5675 - 6.8091	36
6.8091 - 7.0507	44
7.0507 - 7.2923	50
7.2923 - 7.5339	36
7.5339 - 7.7755	24
7.7755 - 8.017	17
8.017 - 8.2586	10
8.2586 - 8.5002	4
8.5002 - 8.7418	4
8.7418 - 8.9834	3
8.9834 - 9.225	1
<i>Missings</i>	0

F.1.5 Variable Valve-air

La media de esta variable es de 41.212 % con una desviación de 8.688 %, con unos valores que van desde el 28.604% al 69.898%, lo cual quiere decir que en un año e funcionamiento de la planta la válvula nunca se ha cerrado ni abierto por completo.

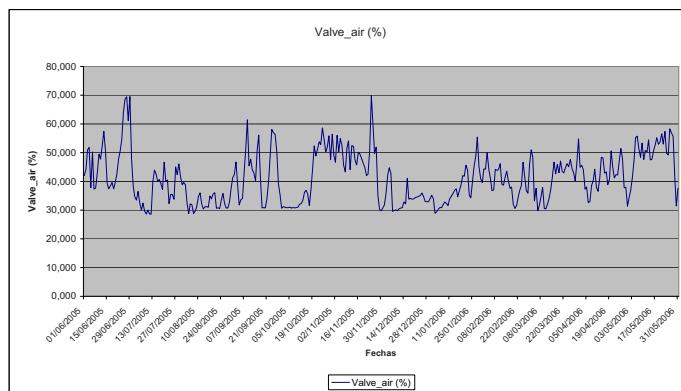


Figura F.9: Serie temporal para variable Valve-air.

En el Histograma, que se muestra en la Figura F.10, se observa que un número importante de valores están incluidos en el intervalo que va desde 30.481% a 32.358 %, con 48

observaciones lo que se observa en el pico que se destaca por sobre los demás (13% del total).

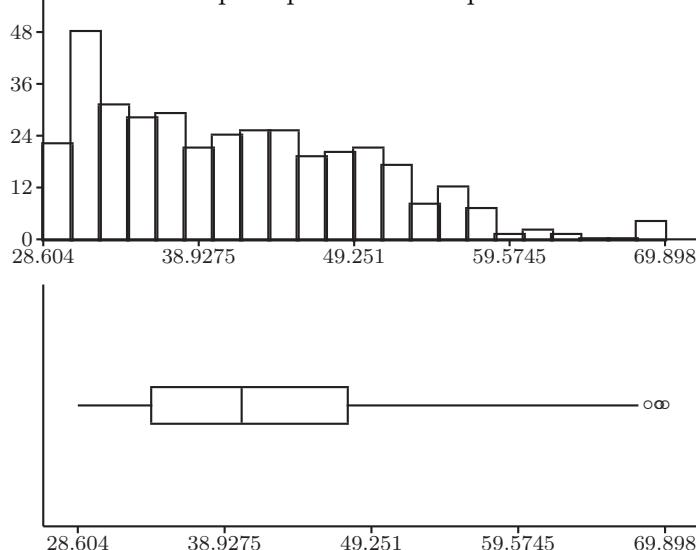


Figura F.10: Histograma y Boxplot de la variable Variable Valve-air.

Tabla de frecuencias	
Modalidades	Freq. absolut.
28.604 - 30.481	22
30.481 - 32.358	48
32.358 - 34.235	31
34.235 - 36.112	28
36.112 - 37.989	29
37.989 - 39.866	21
39.866 - 41.743	24
41.743 - 43.62	25
43.62 - 45.497	25
45.497 - 47.374	19
47.374 - 49.251	20
49.251 - 51.128	21
51.128 - 53.005	17
53.005 - 54.882	8
54.882 - 56.759	12
56.759 - 58.636	7
58.636 - 60.513	1
60.513 - 62.39	2
62.39 - 64.267	1
64.267 - 66.144	0
66.144 - 68.021	0
68.021 - 69.898	4
Missings	0

Es una distribución asimétrica aplana a la derecha.

La media aparece inclinada hacia el lado izquierdo de la caja, ver Figura F.10, que es donde se concentran la mayor cantidad de observaciones

Son tan pocos los valores extremos (8 valores) que casi no se observa la/s categoría/s que los incluyen, ver Figura F.10. La variabilidad de los datos no se mantiene constante, presnetandose un coeficiente de variación alto 21%.

Existen 3 valores por encima de 69% que pueden considerarse atípicos, ellos son las observaciones:

- n°:27 el día 27-jun-2005
- n°:29 el día 29-jun-2005
- n°:177 el día 24-nov-2005

F.1.6 Variable Q-air

En esta variable la variabilidad de los datos no se mantiene constante, y se presenta un coeficiente de variabilidad de 29%.

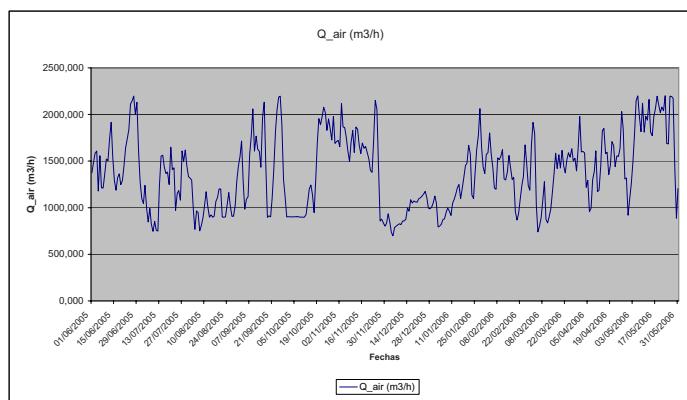


Figura F.11: Serie temporal para variable Q-air.

En el Histograma que se muestra en la Figura F.12 se observa claramente como esta variable queda al margen del comportamiento normal y cómo lo anteriormente comentado convierte la distribución en asimétrica con 2 picos que sobresalen de los demás.

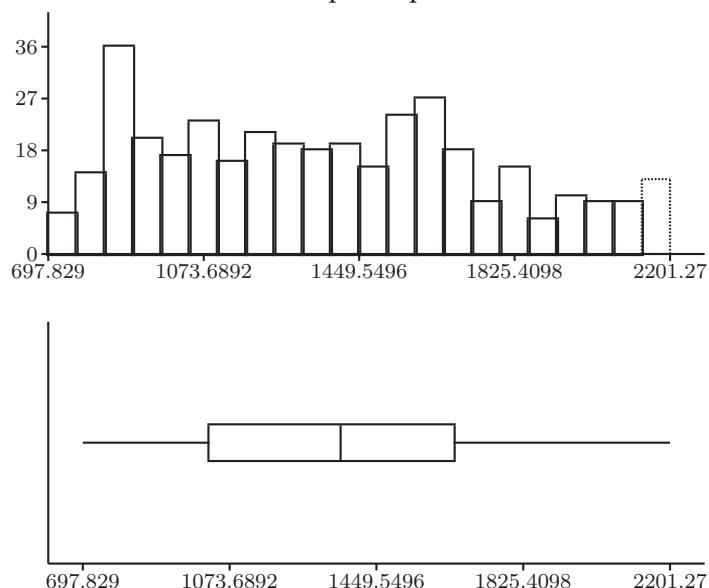


Figura F.12: Histograma y Boxplot de la variable Variable Q-air.

Uno de ellos es el intervalo que va desde 834.5055 metros cúbicos por hora a 902.8438 metros cúbicos por hora, con 36 observaciones y intervalo que va desde 1586.2263 metros cúbicos por hora a 1654.5646 metros cúbicos por hora, con 27 observaciones.

Tenemos una media de 1374.495 metros cúbicos por hora en el caudal de aire, en la planta piloto, con una desviación de 395.9364 metros cúbicos por hora, con unos valores que van desde 697.829 a 2201.27 metros cúbicos por hora.

En esta variable no se observan ningún missing.

No existen datos atípicos.

Tabla de frecuencias		
Modalidades	Freq. absol.	
697.829 - 766.1672	7	
766.1672 - 834.5055	14	
834.5055 - 902.8438	36	
902.8438 - 971.182	20	
971.182 - 1039.5203	17	
1039.5203 - 1107.8585	23	
1107.8585 - 1176.1968	16	
1176.1968 - 1244.535	21	
1244.535 - 1312.8733	19	
1312.8733 - 1381.2115	18	
1381.2115 - 1449.5498	19	
1449.5498 - 1517.8881	15	
1517.8881 - 1586.2263	24	
1586.2263 - 1654.5646	27	
1654.5646 - 1722.9028	18	
1722.9028 - 1791.2411	9	
1791.2411 - 1859.5793	15	
1859.5793 - 1927.9176	6	
1927.9176 - 1996.2559	10	
1996.2559 - 2064.594	9	
2064.594 - 2132.9321	9	
2132.9321 - 2201.2703	13	
Missings	0	

F.1.7 Variable h-ww

La variabilidad de los datos se mantiene constante, con un coeficiente de variación bajo, 1%.

Tenemos una media de 6.876 metros de altura del agua residual en el segundo tanque aerobico, en la planta piloto, promedio por día con una desviación de 0.031 metros de altura del agua residual, con unos valores que van de 2.66 a 3.098 metros de altura del agua residual.

En esta variable no se observan ningún missing, la variabilidad de los datos se mantiene constante. Existen 2 claros datos atípicos uno en la observación nº:164 el día 11-nov-2005 y otro en la observación nº:11 el día 11-jun-2005.

A partir del histograma que se muestra en la Figura F.14 se puede observar que el pico mas alto corresponde al 39% del total de observaciones (142 valores) para el intervalo comprendido entre 2.9985 y 3.0184 metros de altura del agua residual.

Si omitiéramos los datos atípicos obtendríamos una variable con un claro comportamiento normal.

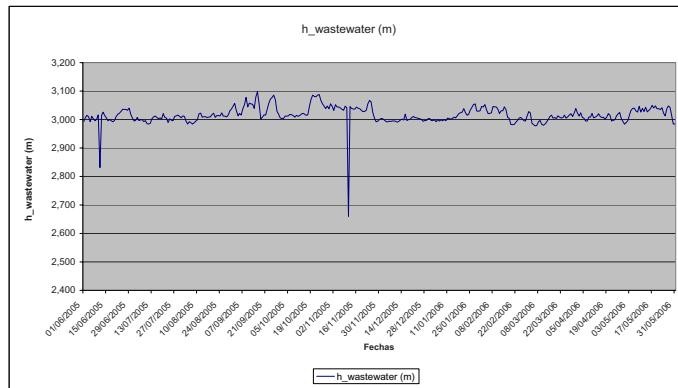


Figura F.13: Serie temporal para variable h-wastewater.

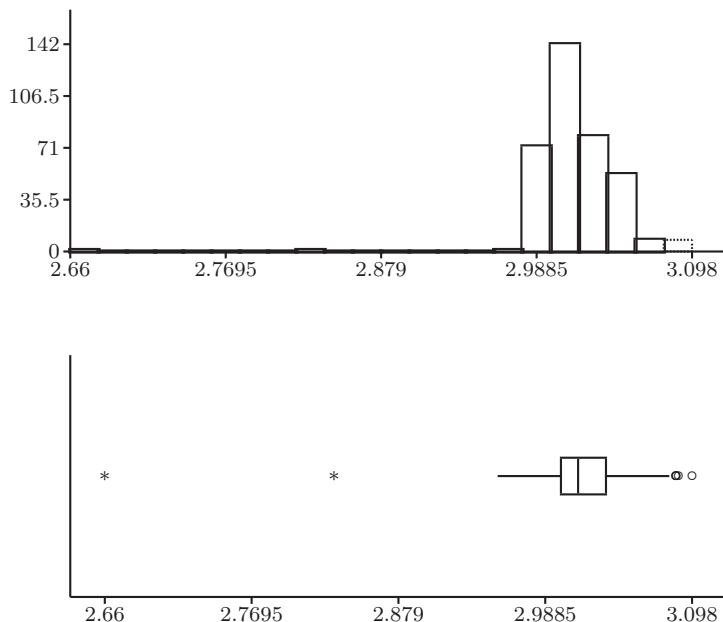


Figura F.14: Histograma y Boxplot de la variable Variable h-ww.

Tabla de frecuencias	
Modalidades	Freq. absol.
2.66 - 2.6799	1
2.6799 - 2.6998	0
2.6998 - 2.7197	0
2.7197 - 2.7396	0
2.7396 - 2.7595	0
2.7595 - 2.7795	0
2.7795 - 2.7994	0
2.7994 - 2.8193	0
2.8193 - 2.8392	1
2.8392 - 2.8591	0
2.8591 - 2.879	0
2.879 - 2.8989	0
2.8989 - 2.9188	0
2.9188 - 2.9387	0
2.9387 - 2.9586	0
2.9586 - 2.9785	1
2.9785 - 2.9985	72
2.9985 - 3.0184	142
3.0184 - 3.0383	79
3.0383 - 3.0582	53
3.0582 - 3.0781	8
3.0781 - 3.098	8
Missings	0

F.1.8 Variable Q-influent

El caudal de entrada a la planta piloto, que proviene del tratamiento mecánico, de 67.4106 m³/día en media, con una desviación de 10.46 m³/día y unos valores que van de 49.706 a 85.50. No existe ningún dato missing para esta variable.

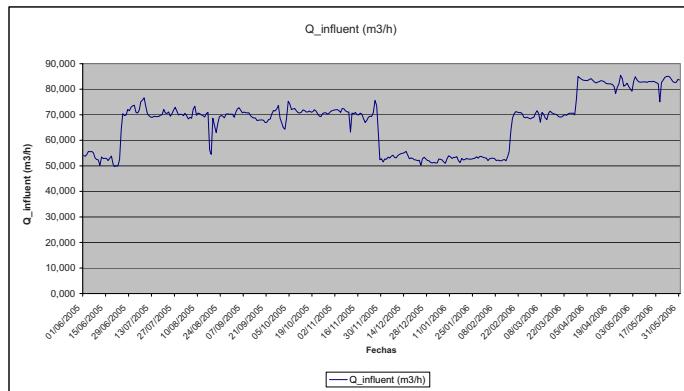


Figura F.15: Serie temporal para variable Q-influent.

La distribución que se muestra en la Figura F.16 presenta varios picos correspondientes a los distintos niveles del caudal de entrada.

Se observa un comportamiento de los datos bastante cambiante, concretamente encontramos 4 cambios de media, podemos distinguir un primer período que va desde el 1-jun-2005 al 23-jun-2005 con una media de 52,881 m³/día.

Un segundo período que comprende un intervalo que esta entre los días 24-jun-2005 y 28-nov-2005, con 158 observaciones (días), se observa una media 70,323 m³/día con valores que se encuentran entre 63,038 m³/día y 76,639 m³/día, quitando 2 observaciones que se podrían considerar como outlier dentro de este periodo, estas son: día 17-agosto-2005 (56,119 m³/día) y el día 18-agosto-2005 (54,497 m³/día).

A continuación encontramos un tercer periodo que comienza el día 29-nov-2005 y finaliza el día 16-feb-2006 (80 observaciones (días)), con una media de 52,825 m³/día y valores que se encuantran entre 50,245 m³/día y 55,666 m³/día. Un cuarto período que va desde el 17-febrero-2006 al 28-marzo-2006 con 44 bservaciones (días), una media de 69,733 m³/día, un mínimo de 63,626 m³/día y un máximo de 71,555 m³/día

Finalmente encontramos un último periodo que comienza el día 29-marzo-2006 m³/día y finaliza el día 31-mayo-2006 (64 observaciones (días)) con una media de 82,693 m³/día y un mínimo 75,068 m³/día y máximo 85,500 m³/día.

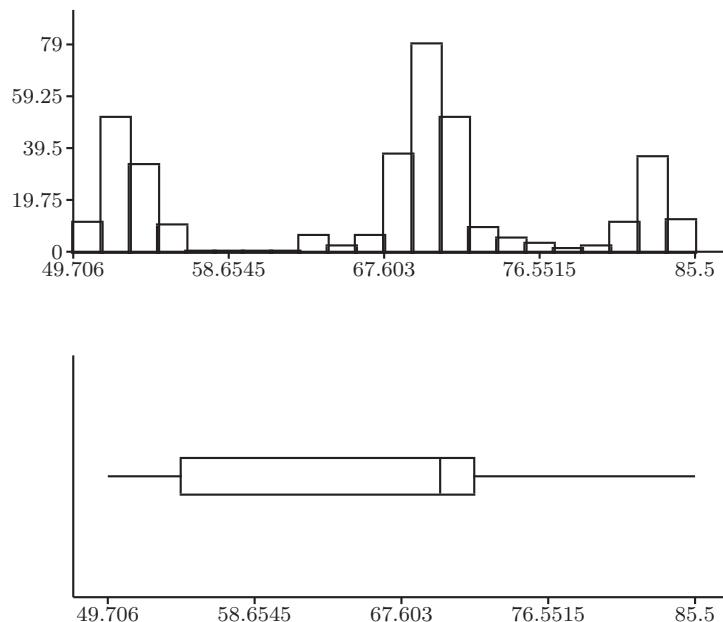


Figura F.16: Histograma y Boxplot de la variable Variable Q-influent.

Este hecho nos podría hacer pensar que esta variable cambia según la estación del año en la que nos encontremos pero, aunque no encontramos los cambios de media próximos a los cambios de estación (sino en la parte central de estos), si que observamos valores más elevados para verano y más bajos para invierno.

Sin embargo al tratarse de una planta piloto, el caudal de entrada que proviene del tratamiento mecánico puede ser periódicamente ajustado para observar el funcionamiento de la planta en diferentes momentos de retención hidráulica, mediante la valvula 1, ver Figure 19.10.

La variabilidad de los datos es muy cambiante debido al patrón anteriormente comentado y por lo tanto no podríamos establecer un comportamiento estacional claro de esta variable.

Estadísticos	
Número de observaciones	365
Número de missings	0
Número de observaciones útiles	365
Media	67.4106
Mediana	69.963
Primer quartil (Q1)	54.214
Tercer quartil (Q3)	71.9765
Mínimo	49.706
Máximo	85.5
Quasi-desviación típica	10.4641
Coeficiente de variación	0.155

Tabla de frecuencias	
Modalidades	Freq. absolut.
49.706 - 51.333	11
51.333 - 52.96	51
52.96 - 54.587	33
54.587 - 56.214	10
56.214 - 57.841	0
57.841 - 59.468	0
59.468 - 61.095	0
61.095 - 62.722	0
62.722 - 64.349	6
64.349 - 65.976	2
65.976 - 67.603	6
67.603 - 69.23	37
69.23 - 70.857	79
70.857 - 72.484	51
72.484 - 74.111	9
74.111 - 75.738	5
75.738 - 77.365	3
77.365 - 78.992	1
78.992 - 80.619	2
80.619 - 82.246	11
82.246 - 83.873	36
83.873 - 85.5	12
Missings	0

F.1.9 Variable FR1-DOTOK

La media de esta variable es de 48.48(Hz) por año con una desviación de 2.427, con unos valores que van de 39.153(Hz) a 50.733(Hz), con un coeficiente de variaación del 0.05, por lo que corresponde a una variable con muy poca variabilidad y un rango de variación también muy pequeño. No existen missings para esta variable.

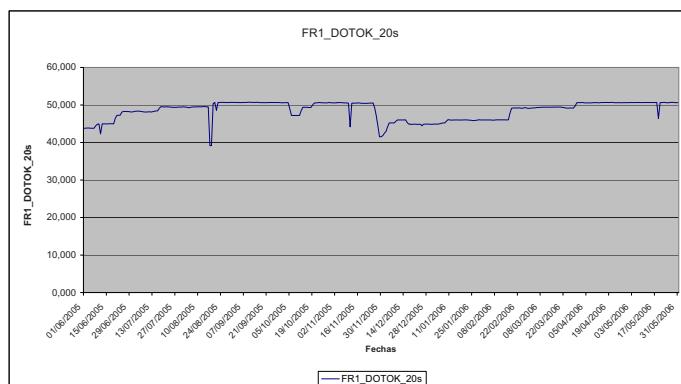


Figura F.17: Serie temporal para variable FR1-DOTOK-20s.

En el Histograma, que se muestra en la Figura F.18, se observa que la mayor parte de

valores están concentrados en un pequeño intervalo con 144 observaciones entre 50.25 (Hz) y 50.35 (Hz), tal y como esperábamos. Se observan que existen 2 observaciones que se podrían considerar como outlier; la número 78 el día 17-agosto-05 bajando a un valor de 39,153 y la número 79 el día 18-agosto-05 con una valor de 39,184. Existe un grupo de observaciones (5 observaciones) que también toman valores bajos, desde el 29-noviembre-05 al 3-diciembre-05 (41.5, 41.538, 41.791, 42.449 42.965).

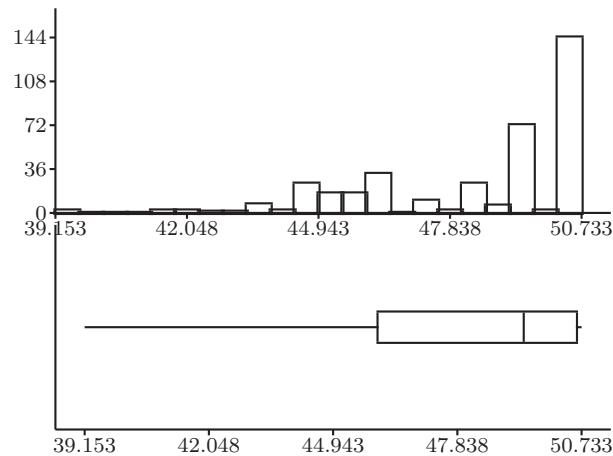


Figura F.18: Histograma y Boxplot de la variable Variable FR1-DOTOK.

Tabla de frecuencias	
Modalidades	Freq. absol.
39.153 - 39.6794	2
39.6794 - 40.2057	0
40.2057 - 40.7321	0
40.7321 - 41.2585	0
41.2585 - 41.7848	2
41.7848 - 42.3112	2
42.3112 - 42.8375	1
42.8375 - 43.3639	1
43.3639 - 43.8903	7
43.8903 - 44.4166	2
44.4166 - 44.943	24
44.943 - 45.4694	16
45.4694 - 45.9957	16
45.9957 - 46.5221	32
46.5221 - 47.0485	0
47.0485 - 47.5748	10
47.5748 - 48.1012	2
48.1012 - 48.6275	24
48.6275 - 49.1539	6
49.1539 - 49.6803	72
49.6803 - 50.2066	2
50.2066 - 50.733	144
<i>Missings</i>	0

F.1.10 Variable Freq-rec

En el Histograma que se muestra en la Figura F.20, se observa que la mayor parte de valores están concentrados en 3 grupos (intervalos). El primero de ellos esta comprendido comprendido entre 24.7774 (Hz) y 25.6927(Hz) con 101 observaciones, para fechas entre el 18 de febrero de 2006 y el 31 de mayo de 2006.

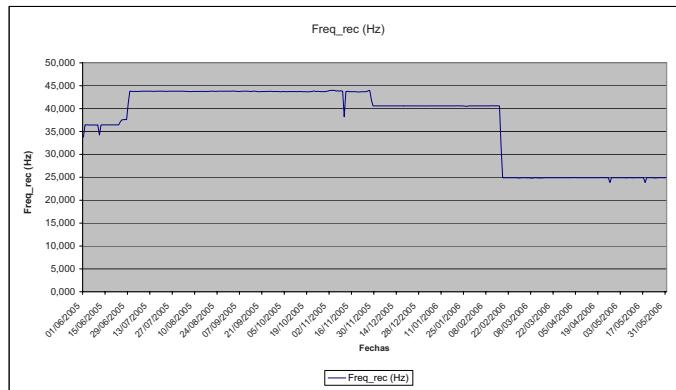


Figura F.19: Serie temporal para variable Freq-rec.

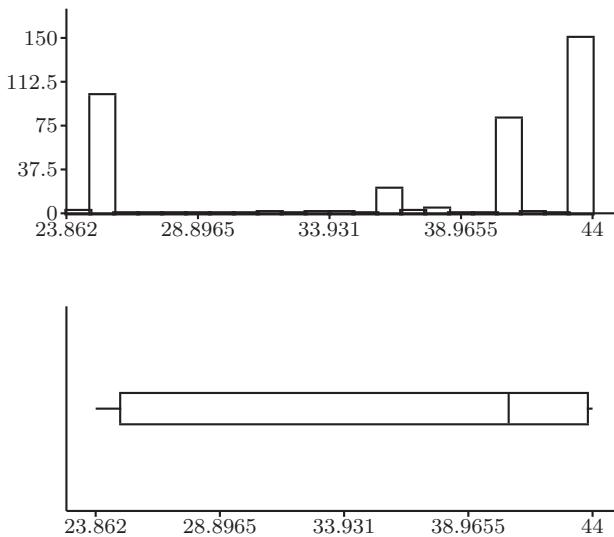


Figura F.20: Histograma y Boxplot de la variable Variable Freq-rec.

El segundo de ellos comprendido entre 40.3385 (Hz) y 41.2539(Hz) con 81 observaciones y el tercer intervalo, comprendido entre 43.0846 (Hz) y 44(Hz) con 150 observaciones.

En conclusion el 92% de las observaciones están concentradas en estos 3 intervalos (o grupos).

Con esto podemos decir que la reticulación en controlada a 3 niveles:

- nivel 1: 24.7774 - 25.6927
- nivel 2: 40.3385 - 41.2539
- nivel 3: 43.0846 - 44

Es importante comentar un outlier que aparece el día 11 de noviembre de 2005, que hace que desde el valor 43,786(Hz)(10-nov-2005), caiga a 38,206(Hz) y luego vuelva a subir a 43,699(Hz)(12-nov-2005).

La media de esta variable es de 37.12(Hz) por año con una desviación de 7.976, con unos valores que van de 22.862(Hz) a 44(Hz), por lo que corresponde a una variable con muy poca variabilidad y un rango de variación también pequeño. No existen missings para esta variable.

Tabla de frecuencias	
Modalidades	Freq. absol.
23.862 - 24.7774	2
24.7774 - 25.6927	101
25.6927 - 26.6081	0
26.6081 - 27.5235	0
27.5235 - 28.4388	0
28.4388 - 29.3542	0
29.3542 - 30.2695	0
30.2695 - 31.1849	0
31.1849 - 32.1003	1
32.1003 - 33.0156	0
33.0156 - 33.931	1
33.931 - 34.8464	1
34.8464 - 35.7617	0
35.7617 - 36.6771	21
36.6771 - 37.5924	2
37.5924 - 38.5078	4
38.5078 - 39.4232	0
39.4232 - 40.3385	0
40.3385 - 41.2539	81
41.2539 - 42.1693	1
42.1693 - 43.0846	0
43.0846 - 44	150
Missings	0

F.1.11 Variable TN-influent

La media de la concentración total de nitrógeno en la entrada es de 45.0675 mg/l con una desviación de 7.976, con unos valores que van de 0 a 83.792 mg/l.

No existen missings para esta variable. La media de los datos se mantienen constantes, aunque la variabilidad es bastante alta como se aprecia en el coeficiente de variación y el el time series plot de la Figura F.21.

En el histograma, que se muestra en la Figura F.22, se observa que no predomina ningún grupo de valores y existen muy pocos valores que se salgan del comportamiento normal, pudiéndose ajustar una curva normal.

Existen 2 claros outliers, uno alcanzando un pico de 83,792 mg/l el 24-nov-2005 y el otro con un valor mínimo de 0,000 mg/l, el día 23-abr-2006.

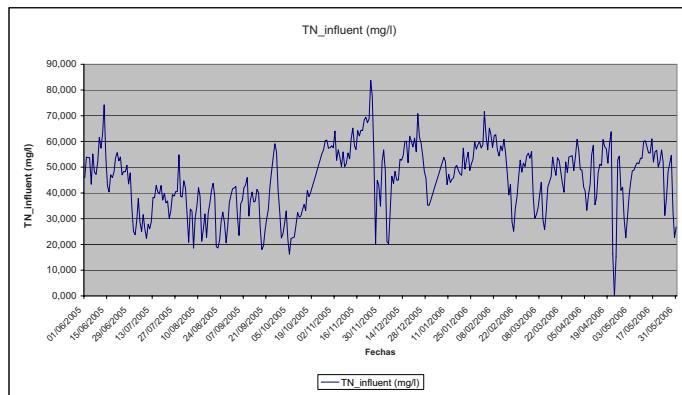


Figura F.21: Serie temporal para variable TN-influent.

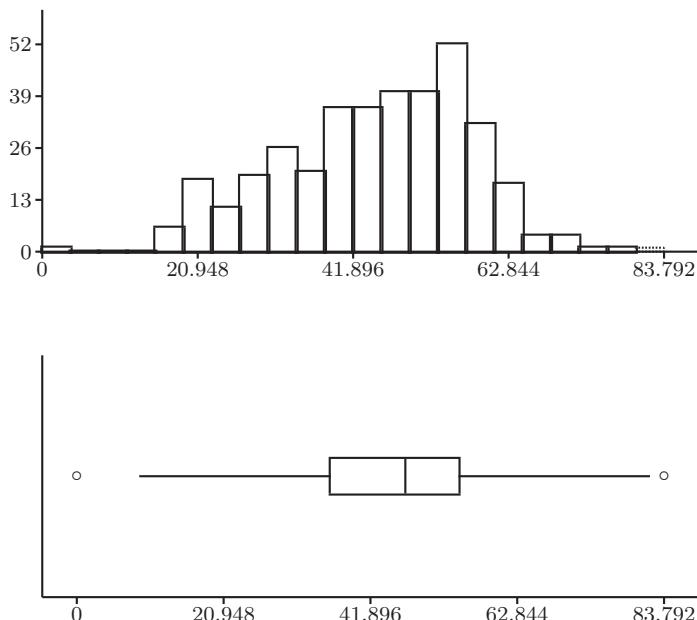


Figura F.22: Histograma y Boxplot de la variable Variable TN-influent.

Estadísticos	
Número de observaciones	365
Número de missings	0
Número de observaciones útiles	365
Media	45.0675
Mediana	46.876
Primer quartil (Q1)	36.239
Tercer quartil (Q3)	54.4575
Mínimo	0
Máximo	83.792
Quasi-desviación típica	12.81
Coeficiente de variación	0.2838

F.1.12 Variable TN-effluent

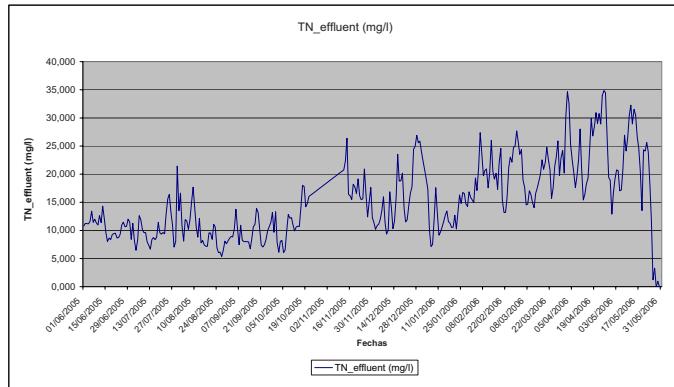


Figura F.23: Serie temporal para variable TN-effluent.

En el histograma, que se muestra en la Figura F.24, se observa que no predomina ning n grupo de valores y existen muy pocos valores que se salgan del comportamiento normal.

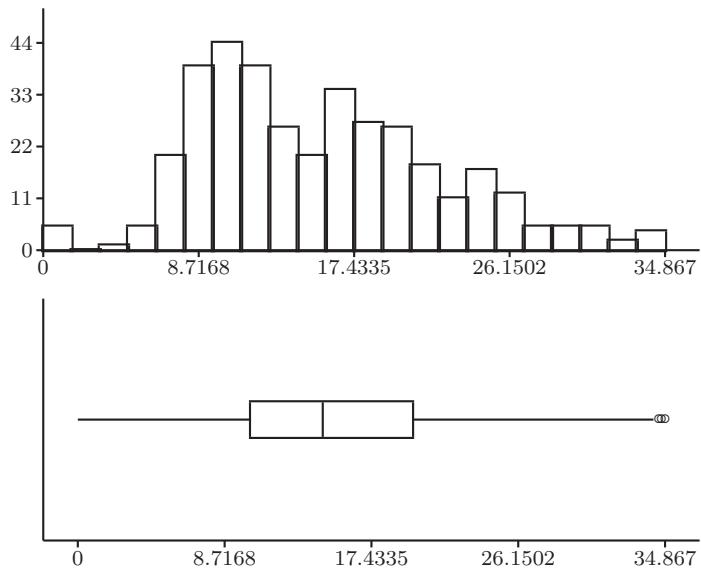


Figura F.24: Histograma y Boxplot de la variable Variable TN-effluent.

La media de la concentraci n total de nitr geno en la entrada es de 15.45 mg/l con una desviaci n de 6.73, con unos valores que van de 0 a 34.867 mg/l. No existen missings para esta variable. La media de los datos se mantienen constantes, aunque la variabilidad es bastante alta como se aprecia en el coeficiente de variaci n y el el time series plot de la Figura F.23.

Tabla de frecuencias	
Modalidades	Freq. absol.
0 - 1.5849	5
1.5849 - 3.1697	0
3.1697 - 4.7546	1
4.7546 - 6.3395	5
6.3395 - 7.9243	20
7.9243 - 9.5092	39
9.5092 - 11.094	44
11.094 - 12.6789	39
12.6789 - 14.2638	26
14.2638 - 15.8486	20
15.8486 - 17.4335	34
17.4335 - 19.0184	27
19.0184 - 20.6032	26
20.6032 - 22.1881	18
22.1881 - 23.773	11
23.773 - 25.3578	17
25.3578 - 26.9427	12
26.9427 - 28.5275	5
28.5275 - 30.1124	5
30.1124 - 31.6973	5
31.6973 - 33.2821	2
33.2821 - 34.867	4
<i>Missings</i>	0

F.1.13 Variable TOC-influent

La concentración de carbono orgánico en la entrada de la planta piloto, presenta una alta variabilidad, como se observa en el valor del coeficiente de variación y en el Time series plot de la Figura F.25.

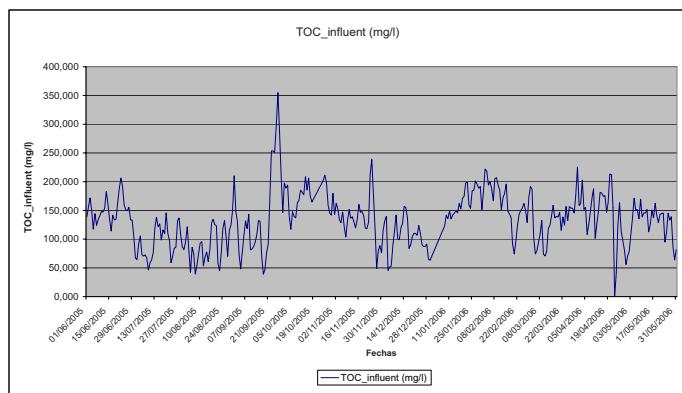


Figura F.25: Serie temporal para variable TOC-influent.

En el Histograma que se muestra en la Figura F.26 se observa claramente que su distribución es bastante simétrica y próxima a una normal. Hemos comprobado que la concentración maxima de carbono orgánico en la entrada de la planta piloto es de 3.55 miligramos

por litro.

Tenemos una media de 133.4 miligramos por litro de la concentración de carbono orgánico en la entrada de la planta piloto en por día, con una desviación de 46.96 miligramos por litro, con unos valores que van de 0 a 355 miligramos por litro.

En esta variable no se observan missing.

Se pueden considerar 2 datos atípicos, uno el día 27 de septiembre de 2005 alcanzando un valor máximo de 355 miligramos por litro y otro el día 23-abr-2006, con un valor mínimo de 0 miligramos por litro.

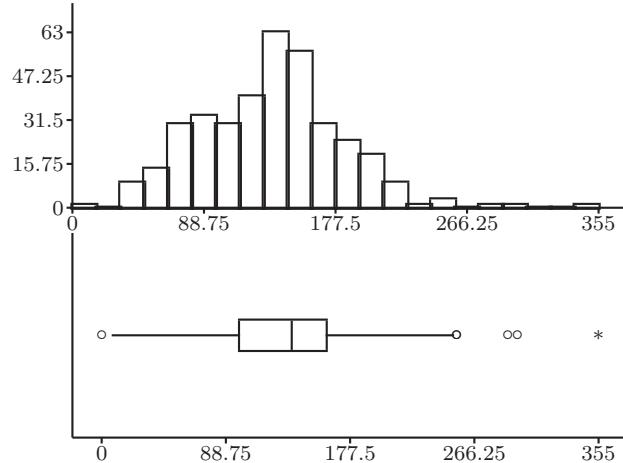


Figura F.26: Histograma y Boxplot de la variable Variable TOC-influent.

Tabla de frecuencias	
Modalidades	Freq. absolut.
0 - 16.1364	1
16.1364 - 32.2727	0
32.2727 - 48.4091	9
48.4091 - 64.5455	14
64.5455 - 80.6818	30
80.6818 - 96.8182	33
96.8182 - 112.9546	30
112.9546 - 129.0909	40
129.0909 - 145.2273	63
145.2273 - 161.3637	56
161.3637 - 177.5	30
177.5 - 193.6364	24
193.6364 - 209.7728	19
209.7728 - 225.9091	9
225.9091 - 242.0455	1
242.0455 - 258.1819	3
258.1819 - 274.3182	0
274.3182 - 290.4546	1
290.4546 - 306.5909	1
306.5909 - 322.7273	0
322.7273 - 338.8636	0
338.8636 - 355	1
Missings	
0	

F.1.14 Variable Nitritox-influent

El Nitritox-influent presenta una alta variabilidad, como se observa en el valor del coeficiente de variación y en el Time series plot de la Figura F.27.

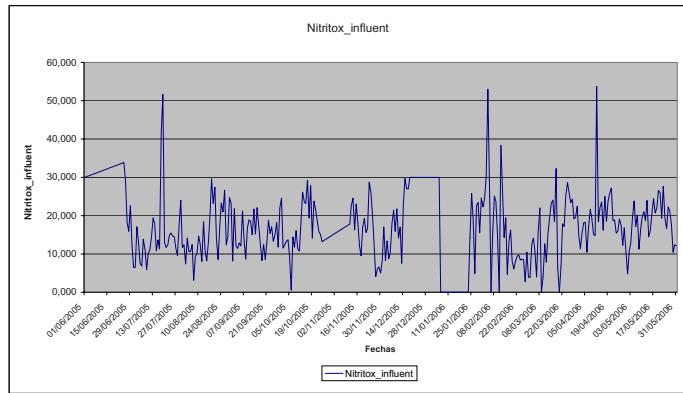


Figura F.27: Serie temporal para variable Nitritox-influent.

La media de esta variable es de 18.5219% con una desviación de 8.5024%, con unos valores que van desde el 28.604% al 69.898%.

En el Histograma, que se muestra en la Figura F.28, se observa que su distribución es bastante simétrica y próxima a una normal.

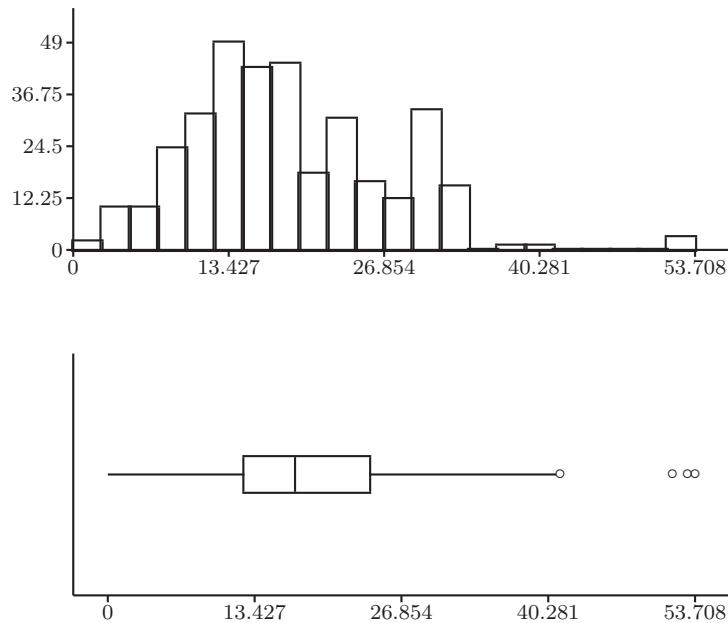


Figura F.28: Histograma y Boxplot de la variable Variable Nitritox-influent.

Esta variable presenta 21 missings, lo cual nos obliga a trabajar con 344 observaciones. Estos son:

- Observaciones comprendidas entre el 06-ene-2006 y el 23-ene-2006 (nº220 a nº237).
- Observación del día 06-dic-2006 (nº251).

- Observación del día 12-feb-2006 (nº256).
- Observación del día 09-mar-2006 (nº282).

Se pueden considerar 3 datos atípicos:

- (nº49) del día 19/07/2005 con un valor de 51,64887432%
- (nº249) del día 04/02/2006 con un valor de 53,00016022%
- (nº316) del día 12/04/2006 con un valor de 53,70812607%

Tabla de frecuencias	
Modalidades	Freq. absol.
0 - 2.4413	2
2.4413 - 4.8825	10
4.8825 - 7.3238	10
7.3238 - 9.7651	24
9.7651 - 12.2064	32
12.2064 - 14.6476	49
14.6476 - 17.0889	43
17.0889 - 19.5302	44
19.5302 - 21.9715	18
21.9715 - 24.4127	31
24.4127 - 26.854	16
26.854 - 29.2953	12
29.2953 - 31.7365	33
31.7365 - 34.1778	15
34.1778 - 36.6191	0
36.6191 - 39.0604	1
39.0604 - 41.5016	1
41.5016 - 43.9429	0
43.9429 - 46.3842	0
46.3842 - 48.8255	0
48.8255 - 51.2667	0
51.2667 - 53.708	3
<i>Missings</i>	21

Estadísticos	
Número de observaciones	365
Número de missings	21
Número de observaciones útiles	344
Media	18.5219
Mediana	17.125
Primer quartil (Q1)	12.4955
Tercer quartil (Q3)	23.8955
Mínimo	0
Màximo	53.708
Quasi-desviación típica	8.5024
Coeficiente de variación	0.4584

F.1.15 Variable TOC-effluent

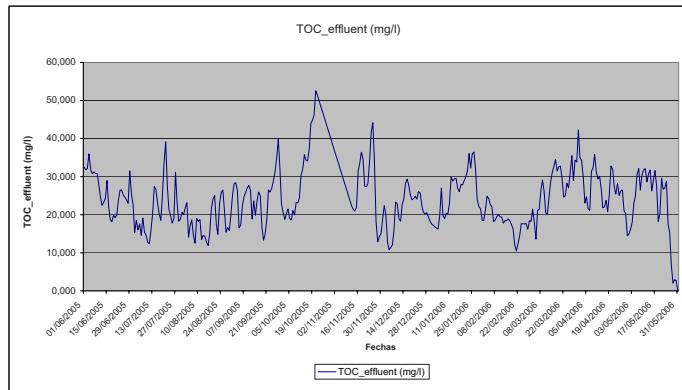


Figura F.29: Serie temporal para variable TOC-effluent.

Tenemos una media de 24.55 miligramos por litro de la concentración de carbono orgánico en la salida de la planta piloto en por día, con una desviación de 8.089 miligramos por litro, con unos valores que van de 0 a 52.57 miligramos por litro. En esta variable no se observan missing.

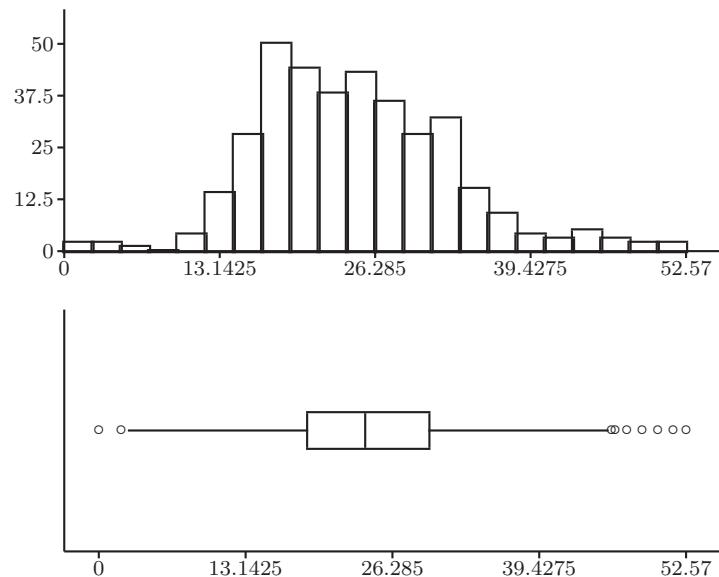


Figura F.30: Histograma y Boxplot de la variable Variable TOC-effluent.

Se pueden considerar 9 datos atípicos:

- 21-oct-2005 alcanzando el valor máximo de 52.57 miligramos por litro
- 31-may-2006 , con un valor mínimo de 0 miligramos por litro.
- 20-oct-2005, el valor 46,220 miligramos por litro.
- 22-oct-2005, el valor 51,401 miligramos por litro.
- 23-oct-2005, el valor 50,018 miligramos por litro.
- 24-oct-2005, el valor 48,635 miligramos por litro.

- 25-oct-2005, el valor 47,252 miligramos por litro.
- 26-oct-2005, el valor 45,870 miligramos por litro.
- 28-may-2006, el valor 2,014 miligramos por litro.

En el Histograma que se muestra en la Figura F.26 se observa claramente que su distribución es bastante simétrica y próxima a una normal. Hemos comprobado que la concentración maxima de carbono orgánico en la entrada de la planta piloto es de 3.55 miligramos por litro.

La concentración de carbono orgánico en la salida de la planta piloto presenta una alta variabilidad, como se observa en el valor del coeficiente de variación y en el Time series plot de la Figura F.29.

Tabla de frecuencias	
Modalidades	Freq. absol.
0 - 2.3895	2
2.3895 - 4.7791	2
4.7791 - 7.1686	1
7.1686 - 9.5582	0
9.5582 - 11.9477	4
11.9477 - 14.3373	14
14.3373 - 16.7268	28
16.7268 - 19.1164	50
19.1164 - 21.5059	44
21.5059 - 23.8955	38
23.8955 - 26.285	43
26.285 - 28.6745	36
28.6745 - 31.0641	28
31.0641 - 33.4536	32
33.4536 - 35.8432	15
35.8432 - 38.2327	9
38.2327 - 40.6223	4
40.6223 - 43.0118	3
43.0118 - 45.4014	5
45.4014 - 47.7909	3
47.7909 - 50.1805	2
50.1805 - 52.57	2
<i>Missings</i>	0

Estadísticos	
Número de observaciones	365
Número de missings	0
Número de observaciones útiles	365
Media	24.5481
Mediana	23.847
Primer quartil (Q1)	18.742
Tercer quartil (Q3)	29.5015
Mínimo	0
Máximo	52.57
Quasi-desviación típica	8.0886
Coeficiente de variación	0.3291

F.2 Análisis bivariate

1. Variables TN-effluent and NH4-2aerobic

Debido a que estas 2 variables se utilizan para construir las reglas utilizadas en la clasificación basada en reglas (((AND (\geq (NH4-2aerobic) 10.0) ($>$ (TN-effluent) 18.0)) -> Mmonia) y ((AND ($<$ (NH4-2aerobic) 10.0) ($>$ (TN-effluent) 18.0)) -> Nitrogen)) hemos realizado el gráfico se muestra en la Figura F.31 que corrobora que la concentración total de nitrógeno en la salida de la planta piloto y la concentración de amonio en el segundo tanque están bastante relacionadas, presentando un coeficiente de correlación bastante alto (0.77). Podríamos decir que existe una relación directa entre ellas a excepción de algunos casos donde la variable NH4-2aerobic toma valores bajos y la TN-effluent valores altos.

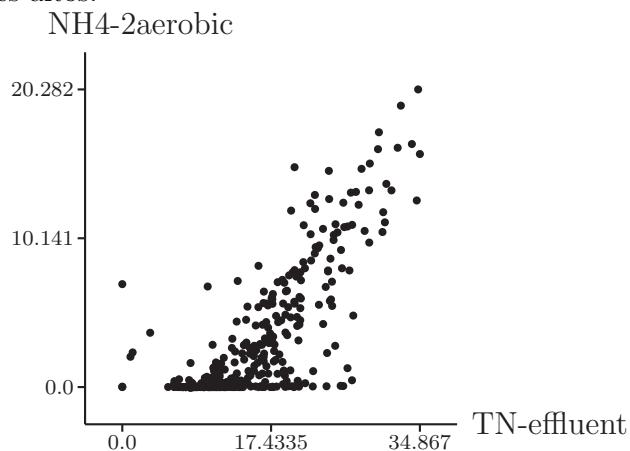


Figura F.31: Diagrama bivariante para las variables TN-effluent and NH4-2aerobic.

Correlació per les variables TN-effluent i NH4-2aerobic		
$r = 0.777$ (coVar = 22.7106)		
Informació sobre dades mancants		
TN-effluent \ NH4-2aerobic	útil	mancant
útil	365	0
mancant	0	0

2. Variables NH4-influent and NH4-2aerobic

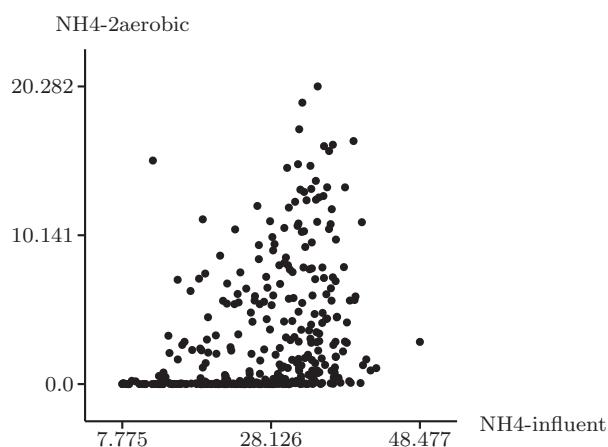


Figura F.32: Diagrama bivariante para las variables NH4-influent and NH4-2aerobic.

Correlació per les Variables NH4-influent i NH4-2aerobic		
$r = 0.37$ (coVar = 13.1024)		
Informació sobre dades mancants		
NH4-influent \ NH4-2aerobic	útil	mancant
útil	364	0
mancant	1	0

3. Variables O2-1aerobic and O2-2aerobic

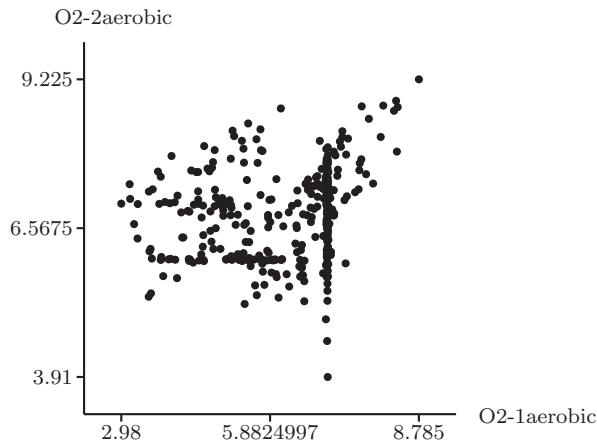


Figura F.33: Diagrama bivariante para las variables O2-1aerobic and O2-2aerobic.

Correlació per les Variables O2-1aerobic i O2-2aerobic		
$r = 0.2447$ (coVar = 0.2294)		
Informació sobre dades mancants		
O2-1aerobic \ O2-2aerobic	útil	mancant
útil	365	0
mancant	0	0

4. Variables TN-influent and TN-effluent

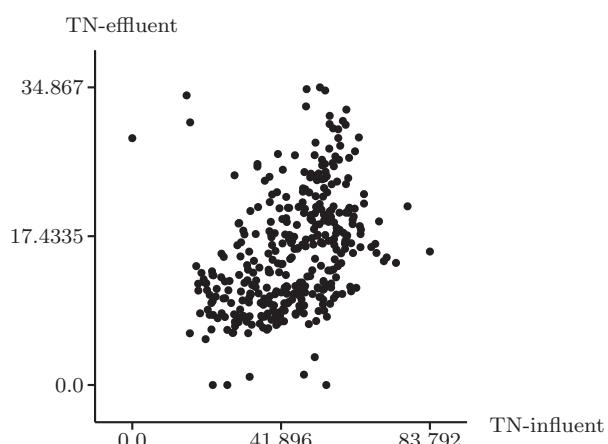


Figura F.34: Diagrama bivariante para las variables TN-influent and TN-effluent.

Correlació per les Variables TN-influent i TN-effluent		
r = 0.3814 (coVar = 32.865)		
Informació sobre dades mancants		
TN-influent \ TN-effluent	útil	mancant
útil	365	0
mancant	0	0

En estas 3 últimas variables que son las mismas pero medidas en distintos puntos, principalmente en la entrada y salida, no existe correlación; presentando valores para el coeficiente de correlación por debajo de 0.4.

Anexo G

Análisis descriptivo por clases, planta eslovena

G.1 Clustering Based on Rules (Family 3) planta eslovena

G.1.1 Class variable: 2-classes [$P2_{Lj3,R2}^{EnW,G}$]

Classe	Objectes
classer361	1-jun-05, 2-jun-05, 3-jun-05, 4-jun-05, 5-jun-05, 6-jun-05, 7-jun-05, 8-jun-05, 9-jun-05, 10-jun-05, 11-jun-05, 12-jun-05, 13-jun-05, 14-jun-05, 15-jun-05, 16-jun-05, 17-jun-05, 18-jun-05, 19-jun-05, 20-jun-05, 21-jun-05, 22-jun-05, 23-jun-05, 24-jun-05, 25-jun-05, 26-jun-05, 27-jun-05, 28-jun-05, 29-jun-05, 30-jul-05, 9-sep-05, 15-sep-05, 16-sep-05, 23-sep-05, 24-sep-05, 25-sep-05, 26-sep-05, 27-sep-05, 17-oct-05, 19-oct-05, 20-oct-05, 21-oct-05, 22-oct-05, 23-oct-05, 24-oct-05, 25-oct-05, 26-oct-05, 27-oct-05, 28-oct-05, 29-oct-05, 30-oct-05, 31-oct-05, 1-nov-05, 2-nov-05, 3-nov-05, 4-nov-05, 5-nov-05, 6-nov-05, 7-nov-05, 8-nov-05, 9-nov-05, 10-nov-05, 11-nov-05, 12-nov-05, 13-nov-05, 14-nov-05, 15-nov-05, 16-nov-05, 17-nov-05, 18-nov-05, 19-nov-05, 20-nov-05, 21-nov-05, 22-nov-05, 23-nov-05, 24-nov-05, 25-nov-05, 26-nov-05, 15-dic-05, 16-dic-05, 17-dic-05, 18-dic-05, 19-dic-05, 20-dic-05, 21-dic-05, 22-dic-05, 23-dic-05, 24-dic-05, 25-dic-05, 26-dic-05, 27-dic-05, 28-dic-05, 29-dic-05, 30-dic-05, 31-dic-05, 1-ene-06, 2-ene-06, 3-ene-06, 15-ene-06, 16-ene-06, 18-ene-06, 19-ene-06, 20-ene-06, 21-ene-06, 22-ene-06, 23-ene-06, 24-ene-06, 25-ene-06, 26-ene-06, 27-ene-06, 28-ene-06, 29-ene-06, 30-ene-06, 31-ene-06, 1-feb-06, 2-feb-06, 3-feb-06, 4-feb-06, 7-feb-06, 8-feb-06, 9-feb-06, 10-feb-06, 11-feb-06, 12-feb-06, 14-feb-06, 15-feb-06, 16-feb-06, 19-feb-06, 24-feb-06, 25-feb-06, 26-feb-06, 2-mar-06, 4-mar-06, 5-mar-06, 15-mar-06, 16-mar-06, 17-mar-06, 18-mar-06, 21-mar-06, 22-mar-06, 25-mar-06, 26-mar-06, 29-mar-06, 6-abr-06, 8-abr-06, 11-abr-06, 14-abr-06, 15-abr-06, 27-abr-06, 28-abr-06, 29-abr-06, 2-may-06, 3-may-06, 4-may-06, 5-may-06, 6-may-06, 7-may-06, 9-may-06, 13-may-06, 16-may-06, 17-may-06, 18-may-06, 19-may-06, 20-may-06, 21-may-06, 22-may-06, 23-may-06, 24-may-06, 25-may-06, 26-may-06, 27-may-06, 28-may-06
classer362	30-jun-05, 1-jul-05, 2-jul-05, 3-jul-05, 4-jul-05, 5-jul-05, 6-jul-05, 7-jul-05, 8-jul-05, 9-jul-05, 10-jul-05, 11-jul-05, 12-jul-05, 13-jul-05, 14-jul-05, 15-jul-05, 16-jul-05, 17-jul-05, 18-jul-05, 19-jul-05, 20-jul-05, 21-jul-05, 22-jul-05, 23-jul-05, 24-jul-05, 25-jul-05, 26-jul-05, 27-jul-05, 28-jul-05, 29-jul-05, 31-jul-05, 1-agosto-05, 2-agosto-05, 3-agosto-05, 4-agosto-05, 5-agosto-05, 6-agosto-05, 7-agosto-05, 8-agosto-05, 9-agosto-05, 10-agosto-05, 11-agosto-05, 12-agosto-05, 13-agosto-05, 14-agosto-05, 15-agosto-05, 16-agosto-05, 17-agosto-05, 18-agosto-05, 19-agosto-05, 20-agosto-05, 21-agosto-05, 22-agosto-05, 23-agosto-05, 24-agosto-05, 25-agosto-05, 26-agosto-05, 27-agosto-05, 28-agosto-05, 29-agosto-05, 30-agosto-05, 31-agosto-05, 1-septiembre-05, 2-septiembre-05, 3-septiembre-05, 4-septiembre-05, 5-septiembre-05, 6-septiembre-05, 7-septiembre-05, 8-septiembre-05, 10-septiembre-05, 11-septiembre-05, 12-septiembre-05, 13-septiembre-05, 14-septiembre-05, 17-septiembre-05, 18-septiembre-05, 19-septiembre-05, 20-septiembre-05, 21-septiembre-05, 22-septiembre-05, 28-septiembre-05, 29-septiembre-05, 30-septiembre-05, 1-octubre-05, 2-octubre-05, 3-octubre-05, 4-octubre-05, 5-octubre-05, 6-octubre-05, 7-octubre-05, 8-octubre-05, 9-octubre-05, 10-octubre-05, 11-octubre-05, 12-octubre-05, 13-octubre-05, 14-octubre-05, 15-octubre-05, 16-octubre-05, 18-octubre-05, 27-noviembre-05, 28-noviembre-05, 29-noviembre-05, 30-noviembre-05, 1-diciembre-05, 2-diciembre-05, 3-diciembre-05, 4-diciembre-05, 5-diciembre-05, 6-diciembre-05, 7-diciembre-05, 8-diciembre-05, 10-diciembre-05, 11-diciembre-05, 12-diciembre-05, 13-diciembre-05, 14-diciembre-05, 4-ene-06, 5-ene-06, 6-ene-06, 7-ene-06, 8-ene-06, 9-ene-06, 10-ene-06, 11-ene-06, 12-ene-06, 13-ene-06, 14-ene-06, 17-ene-06, 5-febrero-06, 6-febrero-06, 13-febrero-06, 17-febrero-06, 18-febrero-06, 20-febrero-06, 21-febrero-06, 22-febrero-06, 23-febrero-06, 27-febrero-06, 28-febrero-06, 1-marzo-06, 3-marzo-06, 6-marzo-06, 7-marzo-06, 8-marzo-06, 9-marzo-06, 10-marzo-06, 11-marzo-06, 12-marzo-06, 13-marzo-06, 14-marzo-06, 19-marzo-06, 20-marzo-06, 23-marzo-06, 24-marzo-06, 27-marzo-06, 28-marzo-06, 30-marzo-06, 31-marzo-06, 1-abril-06, 2-abril-06, 3-abril-06, 4-abril-06, 5-abril-06, 7-abril-06, 9-abril-06, 10-abril-06, 12-abril-06, 13-abril-06, 16-abril-06, 17-abril-06, 18-abril-06, 19-abril-06, 20-abril-06, 21-abril-06, 22-abril-06, 23-abril-06, 24-abril-06, 25-abril-06, 26-abril-06, 30-abril-06, 1-mayo-06, 8-mayo-06, 10-mayo-06, 11-mayo-06, 12-mayo-06, 14-mayo-06, 15-mayo-06, 29-mayo-06, 30-mayo-06, 31-mayo-06

	CLASSE	classer361	classer362
VARIABLE	N = 365	$n_c = 172$	$n_c = 193$
NH4-influent	\bar{X}	31.5278	22.408
	S	5.6473	7.622
	min	7.972	7.775
	max	48.477	40.541
	N*	1	0
NH4-2aerobic	\bar{X}	2.9458	3.2118
	S	2.9868	5.2739
	min	0	0
	max	9.846	20.282
	N*	0	0
O2-1aerobic	\bar{X}	6.3347	5.931
	S	0.9668	1.3145
	min	3.297	2.98
	max	8.785	8.371
	N*	0	0
O2-2aerobic	\bar{X}	6.5467	7.0434
	S	0.7665	0.7469
	min	3.91	4.94
	max	9.225	8.842
	N*	0	0
Valve-air	\bar{X}	46.0302	36.9181
	S	8.5624	6.2026
	min	28.934	28.604
	max	69.898	54.777
	N*	0	0
Q-air	\bar{X}	1586.6061	1185.4628
	S	366.3921	318.4674
	min	739.819	697.829
	max	2201.27	2030.77
	N*	0	0
h-ww	\bar{X}	3.0246	3.0088
	S	0.04	0.0163
	min	2.66	2.978
	max	3.098	3.058
	N*	0	0
Q-influent	\bar{X}	65.0873	69.4812
	S	11.4917	8.9902
	min	49.706	50.99
	max	85.092	85.5
	N*	0	0
FR1-DOTOK	\bar{X}	48.0964	48.815
	S	2.5447	2.2709
	min	42.276	39.153
	max	50.7	50.733
	N*	0	0

	CLASSE	classer361	classer362
VARIABLE	N = 365	$n_c = 172$	$n_c = 193$
Freq-rec	\bar{X}	36.6861	37.5108
	S	7.4469	8.4201
	min	23.862	23.863
	max	44	43.97
	N*	0	0
TN-influent	\bar{X}	52.4629	38.4768
	S	8.9897	12.1023
	min	28.792	0
	max	83.792	65.25
	N*	0	0
TN-effluent	\bar{X}	16.8673	14.194
	S	5.3788	7.5263
	min	0	0
	max	28.933	34.867
	N*	0	0
Temp-ww	\bar{X}	15.9545	16.516
	S	3.4783	4.0217
	min	8.217	8.472
	max	22.583	21.896
	N*	0	0
TOC-influent	\bar{X}	150.0303	118.5904
	S	42.3268	46.0193
	min	63.22	0
	max	355	290.212
	N*	0	0
Nitritox-influent	\bar{X}	22.0272	15.3653
	S	7.9498	7.7284
	min	3.833	0
	max	53	53.708
	N*	9	12
TOC-effluent	\bar{X}	27.2546	22.1361
	S	8.2862	7.1047
	min	2.014	0
	max	52.57	44.053
	N*	0	0

G.1.2 Class variable: 3-classes $[P3_{Lj3,R2}^{EnW,G}]$

Classe	Objectes
classer357	1-jun-05, 2-jun-05, 3-jun-05, 4-jun-05, 5-jun-05, 6-jun-05, 7-jun-05, 8-jun-05, 9-jun-05, 10-jun-05, 11-jun-05, 12-jun-05, 13-jun-05, 14-jun-05, 15-jun-05, 16-jun-05, 17-jun-05, 18-jun-05, 19-jun-05, 20-jun-05, 21-jun-05, 22-jun-05, 23-jun-05, 15-dic-05, 20-dic-05, 21-dic-05, 22-dic-05, 23-dic-05, 24-dic-05, 25-dic-05, 15-ene-06, 16-ene-06, 18-ene-06, 19-ene-06, 20-ene-06, 21-ene-06, 22-ene-06, 23-ene-06, 24-ene-06, 25-ene-06, 26-ene-06, 27-ene-06, 28-ene-06, 29-ene-06, 30-ene-06, 31-ene-06, 1-feb-06, 2-feb-06, 4-feb-06, 11-feb-06
classer353	24-jun-05, 25-jun-05, 26-jun-05, 27-jun-05, 28-jun-05, 29-jun-05, 30-jul-05, 9-sep-05, 15-sep-05, 16-sep-05, 23-sep-05, 24-sep-05, 25-sep-05, 26-sep-05, 27-sep-05, 17-oct-05, 19-oct-05, 20-oct-05, 21-oct-05, 22-oct-05, 23-oct-05, 24-oct-05, 25-oct-05, 26-oct-05, 27-oct-05, 28-oct-05, 29-oct-05, 30-oct-05, 31-oct-05, 1-nov-05, 2-nov-05, 3-nov-05, 4-nov-05, 5-nov-05, 6-nov-05, 7-nov-05, 8-nov-05, 9-nov-05, 10-nov-05, 11-nov-05, 12-nov-05, 13-nov-05, 14-nov-05, 15-nov-05, 16-nov-05, 17-nov-05, 18-nov-05, 19-nov-05, 20-nov-05, 21-nov-05, 22-nov-05, 23-nov-05, 24-nov-05, 25-nov-05, 26-nov-05, 16-dic-05, 17-dic-05, 18-dic-05, 19-dic-05, 26-dic-05, 27-dic-05, 28-dic-05, 29-dic-05, 30-dic-05, 31-dic-05, 1-ene-06, 2-ene-06, 3-ene-06, 3-feb-06, 7-feb-06, 8-feb-06, 9-feb-06, 10-feb-06, 12-feb-06, 14-feb-06, 15-feb-06, 16-feb-06, 19-feb-06, 24-feb-06, 25-feb-06, 26-feb-06, 2-mar-06, 4-mar-06, 5-mar-06, 15-mar-06, 16-mar-06, 17-mar-06, 18-mar-06, 21-mar-06, 22-mar-06, 25-mar-06, 26-mar-06, 29-mar-06, 6-abr-06, 8-abr-06, 11-abr-06, 14-abr-06, 15-abr-06, 27-abr-06, 28-abr-06, 29-abr-06, 2-may-06, 3-may-06, 4-may-06, 5-may-06, 6-may-06, 7-may-06, 9-may-06, 13-may-06, 16-may-06, 17-may-06, 18-may-06, 19-may-06, 20-may-06, 21-may-06, 22-may-06, 23-may-06, 24-may-06, 25-may-06, 26-may-06, 27-may-06, 28-may-06
classer362	30-jun-05, 1-jul-05, 2-jul-05, 3-jul-05, 4-jul-05, 5-jul-05, 6-jul-05, 7-jul-05, 8-jul-05, 9-jul-05, 10-jul-05, 11-jul-05, 12-jul-05, 13-jul-05, 14-jul-05, 15-jul-05, 16-jul-05, 17-jul-05, 18-jul-05, 19-jul-05, 20-jul-05, 21-jul-05, 22-jul-05, 23-jul-05, 24-jul-05, 25-jul-05, 26-jul-05, 27-jul-05, 28-jul-05, 29-jul-05, 31-jul-05, 1-ago-05, 2-ago-05, 3-ago-05, 4-ago-05, 5-ago-05, 6-ago-05, 7-ago-05, 8-ago-05, 9-ago-05, 10-ago-05, 11-ago-05, 12-ago-05, 13-ago-05, 14-ago-05, 15-ago-05, 16-ago-05, 17-ago-05, 18-ago-05, 19-ago-05, 20-ago-05, 21-ago-05, 22-ago-05, 23-ago-05, 24-ago-05, 25-ago-05, 26-ago-05, 27-ago-05, 28-ago-05, 29-ago-05, 30-ago-05, 31-ago-05, 1-sep-05, 2-sep-05, 3-sep-05, 4-sep-05, 5-sep-05, 6-sep-05, 7-sep-05, 8-sep-05, 10-sep-05, 11-sep-05, 12-sep-05, 13-sep-05, 14-sep-05, 17-sep-05, 18-sep-05, 19-sep-05, 20-sep-05, 21-sep-05, 22-sep-05, 28-sep-05, 29-sep-05, 30-sep-05, 1-oct-05, 2-oct-05, 3-oct-05, 4-oct-05, 5-oct-05, 6-oct-05, 7-oct-05, 8-oct-05, 9-oct-05, 10-oct-05, 11-oct-05, 12-oct-05, 13-oct-05, 14-oct-05, 15-oct-05, 16-oct-05, 18-oct-05, 27-nov-05, 28-nov-05, 29-nov-05, 30-nov-05, 1-dic-05, 2-dic-05, 3-dic-05, 4-dic-05, 5-dic-05, 6-dic-05, 7-dic-05, 8-dic-05, 9-dic-05, 10-dic-05, 11-dic-05, 12-dic-05, 13-dic-05, 14-dic-05, 4-ene-06, 5-ene-06, 6-ene-06, 7-ene-06, 8-ene-06, 9-ene-06, 10-ene-06, 11-ene-06, 12-ene-06, 13-ene-06, 14-ene-06, 17-ene-06, 5-feb-06, 6-feb-06, 13-feb-06, 17-feb-06, 18-feb-06, 20-feb-06, 21-feb-06, 22-feb-06, 23-feb-06, 27-feb-06, 28-feb-06, 1-mar-06, 3-mar-06, 6-mar-06, 7-mar-06, 8-mar-06, 9-mar-06, 10-mar-06, 11-mar-06, 12-mar-06, 13-mar-06, 14-mar-06, 19-mar-06, 20-mar-06, 23-mar-06, 24-mar-06, 27-mar-06, 28-mar-06, 30-mar-06, 31-mar-06, 1-abr-06, 2-abr-06, 3-abr-06, 4-abr-06, 5-abr-06, 7-abr-06, 9-abr-06, 10-abr-06, 12-abr-06, 13-abr-06, 16-abr-06, 17-abr-06, 18-abr-06, 19-abr-06, 20-abr-06, 21-abr-06, 22-abr-06, 23-abr-06, 24-abr-06, 25-abr-06, 26-abr-06, 30-abr-06, 1-may-06, 8-may-06, 10-may-06, 11-may-06, 12-may-06, 14-may-06, 15-may-06, 29-may-06, 30-may-06, 31-may-06

	CLASSE	classer357	classer353	classer362
VARIABLE	N = 365	n _c = 50	n _c = 122	n _c = 193
NH4-influent	̄X	32.2268	31.2389	22.408
	S	4.5997	6.0211	7.622
	min	23.23	7.972	7.775
	max	48.477	42.511	40.541
	N*	0	1	0
NH4-2aerobic	̄X	1.2776	3.6295	3.2118
	S	1.6932	3.1345	5.2739
	min	0	0	0
	max	8.262	9.846	20.282
	N*	0	0	0
O2-1aerobic	̄X	6.1719	6.4014	5.931
	S	1.0857	0.91	1.3145
	min	3.75	3.297	2.98
	max	7.018	8.785	8.371
	N*	0	0	0
O2-2aerobic	̄X	6.3738	6.6176	7.0434
	S	0.574	0.8243	0.7469
	min	5.675	3.91	4.94
	max	7.714	9.225	8.842
	N*	0	0	0
Valve-air	̄X	42.2756	47.569	36.9181
	S	6.2214	8.9287	6.2026
	min	32.199	28.934	28.604
	max	57.442	69.898	54.777
	N*	0	0	0
Q-air	̄X	1414.6736	1657.0702	1185.4628
	S	240.8154	386.0172	318.4674
	min	962.514	739.819	697.829
	max	2062.554	2201.27	2030.77
	N*	0	0	0
h-ww	̄X	3.0125	3.0296	3.0088
	S	0.0308	0.0424	0.0163
	min	2.831	2.66	2.978
	max	3.055	3.098	3.058
	N*	0	0	0
Q-influent	̄X	52.8412	70.1061	69.4812
	S	1.39	9.9238	8.9902
	min	49.706	51.123	50.99
	max	55.666	85.092	85.5
	N*	0	0	0

	CLASSE	classer357	classer353	classer362
VARIABLE	N = 365	$n_c = 50$	$n_c = 122$	$n_c = 193$
FR1-DOTOK	\bar{X}	45.3106	49.2381	48.815
	S	1.0132	2.0529	2.2709
	min	42.276	44.167	39.153
	max	47.265	50.7	50.733
	N*	0	0	0
Freq-rec	\bar{X}	38.5903	35.9056	37.5108
	S	2.2435	8.6152	8.4201
	min	33.778	23.862	23.863
	max	40.633	44	43.97
	N*	0	0	0
TN-influent	\bar{X}	54.5931	51.5898	38.4768
	S	7.0002	9.5771	12.1023
	min	40.417	28.792	0
	max	74.249	83.792	65.25
	N*	0	0	0
TN-effluent	\bar{X}	12.6796	18.5835	14.194
	S	2.7506	5.2542	7.5263
	min	8.032	0	0
	max	17.837	28.933	34.867
	N*	0	0	0
Temp-ww	\bar{X}	15.8939	15.9793	16.516
	S	3.5328	3.47	4.0217
	min	11.235	8.217	8.472
	max	21.034	22.583	21.896
	N*	0	0	0
TOC-influent	\bar{X}	157.4615	146.9847	118.5904
	S	30.9108	45.965	46.0193
	min	89.833	63.22	0
	max	222.043	355	290.212
	N*	0	0	0
Nitritox-influent	\bar{X}	29.0309	19.6734	15.3653
	S	7.1261	6.7543	7.7284
	min	4.875	3.833	0
	max	53	38.333	53.708
	N*	9	0	12
TOC-effluent	\bar{X}	26.5675	27.5362	22.1361
	S	5.1903	9.2655	7.1047
	min	18.153	2.014	0
	max	36.476	52.57	44.053
	N*	0	0	0

G.1.3 Class variable: 4-classes $[P4_{Lj3,R2}^{EnW,G}]$

Classe	Objectes
classer357	1-jun-05, 2-jun-05, 3-jun-05, 4-jun-05, 5-jun-05, 6-jun-05, 7-jun-05, 8-jun-05, 9-jun-05, 10-jun-05, 11-jun-05, 12-jun-05, 13-jun-05, 14-jun-05, 15-jun-05, 16-jun-05, 17-jun-05, 18-jun-05, 19-jun-05, 20-jun-05, 21-jun-05, 22-jun-05, 23-jun-05, 15-dic-05, 20-dic-05, 21-dic-05, 22-dic-05, 23-dic-05, 24-dic-05, 25-dic-05, 15-ene-06, 16-ene-06, 18-ene-06, 19-ene-06, 20-ene-06, 21-ene-06, 22-ene-06, 23-ene-06, 24-ene-06, 25-ene-06, 26-ene-06, 27-ene-06, 28-ene-06, 29-ene-06, 30-ene-06, 31-ene-06, 1-feb-06, 2-feb-06, 4-feb-06, 11-feb-06
classer353	24-jun-05, 25-jun-05, 26-jun-05, 27-jun-05, 28-jun-05, 29-jun-05, 30-jul-05, 9-sep-05, 15-sep-05, 16-sep-05, 23-sep-05, 24-sep-05, 25-sep-05, 26-sep-05, 27-sep-05, 17-oct-05, 19-oct-05, 20-oct-05, 21-oct-05, 22-oct-05, 23-oct-05, 24-oct-05, 25-oct-05, 26-oct-05, 27-oct-05, 28-oct-05, 29-oct-05, 30-oct-05, 31-oct-05, 1-nov-05, 2-nov-05, 3-nov-05, 4-nov-05, 5-nov-05, 6-nov-05, 7-nov-05, 8-nov-05, 9-nov-05, 10-nov-05, 11-nov-05, 12-nov-05, 13-nov-05, 14-nov-05, 15-nov-05, 16-nov-05, 17-nov-05, 18-nov-05, 19-nov-05, 20-nov-05, 21-nov-05, 22-nov-05, 23-nov-05, 24-nov-05, 25-nov-05, 26-nov-05, 16-dic-05, 17-dic-05, 18-dic-05, 19-dic-05, 26-dic-05, 27-dic-05, 28-dic-05, 29-dic-05, 30-dic-05, 31-dic-05, 1-ene-06, 2-ene-06, 3-ene-06, 3-feb-06, 7-feb-06, 8-feb-06, 9-feb-06, 10-feb-06, 12-feb-06, 14-feb-06, 15-feb-06, 16-feb-06, 19-feb-06, 24-feb-06, 25-feb-06, 26-feb-06, 2-mar-06, 4-mar-06, 5-mar-06, 15-mar-06, 16-mar-06, 17-mar-06, 18-mar-06, 21-mar-06, 22-mar-06, 25-mar-06, 26-mar-06, 29-mar-06, 6-abr-06, 8-abr-06, 11-abr-06, 14-abr-06, 15-abr-06, 27-abr-06, 28-abr-06, 29-abr-06, 2-may-06, 3-may-06, 4-may-06, 5-may-06, 6-may-06, 7-may-06, 9-may-06, 13-may-06, 16-may-06, 17-may-06, 18-may-06, 19-may-06, 20-may-06, 21-may-06, 22-may-06, 23-may-06, 24-may-06, 25-may-06, 26-may-06, 27-may-06, 28-may-06
classer360	30-jun-05, 1-jul-05, 2-jul-05, 3-jul-05, 4-jul-05, 5-jul-05, 6-jul-05, 7-jul-05, 8-jul-05, 9-jul-05, 10-jul-05, 11-jul-05, 12-jul-05, 13-jul-05, 14-jul-05, 15-jul-05, 16-jul-05, 17-jul-05, 18-jul-05, 19-jul-05, 20-jul-05, 21-jul-05, 22-jul-05, 23-jul-05, 24-jul-05, 25-jul-05, 26-jul-05, 27-jul-05, 28-jul-05, 29-jul-05, 31-jul-05, 1-agosto-05, 2-agosto-05, 3-agosto-05, 4-agosto-05, 5-agosto-05, 6-agosto-05, 7-agosto-05, 8-agosto-05, 9-agosto-05, 10-agosto-05, 11-agosto-05, 12-agosto-05, 13-agosto-05, 14-agosto-05, 15-agosto-05, 16-agosto-05, 19-agosto-05, 20-agosto-05, 21-agosto-05, 22-agosto-05, 23-agosto-05, 24-agosto-05, 25-agosto-05, 26-agosto-05, 27-agosto-05, 28-agosto-05, 29-agosto-05, 30-agosto-05, 31-agosto-05, 1-sept-05, 2-sept-05, 3-sept-05, 4-sept-05, 5-sept-05, 6-sept-05, 7-sept-05, 8-sept-05, 10-sept-05, 11-sept-05, 12-sept-05, 13-sept-05, 14-sept-05, 17-sept-05, 18-sept-05, 19-sept-05, 20-sept-05, 21-sept-05, 22-sept-05, 28-sept-05, 29-sept-05, 30-sept-05, 1-oct-05, 2-oct-05, 3-oct-05, 4-oct-05, 5-oct-05, 6-oct-05, 7-oct-05, 8-oct-05, 9-oct-05, 10-oct-05, 11-oct-05, 12-oct-05, 13-oct-05, 14-oct-05, 15-oct-05, 16-oct-05, 18-oct-05, 12-abr-06
classer358	17-agosto-05, 18-agosto-05, 27-noviembre-05, 28-noviembre-05, 29-noviembre-05, 30-noviembre-05, 1-diciembre-05, 2-diciembre-05, 3-diciembre-05, 4-diciembre-05, 5-diciembre-05, 6-diciembre-05, 7-diciembre-05, 8-diciembre-05, 9-diciembre-05, 10-diciembre-05, 11-diciembre-05, 12-diciembre-05, 13-diciembre-05, 14-diciembre-05, 4-ene-06, 5-ene-06, 6-ene-06, 7-ene-06, 8-ene-06, 9-ene-06, 10-ene-06, 11-ene-06, 12-ene-06, 13-ene-06, 14-ene-06, 17-ene-06, 5-febrero-06, 6-febrero-06, 13-febrero-06, 17-febrero-06, 18-febrero-06, 20-febrero-06, 21-febrero-06, 22-febrero-06, 23-febrero-06, 27-febrero-06, 28-febrero-06, 1-marzo-06, 3-marzo-06, 6-marzo-06, 7-marzo-06, 8-marzo-06, 9-marzo-06, 10-marzo-06, 11-marzo-06, 12-marzo-06, 13-marzo-06, 14-marzo-06, 19-marzo-06, 20-marzo-06, 23-marzo-06, 24-marzo-06, 27-marzo-06, 28-marzo-06, 30-marzo-06, 31-marzo-06, 1-abril-06, 2-abril-06, 3-abril-06, 4-abril-06, 5-abril-06, 7-abril-06, 9-abril-06, 10-abril-06, 13-abril-06, 16-abril-06, 17-abril-06, 18-abril-06, 19-abril-06, 20-abril-06, 21-abril-06, 22-abril-06, 23-abril-06, 24-abril-06, 25-abril-06, 26-abril-06, 30-abril-06, 1-mayo-06, 8-mayo-06, 10-mayo-06, 11-mayo-06, 12-mayo-06, 14-mayo-06, 15-mayo-06, 16-mayo-06, 17-mayo-06, 29-mayo-06, 30-mayo-06, 31-mayo-06

	CLASSE	classer357	classer353	classer360	classer358
VARIABLE	N = 365	$n_c = 50$	$n_c = 122$	$n_c = 100$	$n_c = 93$
NH4-influent	\bar{X}	32.2268	31.2389	19.3756	25.6687
	S	4.5997	6.0211	5.977	7.8776
	min	23.23	7.972	7.775	7.79
	max	48.477	42.511	34.353	40.541
	N*	0	1	0	0
NH4-2aerobic	\bar{X}	1.2776	3.6295	0.224	6.4244
	S	1.6932	3.1345	0.5572	6.1299
	min	0	0	0	0
	max	8.262	9.846	2.856	20.282
	N*	0	0	0	0
O2-1aerobic	\bar{X}	6.1719	6.4014	4.8544	7.0887
	S	1.0857	0.91	0.8104	0.5336
	min	3.75	3.297	2.98	4.479
	max	7.018	8.785	6.998	8.371
	N*	0	0	0	0
O2-2aerobic	\bar{X}	6.3738	6.6176	6.8538	7.2473
	S	0.574	0.8243	0.7327	0.711
	min	5.675	3.91	5.889	4.94
	max	7.714	9.225	8.705	8.842
	N*	0	0	0	0
Valve-air	\bar{X}	42.2756	47.569	35.5443	38.3953
	S	6.2214	8.9287	5.3644	6.7131
	min	32.199	28.934	28.604	29.57
	max	57.442	69.898	51.168	54.777
	N*	0	0	0	0
Q-air	\bar{X}	1414.6736	1657.0702	1142.6132	1231.5381
	S	240.8154	386.0172	273.2748	356.5925
	min	962.514	739.819	746.07	697.829
	max	2062.554	2201.27	1770.8521	2030.77
	N*	0	0	0	0
h-ww	\bar{X}	3.0125	3.0296	3.0132	3.0041
	S	0.0308	0.0424	0.017	0.0142
	min	2.831	2.66	2.984	2.978
	max	3.055	3.098	3.058	3.039
	N*	0	0	0	0
Q-influent	\bar{X}	52.8412	70.1061	70.4546	68.4345
	S	1.39	9.9238	2.5696	12.6269
	min	49.706	51.123	63.038	50.99
	max	55.666	85.092	83.013	85.5
	N*	0	0	0	0
FR1-DOTOK	\bar{X}	45.3106	49.2381	49.5442	48.0309
	S	1.0132	2.0529	1.0593	2.8903
	min	42.276	44.167	47.2	39.153
	max	47.265	50.7	50.733	50.692
	N*	0	0	0	0

	CLASSE	classer357	classer353	classer360	classer358
VARIABLE	N = 365	$n_c = 50$	$n_c = 122$	$n_c = 100$	$n_c = 93$
Freq-rec		38.5903	35.9056	43.5834	30.9811
	S	2.2435	8.6152	1.8874	7.8066
	min	33.778	23.862	24.899	23.863
	max	40.633	44	43.851	43.97
	N*	0	0	0	0
TN-influent		54.5931	51.5898	32.3876	45.0243
	S	7.0002	9.5771	7.7482	12.5469
	min	40.417	28.792	16.209	0
	max	74.249	83.792	54.792	65.25
	N*	0	0	0	0
TN-effluent		12.6796	18.5835	10.0487	18.6514
	S	2.7506	5.2542	2.7683	8.4385
	min	8.032	0	5.371	0
	max	17.837	28.933	17.788	34.867
	N*	0	0	0	0
Temp-ww		15.8939	15.9793	19.9831	12.7879
	S	3.5328	3.47	1.238	2.2337
	min	11.235	8.217	13.327	8.472
	max	21.034	22.583	21.896	20.928
	N*	0	0	0	0
TOC-influent		157.4615	146.9847	110.1023	127.7172
	S	30.9108	45.965	47.6237	42.6218
	min	89.833	63.22	38.888	0
	max	222.043	355	290.212	225.293
	N*	0	0	0	0
Nitritox-influent		29.0309	19.6734	15.326	15.4138
	S	7.1261	6.7543	8.2305	7.1101
	min	4.875	3.833	0.542	0
	max	53	38.333	53.708	30
	N*	9	0	0	12
TOC-effluent		26.5675	27.5362	21.7552	22.5457
	S	5.1903	9.2655	6.4925	7.7236
	min	18.153	2.014	11.879	0
	max	36.476	52.57	44.053	42.251
	N*	0	0	0	0

Anexo H

BbD, BbIR y $\mathcal{R}_e(\mathcal{P}_\xi^*)$ planta eslovena

H.1 BbD para $\mathcal{P}_2^* \equiv P2_{Lj3,R2}^{EnW,G}$

Patrón: centro cerrado para la variable NH4-influent de la partición \mathcal{P}_2^*

$$I_1^{NH4-influent,2} = [7.775, 7.972)$$

$$I_2^{NH4-influent,2} = [7.972, 40.541]$$

$$I_3^{NH4-influent,2} = (40.541, 48.477]$$

Patrón: centro cerrado para la variable NH4-2aerobic de la partición \mathcal{P}_2^*

$$I_1^{NH4-2aerobic,2} = [0.0, 0.0)$$

$$I_2^{NH4-2aerobic,2} = [0.0, 9.846]$$

$$I_3^{NH4-2aerobic,2} = (9.846, 20.282]$$

Patrón: centro cerrado para la variable O2-1aerobic de la partición \mathcal{P}_2^*

$$I_1^{O2-1aerobic,2} = [2.98, 3.297)$$

$$I_2^{O2-1aerobic,2} = [3.297, 8.371]$$

$$I_3^{O2-1aerobic,2} = (8.371, 8.785]$$

Patrón: centro cerrado para la variable O2-2aerobic de la partición \mathcal{P}_2^*

$$I_1^{O2-2aerobic,2} = [3.91, 4.94)$$

$$I_2^{O2-2aerobic,2} = [4.94, 8.842]$$

$$I_3^{O2-2aerobic,2} = (8.842, 9.225]$$

Patrón: centro cerrado para la variable Valve-air de la partición \mathcal{P}_2^*

$$I_1^{Valve-air,2} = [28.604, 28.934)$$

$$I_2^{Valve-air,2} = [28.934, 54.777]$$

$$I_3^{Valve-air,2} = (54.777, 69.898]$$

Patrón: centro cerrado para la variable Q-air de la partición \mathcal{P}_2^*

$$I_1^{Q-air,2} = [697.829, 739.819)$$

$$I_2^{Q-air,2} = [739.819, 2030.77]$$

$$I_3^{Q-air,2} = (2030.77, 2201.27]$$

Patrón: centro cerrado para la variable h-ww de la partición \mathcal{P}_2^*

$$I_1^{h-ww,2} = [2.66, 2.978)$$

$$I_2^{h-ww,2} = [2.978, 3.058]$$

$$I_3^{h-ww,2} = (3.058, 3.098]$$

Patrón: centro cerrado para la variable Q-influent de la partición \mathcal{P}_2^*

$$I_1^{Q-influent,2} = [49.706, 50.99)$$

$$I_2^{Q-influent,2} = [50.99, 85.092]$$

$$I_3^{Q-influent,2} = (85.092, 85.5]$$

Patrón: centro cerrado para la variable FR1-DOTOK de la partición \mathcal{P}_2^*

$$I_1^{FR1-DOTOK,2} = [39.153, 42.276)$$

$$I_2^{FR1-DOTOK,2} = [42.276, 50.7]$$

$$I_3^{FR1-DOTOK,2} = (50.7, 50.733]$$

Patrón: centro cerrado para la variable Freq-rec de la partición \mathcal{P}_2^*

$$I_1^{Freq-rec,2} = [23.862, 23.863)$$

$$I_2^{Freq-rec,2} = [23.863, 43.97]$$

$$I_3^{Freq-rec,2} = (43.97, 44.0]$$

Patrón: centro cerrado para la variable TN-influent de la partición \mathcal{P}_2^*

$$I_1^{TN-influent,2} = [0.0, 28.792)$$

$$I_2^{TN-influent,2} = [28.792, 65.25]$$

$$I_3^{TN-influent,2} = (65.25, 83.792]$$

Patrón: centro cerrado para la variable TN-effluent de la partición \mathcal{P}_2^*

$$I_1^{TN-effluent,2} = [0.0, 0.0)$$

$$I_2^{TN-effluent,2} = [0.0, 28.933]$$

$$I_3^{TN-effluent,2} = (28.933, 34.867]$$

Patrón: centro cerrado para la variable Temp-ww de la partición \mathcal{P}_2^*

$$I_1^{Temp-ww,2} = [8.217, 8.472)$$

$$I_2^{Temp-ww,2} = [8.472, 21.896]$$

$$I_3^{Temp-ww,2} = (21.896, 22.583]$$

Patrón: centro cerrado para la variable TOC-influent de la partición \mathcal{P}_2^*

$$I_1^{TOC-influent,2} = [0.0, 63.22)$$

$$I_2^{TOC-influent,2} = [63.22, 290.212]$$

$$I_3^{TOC-influent,2} = (290.212, 355.0]$$

Patrón: centro cerrado para la variable Nitritox-influent de la partición \mathcal{P}_2^*

$$I_1^{Nitritox-influent,2} = [0.0, 3.833)$$

$$I_2^{Nitritox-influent,2} = [3.833, 53.0]$$

$$I_3^{Nitritox-influent,2} = (53.0, 53.708]$$

Patrón: centro cerrado para la variable TOC-effluent de la partición \mathcal{P}_2^*

$$I_1^{TOC-effluent,2} = [0.0, 2.014)$$

$$I_2^{TOC-effluen,2} = [2.014, 44.053]$$

$$I_3^{TOC-effluen,2} = (44.053, 52.57]$$

H.2 $\mathcal{R}(\mathcal{P}_2^*)$

$$\begin{aligned}\mathcal{R}(\mathcal{P}_2^*) = \{ & r_{1, classer361}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.972] \xrightarrow{0.0} classer361, \\ & r_{2, classer361}^{NH4-influent} : x_{NH4-influent,i} \in [7.972, 40.541] \xrightarrow{0.4663} classer361, \\ & r_{3, classer361}^{NH4-influent} : x_{NH4-influent,i} \in (40.541, 48.477] \xrightarrow{1.0} classer361, \\ & r_{1, classer361}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.972) \xrightarrow{0.0} classer361, \\ & r_{2, classer361}^{NH4-influent} : x_{NH4-influent,i} \in [7.972, 40.541] \xrightarrow{0.4663} classer361, \\ & r_{3, classer361}^{NH4-influent} : x_{NH4-influent,i} \in (40.541, 48.477] \xrightarrow{1.0} classer361, \\ & r_{2, classer361}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in [0.0, 9.846] \xrightarrow{0.526} classer361, \\ & r_{3, classer361}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (9.846, 20.282] \xrightarrow{0.0} classer361, \\ & r_{1, classer361}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 3.297] \xrightarrow{0.0} classer361, \\ & r_{2, classer361}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [3.297, 8.371] \xrightarrow{0.475} classer361, \\ & r_{3, classer361}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (8.371, 8.785] \xrightarrow{1.0} classer361, \\ & r_{1, classer361}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 4.94] \xrightarrow{1.0} classer361, \\ & r_{2, classer361}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [4.94, 8.842] \xrightarrow{0.4669} classer361, \\ & r_{3, classer361}^{O2-2aerobic} : x_{O2-2aerobic,i} \in (8.842, 9.225] \xrightarrow{1.0} classer361, \\ & r_{1, classer361}^{Valve-air} : x_{Valve-air,i} \in [28.604, 28.934] \xrightarrow{0.0} classer361, \\ & r_{2, classer361}^{Valve-air} : x_{Valve-air,i} \in [28.934, 54.777] \xrightarrow{0.4337} classer361, \\ & r_{3, classer361}^{Valve-air} : x_{Valve-air,i} \in (54.777, 69.898] \xrightarrow{1.0} classer361, \\ & r_{1, classer361}^{Q-air} : x_{Q-air,i} \in [697.829, 739.819] \xrightarrow{0.0} classer361, \\ & r_{2, classer361}^{Q-air} : x_{Q-air,i} \in [739.819, 2030.77] \xrightarrow{0.4315} classer361, \\ & r_{3, classer361}^{Q-air} : x_{Q-air,i} \in (2030.77, 2201.27] \xrightarrow{1.0} classer361, \\ & r_{1, classer361}^{h-ww} : x_{h-ww,i} \in [2.66, 2.978] \xrightarrow{1.0} classer361, \\ & r_{2, classer361}^{h-ww} : x_{h-ww,i} \in [2.978, 3.058] \xrightarrow{0.4438} classer361, \\ & r_{3, classer361}^{h-ww} : x_{h-ww,i} \in (3.058, 3.098] \xrightarrow{1.0} classer361, \\ & r_{1, classer361}^{Q-influent} : x_{Q-influent,i} \in [49.706, 50.99] \xrightarrow{1.0} classer361, \\ & r_{2, classer361}^{Q-influent} : x_{Q-influent,i} \in [50.99, 85.092] \xrightarrow{0.4637} classer361, \\ & r_{3, classer361}^{Q-influent} : x_{Q-influent,i} \in (85.092, 85.5] \xrightarrow{0.0} classer361,\end{aligned}$$

$$\begin{aligned}
r_{1, \text{classer361}}^{FR1-DOTOK} &: x_{FR1-DOTOK,i} \in [39.153, 42.276] \xrightarrow{0.0} \text{classer361}, \\
r_{2, \text{classer361}}^{FR1-DOTOK} &: x_{FR1-DOTOK,i} \in [42.276, 50.7] \xrightarrow{0.4791} \text{classer361}, \\
r_{3, \text{classer361}}^{FR1-DOTOK} &: x_{FR1-DOTOK,i} \in (50.7, 50.733] \xrightarrow{0.0} \text{classer361}, \\
r_{1, \text{classer361}}^{Freq-rec} &: x_{Freq-rec,i} \in [23.862, 23.863) \xrightarrow{1.0} \text{classer361}, \\
r_{2, \text{classer361}}^{Freq-rec} &: x_{Freq-rec,i} \in [23.863, 43.97] \xrightarrow{0.4654} \text{classer361}, \\
r_{3, \text{classer361}}^{Freq-rec} &: x_{Freq-rec,i} \in (43.97, 44.0] \xrightarrow{1.0} \text{classer361}, \\
r_{1, \text{classer361}}^{TN-influent} &: x_{TN-influent,i} \in [0.0, 28.792) \xrightarrow{0.0} \text{classer361}, \\
r_{2, \text{classer361}}^{TN-influent} &: x_{TN-influent,i} \in [28.792, 65.25] \xrightarrow{0.5224} \text{classer361}, \\
r_{3, \text{classer361}}^{TN-influent} &: x_{TN-influent,i} \in (65.25, 83.792] \xrightarrow{1.0} \text{classer361}, \\
r_{2, \text{classer361}}^{TN-effluent} &: x_{TN-effluent,i} \in [0.0, 28.933] \xrightarrow{0.0} \text{classer361}, \\
r_{3, \text{classer361}}^{TN-effluent} &: x_{TN-effluent,i} \in (28.933, 34.867] \xrightarrow{0.0} \text{classer361}, \\
r_{1, \text{classer361}}^{Temp-ww} &: x_{Temp-ww,i} \in [8.217, 8.472) \xrightarrow{1.0} \text{classer361}, \\
r_{2, \text{classer361}}^{Temp-ww} &: x_{Temp-ww,i} \in [8.472, 21.896] \xrightarrow{0.4624} \text{classer361}, \\
r_{3, \text{classer361}}^{Temp-ww} &: x_{Temp-ww,i} \in (21.896, 22.583] \xrightarrow{1.0} \text{classer361}, \\
r_{1, \text{classer361}}^{TOC-influent} &: x_{TOC-influent,i} \in [0.0, 63.22] \xrightarrow{0.0} \text{classer361}, \\
r_{2, \text{classer361}}^{TOC-influent} &: x_{TOC-influent,i} \in [63.22, 290.212] \xrightarrow{0.4956} \text{classer361}, \\
r_{3, \text{classer361}}^{TOC-influent} &: x_{TOC-influent,i} \in (290.212, 355.0] \xrightarrow{1.0} \text{classer361}, \\
r_{1, \text{classer361}}^{Nitritox-influent} &: x_{Nitritox-influent,i} \in [0.0, 3.833) \xrightarrow{0.0} \text{classer361}, \\
r_{2, \text{classer361}}^{Nitritox-influent} &: x_{Nitritox-influent,i} \in [3.833, 53.0] \xrightarrow{0.4808} \text{classer361}, \\
r_{3, \text{classer361}}^{Nitritox-influent} &: x_{Nitritox-influent,i} \in (53.0, 53.708] \xrightarrow{0.0} \text{classer361}, \\
r_{1, \text{classer361}}^{TOC-effluent} &: x_{TOC-effluent,i} \in [0.0, 2.014) \xrightarrow{0.0} \text{classer361}, \\
r_{2, \text{classer361}}^{TOC-effluent} &: x_{TOC-effluent,i} \in [2.014, 44.053] \xrightarrow{0.4576} \text{classer361}, \\
r_{3, \text{classer361}}^{TOC-effluent} &: x_{TOC-effluent,i} \in (44.053, 52.57] \xrightarrow{1.0} \text{classer361}, \\
r_{1, \text{classer362}}^{NH4-influent} &: x_{NH4-influent,i} \in [7.775, 7.972) \xrightarrow{1.0} \text{classer362}, \\
r_{2, \text{classer362}}^{NH4-influent} &: x_{NH4-influent,i} \in [7.972, 40.541] \xrightarrow{0.5337} \text{classer362}, \\
r_{3, \text{classer362}}^{NH4-influent} &: x_{NH4-influent,i} \in (40.541, 48.477] \xrightarrow{0.0} \text{classer362}, \\
r_{1, \text{classer362}}^{NH4-influent} &: x_{NH4-influent,i} \in [7.775, 7.972) \xrightarrow{1.0} \text{classer362}, \\
r_{2, \text{classer362}}^{NH4-influent} &: x_{NH4-influent,i} \in [7.972, 40.541] \xrightarrow{0.5337} \text{classer362}, \\
r_{3, \text{classer362}}^{NH4-influent} &: x_{NH4-influent,i} \in (40.541, 48.477] \xrightarrow{0.0} \text{classer362}, \\
r_{2, \text{classer362}}^{NH4-2aerobic} &: x_{NH4-2aerobic,i} \in [0.0, 9.846] \xrightarrow{0.474} \text{classer362}, \\
r_{3, \text{classer362}}^{NH4-2aerobic} &: x_{NH4-2aerobic,i} \in (9.846, 20.282] \xrightarrow{1.0} \text{classer362}, \\
r_{1, \text{classer362}}^{O2-1aerobic} &: x_{O2-1aerobic,i} \in [2.98, 3.297) \xrightarrow{1.0} \text{classer362}, \\
r_{2, \text{classer362}}^{O2-1aerobic} &: x_{O2-1aerobic,i} \in [3.297, 8.371] \xrightarrow{0.525} \text{classer362}, \\
r_{3, \text{classer362}}^{O2-1aerobic} &: x_{O2-1aerobic,i} \in (8.371, 8.785] \xrightarrow{0.0} \text{classer362}, \\
r_{1, \text{classer362}}^{O2-2aerobic} &: x_{O2-2aerobic,i} \in [3.91, 4.94) \xrightarrow{0.0} \text{classer362}, \\
r_{2, \text{classer362}}^{O2-2aerobic} &: x_{O2-2aerobic,i} \in [4.94, 8.842] \xrightarrow{0.5331} \text{classer362},
\end{aligned}$$

$$\begin{aligned}
r_{3,\text{classer362}}^{O2-2aerobic} : & x_{O2-2aerobic,i} \in (8.842, 9.225] \xrightarrow{0.0} \text{classer362}, \\
r_{1,\text{classer362}}^{\text{Valve-air}} : & x_{\text{Valve-air},i} \in [28.604, 28.934) \xrightarrow{1.0} \text{classer362}, \\
r_{2,\text{classer362}}^{\text{Valve-air}} : & x_{\text{Valve-air},i} \in [28.934, 54.777] \xrightarrow{0.5663} \text{classer362}, \\
r_{3,\text{classer362}}^{\text{Valve-air}} : & x_{\text{Valve-air},i} \in (54.777, 69.898] \xrightarrow{0.0} \text{classer362}, \\
r_{1,\text{classer362}}^{Q-air} : & x_{Q-air,i} \in [697.829, 739.819) \xrightarrow{1.0} \text{classer362}, \\
r_{2,\text{classer362}}^{Q-air} : & x_{Q-air,i} \in [739.819, 2030.77] \xrightarrow{0.5685} \text{classer362}, \\
r_{3,\text{classer362}}^{Q-air} : & x_{Q-air,i} \in (2030.77, 2201.27] \xrightarrow{0.0} \text{classer362}, \\
r_{1,\text{classer362}}^{h-ww} : & x_{h-ww,i} \in [2.66, 2.978) \xrightarrow{0.0} \text{classer362}, \\
r_{2,\text{classer362}}^{h-ww} : & x_{h-ww,i} \in [2.978, 3.058] \xrightarrow{0.5562} \text{classer362}, \\
r_{3,\text{classer362}}^{h-ww} : & x_{h-ww,i} \in (3.058, 3.098] \xrightarrow{0.0} \text{classer362}, \\
r_{1,\text{classer362}}^{Q-influent} : & x_{Q-influent,i} \in [49.706, 50.99] \xrightarrow{0.0} \text{classer362}, \\
r_{2,\text{classer362}}^{Q-influent} : & x_{Q-influent,i} \in [50.99, 85.092] \xrightarrow{0.5363} \text{classer362}, \\
r_{3,\text{classer362}}^{Q-influent} : & x_{Q-influent,i} \in (85.092, 85.5] \xrightarrow{1.0} \text{classer362}, \\
r_{1,\text{classer362}}^{FR1-DOTOK} : & x_{FR1-DOTOK,i} \in [39.153, 42.276) \xrightarrow{1.0} \text{classer362}, \\
r_{2,\text{classer362}}^{FR1-DOTOK} : & x_{FR1-DOTOK,i} \in [42.276, 50.7] \xrightarrow{0.5209} \text{classer362}, \\
r_{3,\text{classer362}}^{FR1-DOTOK} : & x_{FR1-DOTOK,i} \in (50.7, 50.733] \xrightarrow{1.0} \text{classer362}, \\
r_{1,\text{classer362}}^{Freq-rec} : & x_{Freq-rec,i} \in [23.862, 23.863) \xrightarrow{0.0} \text{classer362}, \\
r_{2,\text{classer362}}^{Freq-rec} : & x_{Freq-rec,i} \in [23.863, 43.97] \xrightarrow{0.5346} \text{classer362}, \\
r_{3,\text{classer362}}^{Freq-rec} : & x_{Freq-rec,i} \in (43.97, 44.0] \xrightarrow{0.0} \text{classer362}, \\
r_{1,\text{classer362}}^{TN-influent} : & x_{TN-influent,i} \in [0.0, 28.792) \xrightarrow{1.0} \text{classer362}, \\
r_{2,\text{classer362}}^{TN-influent} : & x_{TN-influent,i} \in [28.792, 65.25] \xrightarrow{0.4776} \text{classer362}, \\
r_{3,\text{classer362}}^{TN-influent} : & x_{TN-influent,i} \in (65.25, 83.792] \xrightarrow{0.0} \text{classer362}, \\
r_{2,\text{classer362}}^{TN-effluent} : & x_{TN-effluent,i} \in [0.0, 28.933] \xrightarrow{1.0} \text{classer362}, \\
r_{3,\text{classer362}}^{TN-effluent} : & x_{TN-effluent,i} \in (28.933, 34.867] \xrightarrow{1.0} \text{classer362}, \\
r_{1,\text{classer362}}^{Temp-ww} : & x_{Temp-ww,i} \in [8.217, 8.472) \xrightarrow{0.0} \text{classer362}, \\
r_{2,\text{classer362}}^{Temp-ww} : & x_{Temp-ww,i} \in [8.472, 21.896] \xrightarrow{0.5376} \text{classer362}, \\
r_{3,\text{classer362}}^{Temp-ww} : & x_{Temp-ww,i} \in (21.896, 22.583] \xrightarrow{0.0} \text{classer362}, \\
r_{1,\text{classer362}}^{TOC-influent} : & x_{TOC-influent,i} \in [0.0, 63.22) \xrightarrow{1.0} \text{classer362}, \\
r_{2,\text{classer362}}^{TOC-influent} : & x_{TOC-influent,i} \in [63.22, 290.212] \xrightarrow{0.5044} \text{classer362}, \\
r_{3,\text{classer362}}^{TOC-influent} : & x_{TOC-influent,i} \in (290.212, 355.0] \xrightarrow{0.0} \text{classer362}, \\
r_{1,\text{classer362}}^{Nitritox-influent} : & x_{Nitritox-influent,i} \in [0.0, 3.833) \xrightarrow{1.0} \text{classer362}, \\
r_{2,\text{classer362}}^{Nitritox-influent} : & x_{Nitritox-influent,i} \in [3.833, 53.0] \xrightarrow{0.5192} \text{classer362}, \\
r_{3,\text{classer362}}^{Nitritox-influent} : & x_{Nitritox-influent,i} \in (53.0, 53.708] \xrightarrow{1.0} \text{classer362}, \\
r_{1,\text{classer362}}^{TOC-effluent} : & x_{TOC-effluent,i} \in [0.0, 2.014) \xrightarrow{1.0} \text{classer362}, \\
r_{2,\text{classer362}}^{TOC-effluent} : & x_{TOC-effluent,i} \in [2.014, 44.053] \xrightarrow{0.5424} \text{classer362}, \\
r_{3,\text{classer362}}^{TOC-effluent} : & x_{TOC-effluent,i} \in (44.053, 52.57] \xrightarrow{0.0} \text{classer362} \quad \}
\end{aligned}$$

H.3 BbD para $\mathcal{P}_3^* \subseteq P3_{Lj3,R2}^{ENW,G}$

Patrón: centro cerrado para la variable NH4-influent de la partición \mathcal{P}_3^*

$$I_1^{NH4-influent,3} = [7.972, 23.23)$$

$$I_2^{NH4-influent,3} = [23.23, 42.511]$$

$$I_3^{NH4-influent,3} = (42.511, 48.477]$$

Patrón: centro cerrado para la variable NH4-2aerobic de la partición en \mathcal{P}_3^*

$$I_1^{NH4-2aerobic,3} = [0.0, 0.0)$$

$$I_2^{NH4-2aerobic,3} = [0.0, 8.262]$$

$$I_3^{NH4-2aerobic,3} = (8.262, 9.846]$$

Patrón: centro cerrado para la variable O2-1aerobic de la partición en \mathcal{P}_3^*

$$I_1^{O2-1aerobic,3} = [3.297, 3.75)$$

$$I_2^{O2-1aerobic,3} = [3.75, 7.018]$$

$$I_3^{O2-1aerobic,3} = (7.018, 8.785]$$

Patrón: centro cerrado para la variable O2-2aerobic de la partición en \mathcal{P}_3^*

$$I_1^{O2-2aerobic,3} = [3.91, 5.675)$$

$$I_2^{O2-2aerobic,3} = [5.675, 7.714]$$

$$I_3^{O2-2aerobic,3} = (7.714, 9.225]$$

Patrón: centro cerrado para la variable Valve-air de la partición en \mathcal{P}_3^*

$$I_1^{Valve-air,3} = [28.934, 32.199)$$

$$I_2^{Valve-air,3} = [32.199, 57.442]$$

$$I_3^{Valve-air,3} = (57.442, 69.898]$$

Patrón: centro cerrado para la variable Q-air de la partición en \mathcal{P}_3^*

$$I_1^{Q-air,3} = [739.819, 962.514)$$

$$I_2^{Q-air,3} = [962.514, 2062.554]$$

$$I_3^{Q-air,3} = (2062.554, 2201.27]$$

Patrón: centro cerrado para la variable h-ww de la partición en \mathcal{P}_3^*

$$I_1^{h-ww,3} = [2.66, 2.831)$$

$$I_2^{h-ww,3} = [2.831, 3.055]$$

$$I_3^{h-ww,3} = (3.055, 3.098]$$

Patrón: centro cerrado para la variable Q-influent de la partición en \mathcal{P}_3^*

$$I_1^{Q-influent,3} = [49.706, 51.123)$$

$$I_2^{Q-influent,3} = [51.123, 55.666]$$

$$I_3^{Q-influent,3} = (55.666, 85.092]$$

Patrón: centro cerrado para la variable FR1-DOTOK de la partición en \mathcal{P}_3^*

$$I_1^{FR1-DOTOK,3} = [42.276, 44.167)$$

$$I_2^{FR1-DOTOK,3} = [44.167, 47.265]$$

$$I_3^{FR1-DOTOK,3} = (47.265, 50.7]$$

Patrón: centro cerrado para la variable Freq-rec de la partición en \mathcal{P}_3^*

$$I_1^{Freq-rec,3} = [23.862, 33.778]$$

$$I_2^{Freq-rec,3} = [33.778, 40.633]$$

$$I_3^{Freq-rec,3} = (40.633, 44.0]$$

Patrón: centro cerrado para la variable TN-influent de la partición en \mathcal{P}_3^*

$$I_1^{TN-influent,3} = [28.792, 40.417)$$

$$I_2^{TN-influent,3} = [40.417, 74.249]$$

$$I_3^{TN-influent,3} = (74.249, 83.792]$$

Patrón: centro cerrado para la variable TN-effluent de la partición en \mathcal{P}_3^*

$$I_1^{TN-effluent,3} = [0.0, 8.032)$$

$$I_2^{TN-effluent,3} = [8.032, 17.837]$$

$$I_3^{TN-effluent,3} = (17.837, 28.933]$$

Patrón: centro cerrado para la variable Temp-ww de la partición en \mathcal{P}_3^*

$$I_1^{Temp-ww,3} = [8.217, 11.235)$$

$$I_2^{Temp-ww,3} = [11.235, 21.034]$$

$$I_3^{Temp-ww,3} = (21.034, 22.583]$$

Patrón: centro cerrado para la variable TOC-influent de la partición en \mathcal{P}_3^*

$$I_1^{TOC-influent,3} = [63.22, 89.833)$$

$$I_2^{TOC-influent,3} = [89.833, 222.043]$$

$$I_3^{TOC-influent,3} = (222.043, 355.0]$$

Patrón: centro cerrado para la variable Nitritox-influent de la partición en \mathcal{P}_3^*

$$I_1^{Nitritox-influent,3} = [3.833, 4.875)$$

$$I_2^{Nitritox-influent,3} = [4.875, 38.333]$$

$$I_3^{Nitritox-influent,3} = (38.333, 53.0]$$

Patrón: centro cerrado para la variable TOC-effluent de la partición en \mathcal{P}_3^*

$$I_1^{TOC-effluent,3} = [2.014, 18.153)$$

$$I_2^{TOC-effluent,3} = [18.153, 36.476]$$

$$I_3^{TOC-effluent,3} = (36.476, 52.57]$$

H.4 $\mathcal{R}(\mathcal{P}_3^*)$

$$\begin{aligned}\mathcal{R}(\mathcal{P}_3^*) = \{ & r_{1, \text{classer353}}^{NH4-influent} : x_{NH4-influent,i} \in [7.972, 23.23] \xrightarrow{1.0} \text{classer353}, \\ & r_{2, \text{classer353}}^{NH4-influent} : x_{NH4-influent,i} \in [23.23, 42.511] \xrightarrow{0.6918} \text{classer353}, \\ & r_{3, \text{classer353}}^{NH4-influent} : x_{NH4-influent,i} \in (42.511, 48.477] \xrightarrow{0.0} \text{classer353}, \\ & r_{2, \text{classer353}}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in [0.0, 8.262] \xrightarrow{0.6933} \text{classer353}, \\ & r_{3, \text{classer353}}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (8.262, 9.846] \xrightarrow{1.0} \text{classer353}, \\ & r_{1, \text{classer353}}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [3.297, 3.75] \xrightarrow{1.0} \text{classer353}, \\ & r_{2, \text{classer353}}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [3.75, 7.018] \xrightarrow{0.6875} \text{classer353}, \\ & r_{3, \text{classer353}}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (7.018, 8.785] \xrightarrow{1.0} \text{classer353}, \\ & r_{1, \text{classer353}}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [3.91, 5.675] \xrightarrow{1.0} \text{classer353}, \\ & r_{2, \text{classer353}}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [5.675, 7.714] \xrightarrow{0.6644} \text{classer353}, \\ & r_{3, \text{classer353}}^{O2-2aerobic} : x_{O2-2aerobic,i} \in (7.714, 9.225] \xrightarrow{1.0} \text{classer353}, \\ & r_{1, \text{classer353}}^{Valve-air} : x_{Valve-air,i} \in [28.934, 32.199] \xrightarrow{1.0} \text{classer353}, \\ & r_{2, \text{classer353}}^{Valve-air} : x_{Valve-air,i} \in [32.199, 57.442] \xrightarrow{0.6815} \text{classer353}, \\ & r_{3, \text{classer353}}^{Valve-air} : x_{Valve-air,i} \in (57.442, 69.898] \xrightarrow{1.0} \text{classer353}, \\ & r_{1, \text{classer353}}^{Q-air} : x_{Q-air,i} \in [739.819, 962.514] \xrightarrow{1.0} \text{classer353}, \\ & r_{2, \text{classer353}}^{Q-air} : x_{Q-air,i} \in [962.514, 2062.554] \xrightarrow{0.6528} \text{classer353}, \\ & r_{3, \text{classer353}}^{Q-air} : x_{Q-air,i} \in (2062.554, 2201.27] \xrightarrow{1.0} \text{classer353}, \\ & r_{1, \text{classer353}}^{h-ww} : x_{h-ww,i} \in [2.66, 2.831] \xrightarrow{1.0} \text{classer353}, \\ & r_{2, \text{classer353}}^{h-ww} : x_{h-ww,i} \in [2.831, 3.055] \xrightarrow{0.6689} \text{classer353}, \\ & r_{3, \text{classer353}}^{h-ww} : x_{h-ww,i} \in (3.055, 3.098] \xrightarrow{1.0} \text{classer353}, \\ & r_{1, \text{classer353}}^{Q-influent} : x_{Q-influent,i} \in [49.706, 51.123] \xrightarrow{0.0} \text{classer353}, \\ & r_{2, \text{classer353}}^{Q-influent} : x_{Q-influent,i} \in [51.123, 55.666] \xrightarrow{0.3333} \text{classer353}, \\ & r_{3, \text{classer353}}^{Q-influent} : x_{Q-influent,i} \in (55.666, 85.092] \xrightarrow{1.0} \text{classer353}, \\ & r_{1, \text{classer353}}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [42.276, 44.167] \xrightarrow{0.0} \text{classer353}, \\ & r_{2, \text{classer353}}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [44.167, 47.265] \xrightarrow{0.3636} \text{classer353}, \\ & r_{3, \text{classer353}}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (47.265, 50.7] \xrightarrow{1.0} \text{classer353},\end{aligned}$$

$$\begin{aligned}
r_{1, \text{classer353}}^{Freq-rec} &: x_{Freq-rec,i} \in [23.862, 33.778] \xrightarrow{1.0} \text{classer353}, \\
r_{2, \text{classer353}}^{Freq-rec} &: x_{Freq-rec,i} \in [33.778, 40.633] \xrightarrow{0.359} \text{classer353}, \\
r_{3, \text{classer353}}^{Freq-rec} &: x_{Freq-rec,i} \in (40.633, 44.0] \xrightarrow{1.0} \text{classer353}, \\
r_{1, \text{classer353}}^{TN-influent} &: x_{TN-influent,i} \in [28.792, 40.417] \xrightarrow{1.0} \text{classer353}, \\
r_{2, \text{classer353}}^{TN-influent} &: x_{TN-influent,i} \in [40.417, 74.249] \xrightarrow{0.6753} \text{classer353}, \\
r_{3, \text{classer353}}^{TN-influent} &: x_{TN-influent,i} \in (74.249, 83.792] \xrightarrow{1.0} \text{classer353}, \\
r_{1, \text{classer353}}^{TN-effluent} &: x_{TN-effluent,i} \in [0.0, 8.032] \xrightarrow{1.0} \text{classer353}, \\
r_{2, \text{classer353}}^{TN-effluent} &: x_{TN-effluent,i} \in [8.032, 17.837] \xrightarrow{0.4253} \text{classer353}, \\
r_{3, \text{classer353}}^{TN-effluent} &: x_{TN-effluent,i} \in (17.837, 28.933] \xrightarrow{1.0} \text{classer353}, \\
r_{1, \text{classer353}}^{Temp-ww} &: x_{Temp-ww,i} \in [8.217, 11.235] \xrightarrow{1.0} \text{classer353}, \\
r_{2, \text{classer353}}^{Temp-ww} &: x_{Temp-ww,i} \in [11.235, 21.034] \xrightarrow{0.6689} \text{classer353}, \\
r_{3, \text{classer353}}^{Temp-ww} &: x_{Temp-ww,i} \in (21.034, 22.583] \xrightarrow{1.0} \text{classer353}, \\
r_{1, \text{classer353}}^{TOC-influent} &: x_{TOC-influent,i} \in [63.22, 89.833] \xrightarrow{1.0} \text{classer353}, \\
r_{2, \text{classer353}}^{TOC-influent} &: x_{TOC-influent,i} \in [89.833, 222.043] \xrightarrow{0.6732} \text{classer353}, \\
r_{3, \text{classer353}}^{TOC-influent} &: x_{TOC-influent,i} \in (222.043, 355.0] \xrightarrow{1.0} \text{classer353}, \\
r_{1, \text{classer353}}^{Nitritox-influent} &: x_{Nitritox-influent,i} \in [3.833, 4.875] \xrightarrow{1.0} \text{classer353}, \\
r_{2, \text{classer353}}^{Nitritox-influent} &: x_{Nitritox-influent,i} \in [4.875, 38.333] \xrightarrow{0.75} \text{classer353}, \\
r_{3, \text{classer353}}^{Nitritox-influent} &: x_{Nitritox-influent,i} \in (38.333, 53.0] \xrightarrow{0.0} \text{classer353}, \\
r_{1, \text{classer353}}^{TOC-effluent} &: x_{TOC-effluent,i} \in [2.014, 18.153] \xrightarrow{1.0} \text{classer353}, \\
r_{2, \text{classer353}}^{TOC-effluent} &: x_{TOC-effluent,i} \in [18.153, 36.476] \xrightarrow{0.6403} \text{classer353}, \\
r_{3, \text{classer353}}^{TOC-effluent} &: x_{TOC-effluent,i} \in (36.476, 52.57] \xrightarrow{1.0} \text{classer353}, \\
r_{1, \text{classer357}}^{NH4-influent} &: x_{NH4-influent,i} \in [7.972, 23.23] \xrightarrow{0.0} \text{classer357}, \\
r_{2, \text{classer357}}^{NH4-influent} &: x_{NH4-influent,i} \in [23.23, 42.511] \xrightarrow{0.3082} \text{classer357}, \\
r_{3, \text{classer357}}^{NH4-influent} &: x_{NH4-influent,i} \in (42.511, 48.477] \xrightarrow{1.0} \text{classer357}, \\
r_{2, \text{classer357}}^{NH4-2aerobic} &: x_{NH4-2aerobic,i} \in [0.0, 8.262] \xrightarrow{0.3067} \text{classer357}, \\
r_{3, \text{classer357}}^{NH4-2aerobic} &: x_{NH4-2aerobic,i} \in (8.262, 9.846] \xrightarrow{0.0} \text{classer357}, \\
r_{1, \text{classer357}}^{O2-1aerobic} &: x_{O2-1aerobic,i} \in [3.297, 3.75] \xrightarrow{0.0} \text{classer357}, \\
r_{2, \text{classer357}}^{O2-1aerobic} &: x_{O2-1aerobic,i} \in [3.75, 7.018] \xrightarrow{0.3125} \text{classer357}, \\
r_{3, \text{classer357}}^{O2-1aerobic} &: x_{O2-1aerobic,i} \in (7.018, 8.785] \xrightarrow{0.0} \text{classer357}, \\
r_{1, \text{classer357}}^{O2-2aerobic} &: x_{O2-2aerobic,i} \in [3.91, 5.675] \xrightarrow{0.0} \text{classer357}, \\
r_{2, \text{classer357}}^{O2-2aerobic} &: x_{O2-2aerobic,i} \in [5.675, 7.714] \xrightarrow{0.3356} \text{classer357}, \\
r_{3, \text{classer357}}^{O2-2aerobic} &: x_{O2-2aerobic,i} \in (7.714, 9.225] \xrightarrow{0.0} \text{classer357}, \\
r_{1, \text{classer357}}^{Valve-air} &: x_{Valve-air,i} \in [28.934, 32.199] \xrightarrow{0.0} \text{classer357}, \\
r_{2, \text{classer357}}^{Valve-air} &: x_{Valve-air,i} \in [32.199, 57.442] \xrightarrow{0.3185} \text{classer357}, \\
r_{3, \text{classer357}}^{Valve-air} &: x_{Valve-air,i} \in (57.442, 69.898] \xrightarrow{0.0} \text{classer357},
\end{aligned}$$

$$\begin{aligned}
r_{1,\text{classer357}}^{Q-\text{air}} : x_{Q-\text{air},i} \in [739.819, 962.514] &\xrightarrow{0.0} \text{classer357}, \\
r_{2,\text{classer357}}^{Q-\text{air}} : x_{Q-\text{air},i} \in [962.514, 2062.554] &\xrightarrow{0.3472} \text{classer357}, \\
r_{3,\text{classer357}}^{Q-\text{air}} : x_{Q-\text{air},i} \in (2062.554, 2201.27] &\xrightarrow{0.0} \text{classer357}, \\
r_{1,\text{classer357}}^{h-\text{ww}} : x_{h-\text{ww},i} \in [2.66, 2.831] &\xrightarrow{0.0} \text{classer357}, \\
r_{2,\text{classer357}}^{h-\text{ww}} : x_{h-\text{ww},i} \in [2.831, 3.055] &\xrightarrow{0.3311} \text{classer357}, \\
r_{3,\text{classer357}}^{h-\text{ww}} : x_{h-\text{ww},i} \in (3.055, 3.098] &\xrightarrow{0.0} \text{classer357}, \\
r_{1,\text{classer357}}^{Q-\text{influent}} : x_{Q-\text{influent},i} \in [49.706, 51.123] &\xrightarrow{1.0} \text{classer357}, \\
r_{2,\text{classer357}}^{Q-\text{influent}} : x_{Q-\text{influent},i} \in [51.123, 55.666] &\xrightarrow{0.6667} \text{classer357}, \\
r_{3,\text{classer357}}^{Q-\text{influent}} : x_{Q-\text{influent},i} \in (55.666, 85.092] &\xrightarrow{0.0} \text{classer357}, \\
r_{1,\text{classer357}}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [42.276, 44.167] &\xrightarrow{1.0} \text{classer357}, \\
r_{2,\text{classer357}}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [44.167, 47.265] &\xrightarrow{0.6364} \text{classer357}, \\
r_{3,\text{classer357}}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (47.265, 50.7] &\xrightarrow{0.0} \text{classer357}, \\
r_{1,\text{classer357}}^{Freq-rec} : x_{Freq-rec,i} \in [23.862, 33.778] &\xrightarrow{0.0} \text{classer357}, \\
r_{2,\text{classer357}}^{Freq-rec} : x_{Freq-rec,i} \in [33.778, 40.633] &\xrightarrow{0.641} \text{classer357}, \\
r_{3,\text{classer357}}^{Freq-rec} : x_{Freq-rec,i} \in (40.633, 44.0] &\xrightarrow{0.0} \text{classer357}, \\
r_{1,\text{classer357}}^{TN-influent} : x_{TN-influent,i} \in [28.792, 40.417] &\xrightarrow{0.0} \text{classer357}, \\
r_{2,\text{classer357}}^{TN-influent} : x_{TN-influent,i} \in [40.417, 74.249] &\xrightarrow{0.3247} \text{classer357}, \\
r_{3,\text{classer357}}^{TN-influent} : x_{TN-influent,i} \in (74.249, 83.792] &\xrightarrow{0.0} \text{classer357}, \\
r_{1,\text{classer357}}^{TN-effluent} : x_{TN-effluent,i} \in [0.0, 8.032] &\xrightarrow{0.0} \text{classer357}, \\
r_{2,\text{classer357}}^{TN-effluent} : x_{TN-effluent,i} \in [8.032, 17.837] &\xrightarrow{0.5747} \text{classer357}, \\
r_{3,\text{classer357}}^{TN-effluent} : x_{TN-effluent,i} \in (17.837, 28.933] &\xrightarrow{0.0} \text{classer357}, \\
r_{1,\text{classer357}}^{Temp-ww} : x_{Temp-ww,i} \in [8.217, 11.235] &\xrightarrow{0.0} \text{classer357}, \\
r_{2,\text{classer357}}^{Temp-ww} : x_{Temp-ww,i} \in [11.235, 21.034] &\xrightarrow{0.3311} \text{classer357}, \\
r_{3,\text{classer357}}^{Temp-ww} : x_{Temp-ww,i} \in (21.034, 22.583] &\xrightarrow{0.0} \text{classer357}, \\
r_{1,\text{classer357}}^{TOC-influent} : x_{TOC-influent,i} \in [63.22, 89.833] &\xrightarrow{0.0} \text{classer357}, \\
r_{2,\text{classer357}}^{TOC-influent} : x_{TOC-influent,i} \in [89.833, 222.043] &\xrightarrow{0.3268} \text{classer357}, \\
r_{3,\text{classer357}}^{TOC-influent} : x_{TOC-influent,i} \in (222.043, 355.0] &\xrightarrow{0.0} \text{classer357}, \\
r_{1,\text{classer357}}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [3.833, 4.875] &\xrightarrow{0.0} \text{classer357}, \\
r_{2,\text{classer357}}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [4.875, 38.333] &\xrightarrow{0.25} \text{classer357}, \\
r_{3,\text{classer357}}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (38.333, 53.0] &\xrightarrow{1.0} \text{classer357}, \\
r_{1,\text{classer357}}^{TOC-effluent} : x_{TOC-effluent,i} \in [2.014, 18.153] &\xrightarrow{0.0} \text{classer357}, \\
r_{2,\text{classer357}}^{TOC-effluent} : x_{TOC-effluent,i} \in [18.153, 36.476] &\xrightarrow{0.3597} \text{classer357}, \\
r_{3,\text{classer357}}^{TOC-effluent} : x_{TOC-effluent,i} \in (36.476, 52.57] &\xrightarrow{0.0} \text{classer357} \quad \}
\end{aligned}$$

H.5 BbD para $\mathcal{P}_4^* \subseteq P4_{Lj3,R2}^{EnW,G}$

Patrón: centro cerrado para la variable NH4-influent de la partición en \mathcal{P}_4^*

$$I_1^{NH4-influent,4} = [7.775, 7.79)$$

$$I_2^{NH4-influent,4} = [7.79, 34.353]$$

$$I_3^{NH4-influent,4} = (34.353, 40.541]$$

Patrón: centro cerrado para la variable NH4-2aerobic de la partición en \mathcal{P}_4^*

$$I_1^{NH4-2aerobic,4} = [0.0, 0.0)$$

$$I_2^{NH4-2aerobic,4} = [0.0, 2.856]$$

$$I_3^{NH4-2aerobic,4} = (2.856, 20.282]$$

Patrón: centro cerrado para la variable O2-1aerobic de la partición en \mathcal{P}_4^*

$$I_1^{O2-1aerobic,4} = [2.98, 4.479)$$

$$I_2^{O2-1aerobic,4} = [4.479, 6.998]$$

$$I_3^{O2-1aerobic,4} = (6.998, 8.371]$$

Patrón: centro cerrado para la variable O2-2aerobic de la partición en \mathcal{P}_4^*

$$I_1^{O2-2aerobic,4} = [4.94, 5.889)$$

$$I_2^{O2-2aerobic,4} = [5.889, 8.705]$$

$$I_3^{O2-2aerobic,4} = (8.705, 8.842]$$

Patrón: centro cerrado para la variable Valve-air de la partición en \mathcal{P}_4^*

$$I_1^{Valve-air,4} = [28.604, 29.57)$$

$$I_2^{Valve-air,4} = [29.57, 51.168]$$

$$I_3^{Valve-air,4} = (51.168, 54.777]$$

Patrón: centro cerrado para la variable Q-air de la partición en \mathcal{P}_4^*

$$I_1^{Q-air,4} = [697.829, 746.07)$$

$$I_2^{Q-air,4} = [746.07, 1770.852]$$

$$I_3^{Q-air,4} = (1770.852, 2030.77]$$

Patrón: centro cerrado para la variable h-ww de la partición en \mathcal{P}_4^*

$$I_1^{h-ww,4} = [2.978, 2.984)$$

$$I_2^{h-ww,4} = [2.984, 3.039]$$

$$I_3^{h-ww,4} = (3.039, 3.058]$$

Patrón: centro cerrado para la variable Q-influent de la partición en \mathcal{P}_4^*

$$I_1^{Q-influent,4} = [50.99, 63.038)$$

$$I_2^{Q-influent,4} = [63.038, 83.013]$$

$$I_3^{Q-influent,4} = (83.013, 85.5]$$

Patrón: centro cerrado para la variable FR1-DOTOK de la partición en \mathcal{P}_4^*

$$I_1^{FR1-DOTOK,4} = [39.153, 47.2)$$

$$I_2^{FR1-DOTOK,4} = [47.2, 50.692]$$

$$I_3^{FR1-DOTOK,4} = (50.692, 50.733]$$

Patrón: centro cerrado para la variable Freq-rec de la partición en \mathcal{P}_4^*

$$I_1^{Freq-rec,4} = [23.863, 24.899]$$

$$I_2^{Freq-rec,4} = [24.899, 43.851]$$

$$I_3^{Freq-rec,4} = (43.851, 43.97]$$

Patrón: centro cerrado para la variable TN-influent de la partición en \mathcal{P}_4^*

$$I_1^{TN-influent,4} = [0.0, 16.209)$$

$$I_2^{TN-influent,4} = [16.209, 54.792]$$

$$I_3^{TN-influent,4} = (54.792, 65.25]$$

Patrón: centro cerrado para la variable TN-effluent de la partición en \mathcal{P}_4^*

$$I_1^{TN-effluent,4} = [0.0, 5.371)$$

$$I_2^{TN-effluent,4} = [5.371, 17.788]$$

$$I_3^{TN-effluent,4} = (17.788, 34.867]$$

Patrón: centro cerrado para la variable Temp-ww de la partición en \mathcal{P}_4^*

$$I_1^{Temp-ww,4} = [8.472, 13.327)$$

$$I_2^{Temp-ww,4} = [13.327, 20.928]$$

$$I_3^{Temp-ww,4} = (20.928, 21.896]$$

Patrón: centro cerrado para la variable TOC-influent de la partición en \mathcal{P}_4^*

$$I_1^{NH4-influent,4} = [0.0, 38.888)$$

$$I_2^{NH4-influent,4} = [38.888, 225.293]$$

$$I_3^{NH4-influent,4} = (225.293, 290.212]$$

Patrón: centro cerrado para la variable Nitritox-influent de la partición en \mathcal{P}_4^*

$$I_1^{Nitritox-influent,4} = [0.0, 0.542)$$

$$I_2^{Nitritox-influent,4} = [0.542, 30.0]$$

$$I_3^{Nitritox-influent,4} = (30.0, 53.708]$$

Patrón: centro cerrado para la variable TOC-effluent de la partición en \mathcal{P}_4^*

$$I_1^{TOC-effluent,4} = [0.0, 11.879)$$

$$I_2^{TOC-effluent,4} = [11.879, 42.251]$$

$$I_3^{TOC-effluent,4} = (42.251, 44.053]$$

H.6 $\mathcal{R}(\mathcal{P}_4^*)$

$$\begin{aligned}\mathcal{R}(\mathcal{P}_4^*) = \{ & r_{1, classer358}^{NH4-influent} : x_{NH4-influent,i} \in [7.775, 7.79] \xrightarrow{0.0} classer358, \\ & r_{2, classer358}^{NH4-influent} : x_{NH4-influent,i} \in [7.79, 34.353] \xrightarrow{0.4438} classer358, \\ & r_{3, classer358}^{NH4-influent} : x_{NH4-influent,i} \in (34.353, 40.541] \xrightarrow{1.0} classer358, \\ & r_{2, classer358}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in [0.0, 2.856] \xrightarrow{0.2806} classer358, \\ & r_{3, classer358}^{NH4-2aerobic} : x_{NH4-2aerobic,i} \in (2.856, 20.282] \xrightarrow{1.0} classer358, \\ & r_{1, classer358}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [2.98, 4.479] \xrightarrow{0.0} classer358, \\ & r_{2, classer358}^{O2-1aerobic} : x_{O2-1aerobic,i} \in [4.479, 6.998] \xrightarrow{0.3514} classer358, \\ & r_{3, classer358}^{O2-1aerobic} : x_{O2-1aerobic,i} \in (6.998, 8.371] \xrightarrow{1.0} classer358, \\ & r_{1, classer358}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [4.94, 5.889] \xrightarrow{1.0} classer358, \\ & r_{2, classer358}^{O2-2aerobic} : x_{O2-2aerobic,i} \in [5.889, 8.705] \xrightarrow{0.4565} classer358, \\ & r_{3, classer358}^{O2-2aerobic} : x_{O2-2aerobic,i} \in (8.705, 8.842] \xrightarrow{1.0} classer358, \\ & r_{1, classer358}^{Valve-air} : x_{Valve-air,i} \in [28.604, 29.57] \xrightarrow{0.0} classer358, \\ & r_{2, classer358}^{Valve-air} : x_{Valve-air,i} \in [29.57, 51.168] \xrightarrow{0.4865} classer358, \\ & r_{3, classer358}^{Valve-air} : x_{Valve-air,i} \in (51.168, 54.777] \xrightarrow{1.0} classer358, \\ & r_{1, classer358}^{Q-air} : x_{Q-air,i} \in [697.829, 746.07] \xrightarrow{1.0} classer358, \\ & r_{2, classer358}^{Q-air} : x_{Q-air,i} \in [746.07, 1770.852] \xrightarrow{0.4505} classer358, \\ & r_{3, classer358}^{Q-air} : x_{Q-air,i} \in (1770.852, 2030.77] \xrightarrow{1.0} classer358, \\ & r_{1, classer358}^{h-ww} : x_{h-ww,i} \in [2.978, 2.984] \xrightarrow{1.0} classer358, \\ & r_{2, classer358}^{h-ww} : x_{h-ww,i} \in [2.984, 3.039] \xrightarrow{0.4888} classer358, \\ & r_{3, classer358}^{h-ww} : x_{h-ww,i} \in (3.039, 3.058] \xrightarrow{0.0} classer358, \\ & r_{1, classer358}^{Q-influent} : x_{Q-influent,i} \in [50.99, 63.038] \xrightarrow{1.0} classer358, \\ & r_{2, classer358}^{Q-influent} : x_{Q-influent,i} \in [63.038, 83.013] \xrightarrow{0.3151} classer358, \\ & r_{3, classer358}^{Q-influent} : x_{Q-influent,i} \in (83.013, 85.5] \xrightarrow{1.0} classer358, \\ & r_{1, classer358}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 47.2] \xrightarrow{1.0} classer358, \\ & r_{2, classer358}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [47.2, 50.692] \xrightarrow{0.3782} classer358, \\ & r_{3, classer358}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (50.692, 50.733] \xrightarrow{0.0} classer358,\end{aligned}$$

$$\begin{aligned}
r_{1,\text{classer358}}^{Freq-rec} &: x_{Freq-rec,i} \in [23.863, 24.899] \xrightarrow{1.0} \text{classer358}, \\
r_{2,\text{classer358}}^{Freq-rec} &: x_{Freq-rec,i} \in [24.899, 43.851] \xrightarrow{0.4413} \text{classer358}, \\
r_{3,\text{classer358}}^{Freq-rec} &: x_{Freq-rec,i} \in (43.851, 43.97] \xrightarrow{1.0} \text{classer358}, \\
r_{1,\text{classer358}}^{TN-influent} &: x_{TN-influent,i} \in [0.0, 16.209] \xrightarrow{1.0} \text{classer358}, \\
r_{2,\text{classer358}}^{TN-influent} &: x_{TN-influent,i} \in [16.209, 54.792] \xrightarrow{0.4186} \text{classer358}, \\
r_{3,\text{classer358}}^{TN-influent} &: x_{TN-influent,i} \in (54.792, 65.25] \xrightarrow{1.0} \text{classer358}, \\
r_{1,\text{classer358}}^{TN-effluent} &: x_{TN-effluent,i} \in [0.0, 5.371] \xrightarrow{1.0} \text{classer358}, \\
r_{2,\text{classer358}}^{TN-effluent} &: x_{TN-effluent,i} \in [5.371, 17.788] \xrightarrow{0.3421} \text{classer358}, \\
r_{3,\text{classer358}}^{TN-effluent} &: x_{TN-effluent,i} \in (17.788, 34.867] \xrightarrow{1.0} \text{classer358}, \\
r_{1,\text{classer358}}^{Temp-ww} &: x_{Temp-ww,i} \in [8.472, 13.327] \xrightarrow{1.0} \text{classer358}, \\
r_{2,\text{classer358}}^{Temp-ww} &: x_{Temp-ww,i} \in [13.327, 20.928] \xrightarrow{0.3039} \text{classer358}, \\
r_{3,\text{classer358}}^{Temp-ww} &: x_{Temp-ww,i} \in (20.928, 21.896] \xrightarrow{0.0} \text{classer358}, \\
r_{1,\text{classer358}}^{TOC-influent} &: x_{TOC-influent,i} \in [0.0, 38.888) \xrightarrow{1.0} \text{classer358}, \\
r_{2,\text{classer358}}^{TOC-influent} &: x_{TOC-influent,i} \in [38.888, 225.293] \xrightarrow{0.4817} \text{classer358}, \\
r_{3,\text{classer358}}^{TOC-influent} &: x_{TOC-influent,i} \in (225.293, 290.212] \xrightarrow{0.0} \text{classer358}, \\
r_{1,\text{classer358}}^{Nitritox-influent} &: x_{Nitritox-influent,i} \in [0.0, 0.542) \xrightarrow{1.0} \text{classer358}, \\
r_{2,\text{classer358}}^{Nitritox-influent} &: x_{Nitritox-influent,i} \in [0.542, 30.0] \xrightarrow{0.452} \text{classer358}, \\
r_{3,\text{classer358}}^{Nitritox-influent} &: x_{Nitritox-influent,i} \in (30.0, 53.708] \xrightarrow{0.0} \text{classer358}, \\
r_{1,\text{classer358}}^{TOC-effluent} &: x_{TOC-effluent,i} \in [0.0, 11.879) \xrightarrow{1.0} \text{classer358}, \\
r_{2,\text{classer358}}^{TOC-effluent} &: x_{TOC-effluent,i} \in [11.879, 42.251] \xrightarrow{0.4677} \text{classer358}, \\
r_{3,\text{classer358}}^{TOC-effluent} &: x_{TOC-effluent,i} \in (42.251, 44.053] \xrightarrow{0.0} \text{classer358}, \\
r_{1,\text{classer360}}^{NH4-influent} &: x_{NH4-influent,i} \in [7.775, 7.79) \xrightarrow{1.0} \text{classer360}, \\
r_{2,\text{classer360}}^{NH4-influent} &: x_{NH4-influent,i} \in [7.79, 34.353] \xrightarrow{0.5562} \text{classer360}, \\
r_{3,\text{classer360}}^{NH4-influent} &: x_{NH4-influent,i} \in (34.353, 40.541] \xrightarrow{0.0} \text{classer360}, \\
r_{2,\text{classer360}}^{NH4-2aerobic} &: x_{NH4-2aerobic,i} \in [0.0, 2.856] \xrightarrow{0.7194} \text{classer360}, \\
r_{3,\text{classer360}}^{NH4-2aerobic} &: x_{NH4-2aerobic,i} \in (2.856, 20.282] \xrightarrow{0.0} \text{classer360}, \\
r_{1,\text{classer360}}^{O2-1aerobic} &: x_{O2-1aerobic,i} \in [2.98, 4.479) \xrightarrow{1.0} \text{classer360}, \\
r_{2,\text{classer360}}^{O2-1aerobic} &: x_{O2-1aerobic,i} \in [4.479, 6.998] \xrightarrow{0.6486} \text{classer360}, \\
r_{3,\text{classer360}}^{O2-1aerobic} &: x_{O2-1aerobic,i} \in (6.998, 8.371] \xrightarrow{0.0} \text{classer360}, \\
r_{1,\text{classer360}}^{O2-2aerobic} &: x_{O2-2aerobic,i} \in [4.94, 5.889) \xrightarrow{0.0} \text{classer360}, \\
r_{2,\text{classer360}}^{O2-2aerobic} &: x_{O2-2aerobic,i} \in [5.889, 8.705] \xrightarrow{0.5435} \text{classer360}, \\
r_{3,\text{classer360}}^{O2-2aerobic} &: x_{O2-2aerobic,i} \in (8.705, 8.842] \xrightarrow{0.0} \text{classer360}, \\
r_{1,\text{classer360}}^{Valve-air} &: x_{Valve-air,i} \in [28.604, 29.57) \xrightarrow{1.0} \text{classer360}, \\
r_{2,\text{classer360}}^{Valve-air} &: x_{Valve-air,i} \in [29.57, 51.168] \xrightarrow{0.5135} \text{classer360}, \\
r_{3,\text{classer360}}^{Valve-air} &: x_{Valve-air,i} \in (51.168, 54.777] \xrightarrow{0.0} \text{classer360},
\end{aligned}$$

$$\begin{aligned}
r_{1, \text{classer360}}^{Q-air} : x_{Q-air,i} \in [697.829, 746.07] &\xrightarrow{0.0} \text{classer360}, \\
r_{2, \text{classer360}}^{Q-air} : x_{Q-air,i} \in [746.07, 1770.852] &\xrightarrow{0.5495} \text{classer360}, \\
r_{3, \text{classer360}}^{Q-air} : x_{Q-air,i} \in (1770.852, 2030.77] &\xrightarrow{0.0} \text{classer360}, \\
r_{1, \text{classer360}}^{h-ww} : x_{h-ww,i} \in [2.978, 2.984) &\xrightarrow{0.0} \text{classer360}, \\
r_{2, \text{classer360}}^{h-ww} : x_{h-ww,i} \in [2.984, 3.039] &\xrightarrow{0.5112} \text{classer360}, \\
r_{3, \text{classer360}}^{h-ww} : x_{h-ww,i} \in (3.039, 3.058] &\xrightarrow{1.0} \text{classer360}, \\
r_{1, \text{classer360}}^{Q-influent} : x_{Q-influent,i} \in [50.99, 63.038] &\xrightarrow{0.0} \text{classer360}, \\
r_{2, \text{classer360}}^{Q-influent} : x_{Q-influent,i} \in [63.038, 83.013] &\xrightarrow{0.6849} \text{classer360}, \\
r_{3, \text{classer360}}^{Q-influent} : x_{Q-influent,i} \in (83.013, 85.5] &\xrightarrow{0.0} \text{classer360}, \\
r_{1, \text{classer360}}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [39.153, 47.2) &\xrightarrow{0.0} \text{classer360}, \\
r_{2, \text{classer360}}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in [47.2, 50.692] &\xrightarrow{0.6218} \text{classer360}, \\
r_{3, \text{classer360}}^{FR1-DOTOK} : x_{FR1-DOTOK,i} \in (50.692, 50.733] &\xrightarrow{1.0} \text{classer360}, \\
r_{1, \text{classer360}}^{Freq-rec} : x_{Freq-rec,i} \in [23.863, 24.899] &\xrightarrow{0.0} \text{classer360}, \\
r_{2, \text{classer360}}^{Freq-rec} : x_{Freq-rec,i} \in [24.899, 43.851] &\xrightarrow{0.5587} \text{classer360}, \\
r_{3, \text{classer360}}^{Freq-rec} : x_{Freq-rec,i} \in (43.851, 43.97] &\xrightarrow{0.0} \text{classer360}, \\
r_{1, \text{classer360}}^{TN-influent} : x_{TN-influent,i} \in [0.0, 16.209) &\xrightarrow{0.0} \text{classer360}, \\
r_{2, \text{classer360}}^{TN-influent} : x_{TN-influent,i} \in [16.209, 54.792] &\xrightarrow{0.5814} \text{classer360}, \\
r_{3, \text{classer360}}^{TN-influent} : x_{TN-influent,i} \in (54.792, 65.25] &\xrightarrow{0.0} \text{classer360}, \\
r_{1, \text{classer360}}^{TN-effluent} : x_{TN-effluent,i} \in [0.0, 5.371) &\xrightarrow{0.0} \text{classer360}, \\
r_{2, \text{classer360}}^{TN-effluent} : x_{TN-effluent,i} \in [5.371, 17.788] &\xrightarrow{0.6579} \text{classer360}, \\
r_{3, \text{classer360}}^{TN-effluent} : x_{TN-effluent,i} \in (17.788, 34.867] &\xrightarrow{0.0} \text{classer360}, \\
r_{1, \text{classer360}}^{Temp-ww} : x_{Temp-ww,i} \in [8.472, 13.327) &\xrightarrow{0.0} \text{classer360}, \\
r_{2, \text{classer360}}^{Temp-ww} : x_{Temp-ww,i} \in [13.327, 20.928] &\xrightarrow{0.6961} \text{classer360}, \\
r_{3, \text{classer360}}^{Temp-ww} : x_{Temp-ww,i} \in (20.928, 21.896] &\xrightarrow{1.0} \text{classer360}, \\
r_{1, \text{classer360}}^{TOC-influent} : x_{TOC-influent,i} \in [0.0, 38.888) &\xrightarrow{0.0} \text{classer360}, \\
r_{2, \text{classer360}}^{TOC-influent} : x_{TOC-influent,i} \in [38.888, 225.293] &\xrightarrow{0.5183} \text{classer360}, \\
r_{3, \text{classer360}}^{TOC-influent} : x_{TOC-influent,i} \in (225.293, 290.212] &\xrightarrow{1.0} \text{classer360}, \\
r_{1, \text{classer360}}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.0, 0.542) &\xrightarrow{0.0} \text{classer360}, \\
r_{2, \text{classer360}}^{Nitritox-influent} : x_{Nitritox-influent,i} \in [0.542, 30.0] &\xrightarrow{0.548} \text{classer360}, \\
r_{3, \text{classer360}}^{Nitritox-influent} : x_{Nitritox-influent,i} \in (30.0, 53.708] &\xrightarrow{1.0} \text{classer360}, \\
r_{1, \text{classer360}}^{TOC-effluent} : x_{TOC-effluent,i} \in [0.0, 11.879) &\xrightarrow{0.0} \text{classer360}, \\
r_{2, \text{classer360}}^{TOC-effluent} : x_{TOC-effluent,i} \in [11.879, 42.251] &\xrightarrow{0.5323} \text{classer360}, \\
r_{3, \text{classer360}}^{TOC-effluent} : x_{TOC-effluent,i} \in (42.251, 44.053] &\xrightarrow{1.0} \text{classer360} \quad \}
\end{aligned}$$