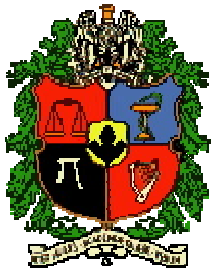
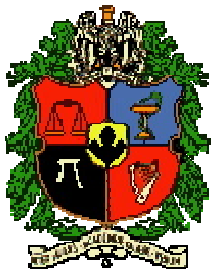
GUÍA METODOLÓGICA PARA LA SELECCIÓN DE TÉCNICAS DE DEPURACIÓN DE DATOS



IVÁN AMÓN URIBE

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE MINAS, ESCUELA DE SISTEMAS
MEDELLÍN
2010

GUÍA METODOLÓGICA PARA LA SELECCIÓN DE TÉCNICAS DE DEPURACIÓN DE DATOS



IVÁN AMÓN URIBE

TESIS DE MAESTRÍA
MAESTRÍA EN INGENIERÍA - INGENIERÍA DE SISTEMAS

Directora:
CLAUDIA JIMÉNEZ RAMÍREZ, Ph.D

UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE MINAS, ESCUELA DE SISTEMAS
MEDELLÍN
2010

AGRADECIMIENTOS

A la profesora Claudia Jiménez Ramírez, adscrita a la Escuela de Sistemas de la Universidad Nacional de Colombia Sede Medellín, por su disponibilidad permanente y acompañamiento continuo durante el desarrollo de este trabajo.

A la Universidad Pontificia Bolivariana, por costear mis estudios de maestría.

A mis familiares, por su apoyo y comprensión.

CONTENIDO

pág.

LISTA DE TABLAS

LISTA DE FIGURAS

RESUMEN

Es una realidad que parte de los datos almacenados por las organizaciones, contienen errores y estos pueden conducir a tomar decisiones erróneas, ocasionando pérdidas de tiempo, dinero y credibilidad. Esta situación ha capturado la atención de los investigadores, llevando al desarrollo de múltiples técnicas para detectar y corregir los problemas en los datos, pero no es trivial decidir cuáles técnicas deben aplicarse a un conjunto de datos particular de la vida real.

Para lograr buenos resultados en procesos de limpieza de datos, la elección de la técnica es fundamental, pero no se conoce de alguna metodología que detalle la forma de realizar dicha selección de técnicas. Es por esto que esta tesis de maestría construye una guía metodológica que oriente al analista de los datos hacia una selección, con mayor rigor científico, de las técnicas adecuadas para aplicar a un conjunto de datos particular de un dominio específico. La guía metodológica construida en este trabajo, orienta la selección de técnicas para tres de los posibles problemas que pueden presentar los datos: detección de duplicados, valores atípicos incorrectos y valores faltantes.

Para la construcción de la guía, se caracterizaron varias técnicas para cada uno de los tres problemas de datos bajo estudio, examinando su eficacia ante diferentes casos o situaciones problemáticas propuestas. Para realizar comparativos y validar la guía, se utilizaron tanto datos de prueba como reales pertenecientes al censo de los Estados Unidos y a la base de datos Scienti de Colciencias.

Analistas de datos que requieran hacer tareas de depuración de datos para los tres problemas mencionados, encontrarán una guía metodológica expresada mediante diagramas de flujo, la cual recomienda una o varias técnicas –de entre algunas estudiadas- para su situación particular.

INTRODUCCIÓN

Ante los grandes volúmenes de datos almacenados hoy en las organizaciones y su creciente papel en la toma de decisiones, es cada vez más importante, que estén libres de errores y evitar así tomar decisiones erróneas.

Idealmente, los datos almacenados no deberían contener errores, pero es innegable que son una realidad y merecen atención. La idea de este trabajo de maestría, es ofrecer apoyo a los analistas de los datos para que puedan elegir, con mayor rigor, las técnicas de depuración a aplicar a un conjunto de datos con características particulares.

Los datos pueden presentar problemas de diferentes tipos como duplicados, valores faltantes (nulos), valores atípicos algunos de ellos incorrectos o variaciones tipográficas, entre muchos otros, y para cada tipo existen técnicas para detectarlos y corregirlos. La calidad de la limpieza lograda sobre los datos, depende de la técnica aplicada y la elección de la técnica debe tener en cuenta la naturaleza de los datos específicos sobre los que se está trabajando.

Contar con una guía metodológica que oriente la selección de la técnica a aplicar en un caso particular, es un apoyo significativo para las personas que deben realizar tareas de depuración a los datos organizacionales. Al no encontrar evidencia de trabajos de investigación en este sentido, en esta tesis de maestría se construyó una aproximación metodológica que conduzca al analista de los datos, hacia una selección con mayor rigor de la técnica apropiada -entre las analizadas- para aplicar a un conjunto de datos particular o dominio específico.

Para la construcción de la guía, se caracterizaron algunas de las técnicas existentes evidenciando sus fortalezas y debilidades para tres tipos de problemas, para luego idear la secuencia de razonamientos que constituye el proceso guiado para recomendar las técnicas a utilizar.

Este documento está organizado como sigue: el capítulo 2 presenta los objetivos y el alcance del trabajo, el capítulo 3 presenta los fundamentos teóricos necesarios para entender el resto del trabajo, el capítulo 4 presenta el tema de la detección de duplicados, el capítulo 5 presenta el tema de los valores ausentes o nulos y el capítulo 6 el tema de los valores atípicos incorrectos. En cada uno de los tres capítulos centrales, se incluye el diseño del experimento realizado, la comparación de las técnicas, la guía metodológica y las conclusiones sobre cada tema individualmente.

1. OBJETIVOS Y ALCANCE

1.1. Objetivo General

El objetivo general a lograr con esta tesis de maestría es:

Diseñar una guía metodológica para la selección de las técnicas de depuración de datos, a ser aplicadas a un conjunto de datos particular o un dominio específico que presente problemas de duplicación de campos tipo texto, valores atípicos incorrectos y/o valores faltantes o nulos.

1.2. Objetivos específicos

Los objetivos específicos mediante los cuales se llegó al cumplimiento del objetivo general son:

- ✓ Identificar al menos tres técnicas para depuración de datos, de cada uno de los tres tipos de problemas a abordar (detección de duplicados, valores atípicos incorrectos y valores faltantes).

- ✓ Caracterizar las técnicas antes seleccionadas, con el fin de conocer sus propiedades, funcionalidad, fortalezas y debilidades.

- ✓ Comparar las técnicas.

- ✓ Diseñar la guía metodológica para la selección de la técnica más adecuada según la naturaleza de los datos a depurar.

1.3. Alcance

El trabajo tiene unas fronteras a precisar, así:

Aplica a datos estructurados, como atributos de tablas de una base de datos, campos de una bodega de datos, columnas de una hoja de cálculo o un archivo

de texto. No aplica a datos no estructurados como páginas Web o correos electrónicos.

Contempla tres de los problemas que pueden presentar los datos almacenados: detección de duplicados, detección de valores atípicos incorrectos y datos faltantes o nulos. Estos tres tipos de problemas, se seleccionaron ya que, según la literatura, son frecuentes en los datos y se dispone de diversas técnicas para tratarlos. No se cubrirán problemas como falta de integridad referencial, valores que entran en conflicto entre sí u otros.

Inicialmente se había planteado abarcar tres técnicas para cada uno de los tres tipos de problemas considerados, pero se logró contemplar nueve técnicas para la detección de duplicados, cinco para valores faltantes y cinco para valores atípicos incorrectos.

Aplica a atributos individuales, es decir, la guía metodológica propone técnicas para detectar los problemas anteriormente citados, pero para un atributo a la vez. Para depurar varios atributos, deberá seguirse varias veces la guía.

Se implementaron los algoritmos de las técnicas y/o se utilizaron programas estadísticos, pero no se construyó un software que incorpore toda la guía metodológica obtenida.

2. FUNDAMENTOS TEÓRICOS

2.1. La Calidad de los Datos

Actualmente, las organizaciones toman sus decisiones cada vez más con base en el conocimiento derivado de los datos almacenados en sus bases o bodegas de datos, aplicando el enfoque denominado Inteligencia del Negocio (*Business Intelligence*, en inglés) y no con el juicio subjetivo o las intuiciones de las directivas. Por tanto, es de vital importancia que los datos contengan la menor cantidad de errores posibles.

Dasu *et al*. en el año 2003 [Dasu *et. al.*, 2003], afirmaban: "no es poco frecuente que las operaciones de bases de datos tengan del 60% al 90% de problemas de calidad de datos". En el mismo sentido, una investigación realizada por la firma Gartner en el año 2007 [Gartner, 2007], indica que más del 25% de los datos críticos en las compañías Fortune 1000, presentan errores. Según Andreas Bitterer, vicepresidente de investigación de Gartner, "No existe una compañía en el planeta que no tenga un problema de calidad de datos y aquellas compañías que reconocen tenerlo, a menudo subestiman el tamaño de éste".

Los datos "sucios" pueden conducir a decisiones erróneas, ocasionando pérdidas de tiempo, dinero y credibilidad. Gartner [Gartner, 2007] afirma: "la mala calidad de los datos de los clientes, lleva a costos importantes, como el sobreestimar el volumen de ventas de los clientes, el exceso de gastos en los procesos de contacto con los clientes y pérdida de oportunidades de ventas. Pero ahora las empresas están descubriendo que la calidad de los datos tiene un impacto significativo en la mayoría de sus iniciativas empresariales estratégicas, no sólo de ventas y marketing. Otras funciones como elaboración de presupuestos, producción y distribución también se ven afectadas".

Los problemas en los datos se pueden presentar al reunir información proveniente de varias fuentes, o al interior de un archivo o una misma tabla en una base de datos relacional. Por ejemplo, el atributo *nombre* puede contener *"Juan Alberto López Gómez"* para un registro y para otro *"Juan A. López G."* haciendo referencia a la misma persona o el *nombre* "Carlos" puede presentar errores de digitación con caracteres sobrantes, faltantes o transpuestos ("Carklos", "Calos", Catlos").

Otro tipo de error corresponde a los datos atípicos, pues aunque pueden aparentar ser inválidos pueden ser correctos y viceversa, como es el caso de un 90% de las ventas destinado a investigación. Los valores atípicos también

son conocidos como *Outliers* y son aquellas observaciones que se desvían significativamente de la mayoría de observaciones. De otra parte, no todos los valores faltantes son problemas, pero datos que deberían tener valor y no lo tienen, si lo son. Las diferentes convenciones utilizadas para los valores faltantes (nulos), también pueden generar problemas en la realización de las tareas de depuración o de minería a los datos.

La importancia de contar con datos confiables, con los cuales se puedan tomar decisiones acertadas, es cada vez mayor. Conceptos como Gestión del Conocimiento, Minería de Datos e Inteligencia de Negocios, se están desarrollando a pasos agigantados, y de poco o nada sirven si se basan en datos errados.

Para ilustrar los posibles errores en los datos y los problemas que originan, se toma como ejemplo la base de datos ScienTI, donde se registra la actividad investigativa de Colombia. En ella, cada investigador actualiza el programa CvLAC (*Curriculum vitae Latinoamericano y el Caribe*) registrando los proyectos de investigación en los que participa. Dado que en un proyecto (i) pueden participar varios investigadores, (ii) cada uno de ellos ingresa sus datos al sistema por separado y (iii) no existe una identificación única de los proyectos, puede suceder que el nombre de un mismo proyecto no se escriba exactamente igual por parte de todos sus integrantes (por ejemplo: *"Guía Metodológica para selección de técnicas para depuración de datos"* vs *"Guía Metodológica para selección de técnicas de depuración de datos")*. Si no se toman las medidas adecuadas, previas a la contabilización de la cantidad de proyectos, se obtendrá un número que sobredimensiona la producción académica de las universidades o centros y se distorsionará la realidad.

Para los múltiples errores que pueden presentar los datos, diversos investigadores han desarrollado técnicas, para detectarlos y corregirlos. Bajo diferentes nombres, como calidad de datos (*Data Quality*), heterogeneidad de datos (*Data Heterogeneity*), limpieza de datos (*Data Cleansing*), reconciliación de datos (*Data Reconciliation*), se tratan temas relacionados con esta problemática.

2.2. Trabajos Relacionados

Múltiples trabajos se han realizado en la temática de calidad de datos. A continuación, se relacionan algunos que son de interés para el propósito de esta tesis de maestría.

En cuanto a trabajos relacionados con la clasificación y la detección de los problemas, son varios los que han realizado clasificaciones de las anomalías de los datos [Rahm y Do, 2000] [Kim *et. al.*, 2003] [Müller y Freytag, 2003].

Oliveira *et. al.* [Oliveira *et.al.*, 2005] no sólo realizan una taxonomía con treinta y cinco problemas de calidad de los datos, sino que plantean métodos semiautomáticos para detectarlos, los cuales representan mediante árboles binarios. Los árboles corresponden al razonamiento que se necesita hacer para detectar un problema particular.

Un tipo de problema que pueden presentar los datos es el conocido como *Record Linkage,* detección de duplicados o deduplicación [Elmagarmid *et. al.*, 2007], el cual tiene como meta identificar registros que se refieren a la misma entidad del mundo real, aun si los datos no son idénticos, esto es, se trata de la detección de atributos o registros que tienen contenidos distintos pero debieran ser el mismo. Una misma entidad del mundo real puede aparecer representada dos o más veces (duplicada) a través de una o varias bases de datos, en tuplas con igual estructura, que no comparten un identificador único y presentan diferencias textuales en sus valores. En el capítulo 4 se trata este tema en detalle.

Otro tipo de problemas es el de los valores atípicos, conocidos como *Outliers* [Chandola *et. al.*, 2007]. Aunque no necesariamente son errores, pueden ser generados por un mecanismo diferente de los datos normales como problemas en los sensores, distorsiones en el proceso, mala calibración de instrumentos y/o errores humanos. También sobre este tema se han realizado múltiples investigaciones, entre las cuales se encuentran trabajos tipo resumen [Chandola *et. al.*, 2007], trabajos comparativos [Bakar *et. al.*, 2006] [Matsumoto *et. al.*, 2007], trabajos sobre técnicas específicas [Angiulli *et. al.*, 2006] [Narita y Kitagawa, 2008], entre muchos otros. En el capítulo 5 se trata este tema en detalle.

En los trabajos realizados, también se encuentran algunos de tipo metodológico. Tierstein presenta una metodología que incorpora dos tareas que se interrelacionan y se traslapan: limpiar los datos de un sistema *legacy* y convertirlos a una base de datos [Tierstein, 2005]. Rosenthal, extiende los sistemas de bases de datos para manejar anotaciones de calidad de los datos en las bases de datos mediante metadatos [Rosenthal, 2001]. Rittman presenta la metodología seguida por el módulo de Oracle encargado de realizar depuración a los datos (*Oracle Warehouse Builder*), para realizar este proceso [Rittman, 2006].

Las técnicas desarrolladas por los investigadores hasta el momento, son variadas y casi siempre aplican a un tipo de problema en particular.

Es así como existen técnicas para tratar el problema de la detección de duplicados, para detección y corrección de valores atípicos, para tratar con los valores faltantes y para cada posible problema que puedan presentar los datos.

Para la detección de duplicados, se encuentran técnicas como la distancia de edición [Ristad y Yianilos, 1998], distancia de brecha afín [Waterman et.al., 1976], distancia de Smith-Waterman [Smith y Waterman, 1981], distancia de Jaro [Jaro, 1976], q-grams [Ullmann, 1977], Whirl [Cohen, 1998] y técnicas fonéticas como *soundex* [Russell, 1918, 1922], NYSIIS [Taft, 1970], ONCA [Gill, 1997], *Metaphone* [Philips, 1990] y *Double Metaphone* [Philips, 2000].

Para la detección de valores atípicos se encuentran técnicas como el diagrama de cajas y bigotes [Chambers *et. al.*, 1983], la prueba de Dixon [Dixon y Massey, 1983], la prueba de Grubbs [Grubbs, 1969] y regresión [Robiah *et. al.*, 2003], entre otras.

Para dar solución a los datos faltantes, existen técnicas como las imputaciones de media, mediana, moda, *Hot Deck* e imputación por regresión [Medina y Galván, 2007], entre otras.

2.3. Necesidad de una Metodología para seleccionar técnicas

Elmagarmid *et al*. [Elmagarmid *et. al.*, 2007] plantean que:

- ✓ Ninguna métrica es adaptable a todos los conjuntos de datos.

- ✓ Es poco probable que se resuelva pronto la pregunta de cuál de los métodos debe utilizarse para una determinada tarea de depuración.

- ✓ La tarea de depuración de datos, es altamente dependiente de los datos y no está claro si nunca se va a ver una técnica que domine a todas las demás en todos los conjuntos de datos.

Lo anterior significa que la calidad de la limpieza lograda sobre los datos, depende de la técnica aplicada y la elección de la técnica está íntimamente ligada con la naturaleza de los datos específicos sobre los que se está trabajando.

La selección de las técnicas para depuración de datos, que permitan la entrega de información de buena calidad para la toma de decisiones, requiere de conocimiento profundo de las propiedades de cada una de las técnicas, sus características y cuándo pueden ser aplicadas con éxito a un conjunto de datos dependiendo de la naturaleza de los mismos.

Para mostrar como la técnica depende del conjunto de observaciones que se deben depurar, se toma como ejemplo la detección de valores atípicos. Ésta puede ser fácil visualmente con el apoyo de una gráfica que muestre la dispersión de los puntos, pero para la detección automática de estos valores por medio de un procedimiento o una función almacenada, se necesita alguna técnica matemática para hallarlos. Comúnmente se usa la fórmula de Tukey, basada en los cuartiles de la distribución o los valores que subdividen el conjunto de datos ordenados en cuatro partes, cada una con el mismo porcentaje de datos. Tomando como referencia la diferencia entre el primer cuartil Q_1 y el tercer cuartil Q_3, o el valor intercuartil, se considera un valor extremo o atípico aquel que se encuentra a 1,5 veces esa distancia de uno de esos cuartiles (atípico leve) o a 3 veces esa distancia (atípico extremo). Sin embargo, dependiendo de la distribución de los datos, este método puede fallar. Si el rango intercuartil resulta ser cero, cualquier valor diferente de cero se tomaría como atípico. Por lo tanto, en estos casos, es recomendable usar otro método.

De otra parte, las herramientas comerciales que realizan depuración a los datos, en general, no realizan autónoma y automáticamente este trabajo, sino que requieren la intervención del usuario. Generalmente, ofrecen un conjunto de opciones entre las cuales se debe elegir la técnica a ser aplicada a los datos, tarea que demanda altos conocimientos técnicos. Algunas herramientas inclusive aplican por defecto técnicas que no son adecuadas en ocasiones, como es el caso de descartar los registros imcompletos [Kalton y Kasprzyk, 1982] [UCLA, 2009].

Teniendo en mente todo lo anterior, surge entonces la pregunta: ¿Cómo determinar las técnicas que deben ser empleadas para realizar procesos de depuración a los datos en un caso particular? En la literatura revisada, esta pregunta no se responde satisfactoriamente, ya que no se encontró evidencia de una metodología que indique claramente el procedimiento para seleccionar la técnica más adecuada -bajo alguna(s) métrica(s) predefinidas- a aplicar en una situación específica, considerando la naturaleza de los datos en cuestión y el tipo de inconveniente o error que presenten los datos.

Los trabajos de investigación mencionados en la Sección 2.2, incluyendo aquellos de tipo metodológico, no se ocupan lo suficiente de la selección de las técnicas para depurar los datos, en un caso particular. El trabajo de Oliveira *et. al.* [Oliveira et.al., 2005], plantea sin mayor detalle, como detectar una anomalía de los datos sin indicar cual técnica usar para su detección y/o corrección. Tierstein, aunque presenta una metodología que intenta cubrir todo el proceso de depuración de datos, se enfoca principalmente hacia el manejo de los datos históricos, no examina las técnicas existentes para depuración, el paso de detección de los defectos. No se ocupa de recomendar una técnica y no tiene en cuenta la naturaleza de los datos [Tierstein, 2005]. Rosenthal *et. al.*, están orientados al enriquecimiento de los sistemas de bases de datos con

metadatos, sin examinar ni recomendar técnicas de depuración [Rosenthal, 2001].

Una metodología ampliamente conocida y usada en proyectos de descubrimiento de conocimiento en bases de datos (*KDD: Knowledge Discovery in Databases*, en inglés) como Crisp-Dm [CRISP-DM, 2000], aunque en su fase de preparación de los datos se ocupa de la transformación y limpieza de los datos, no llega hasta el nivel de recomendar técnicas específicas dependiendo de la naturaleza de los datos. Similar situación sucede con SEMMA [SAS, 2003], otra metodología de *KDD* estrechamente ligada a los productos *SAS*[1]. La metodología seguida por Oracle en [Rittman, 2006], confirma que el software ofrecido para la depuración de los datos no selecciona por el usuario, la técnica a aplicar.

Por lo tanto, la elección de la técnica es esencial, pero no se conoce de alguna metodología que detalle la forma de realizar esta tarea. Es por esto que esta tesis de maestría construye una guía metodológica que oriente al analista de los datos, hacia una selección, con mayor rigor científico de las técnicas adecuadas para aplicar a un conjunto de datos particular de un dominio específico. La guía metodológica construida en este trabajo, orienta la selección de técnicas para tres de los posibles problemas que pueden presentar los datos: detección de duplicados, valores atípicos incorrectos y valores faltantes.

[1] SAS. [En línea]. http://www.sas.com/products/ [Consulta: Febrero 10 de 2009]

3. DETECCIÓN DE DUPLICADOS

Uno de los posibles conflictos en los datos, se origina cuando una misma entidad del mundo real aparece representada dos o más veces (duplicada) a través de una o varias bases de datos, en tuplas con igual estructura, que no comparten un identificador único y presentan diferencias textuales en sus valores. Por ejemplo, en cierta base de datos una persona puede aparecer como:

Nombre	**e-mail**
Jorge Eduardo Rodríguez López	jorge.rodriguez@gmail.com

Mientras que en otra, debido, por ejemplo, a errores ortográficos, la misma persona puede aparecer como

nombre_persona	**email_persona**
Jorje Eduardo Rodrigues Lópes	jorge.rodrigues@jmail.com

Ambas tuplas tienen la misma estructura: dos campos, uno para el nombre completo y otro para el correo electrónico. Además, no comparten un identificador único que permita intuir fácilmente que se trata de la misma persona, como podría ser el número de cédula. Luego, dado que sus valores son similares mas no idénticos, no es fácil identificarlas como la misma persona con una sentencia SQL clásica.

El proceso que detecta este conflicto se conoce con múltiples nombres: *record linkage* o *record matching* entre la comunidad estadística; *database hardening* en el mundo de la Inteligencia Artificial; *merge-purge*, *data deduplication* o *instance identification* en el mundo de las Bases de Datos; otros nombres como *coreference resolution* y *duplicate record detection* también son usados con frecuencia. Aquí se utilizará el término genérico *detección de duplicados*.

El proceso de *record linkage* fue primero planteado por Dunn en 1946 [Dunn, 1946]. Algunos fundamentos probabilísticos fueron luego desarrollados por Newcombe *et al.* [Newcombe, *et. al.*, 1959] [Newcombe y Kennedy, 1962], siendo formalizados por Fellegi y Sunter en 1969 como una regla de decisión probabilística [Fellegi y Sunter, 1969]. Algunas mejoras de este modelo han sido propuestas por Winkler [Winkler, 1989, 1990, 1993, 2000].

En su forma más simple, dicho proceso es como sigue:

Dado un conjunto R de registros a detectarle duplicados:

1) Se define un umbral real $\theta \in [0,1]$.

2) Se compara cada registro de R con el resto; existen variaciones más eficientes de este paso, pero están fuera del alcance de este trabajo.

3) Si la *similitud* entre una pareja de registros es mayor o igual que θ, se asumen duplicados; es decir, se consideran representaciones de una misma entidad real.

Entonces es necesaria alguna *función de similitud* que, dados dos registros con la misma estructura, devuelva un número real en el intervalo [0,1]: igual a uno si ambos, registros son idénticos y menor entre más diferentes sean. Tal valor depende de la similitud entre cada pareja de atributos respectivos; en el ejemplo anterior, de la similitud entre los dos nombres y los dos correos electrónicos. Luego, es necesaria otra función de similitud al nivel de atributo (no de registro), siendo frecuente en este contexto tratar los atributos, sin importar su tipo de datos, como cadenas de texto (*strings*), y desarrollar *funciones de similitud sobre cadenas de texto* que retornan un número real en [0,1]: uno si ambas cadenas son idénticas, menor entre más diferentes sean y cero si, en general, no tienen ni un solo carácter en común.

Estas funciones de similitud han sido tema de investigación por años. Actualmente existen muchas funciones, las cuales pueden ser clasificadas en dos grandes categorías: basadas en caracteres y basadas en tokens. Las primeras consideran cada cadena de texto como una secuencia ininterrumpida de caracteres. Las segundas como un conjunto de subcadenas delimitadas por caracteres especiales, como espacios en blanco, comas y puntos; esto es, como un conjunto de palabras o *tokens*, y calculan la similitud entre cada pareja de *tokens* mediante alguna función basada en caracteres.

El valor de un atributo puede aparecer representado de muchas formas diferentes. Por ejemplo, el nombre “Jorge Eduardo Rodríguez López” también puede aparecer como “Rodríguez López Jorge Eduardo”, “Jorge E Rodríguez López” o “Jorge Eduardo Rodríguez L.”; con errores de escritura, como “Jorje Eduardo Rodrigues Lópes”; con información adicional, como “PhD Jorge Eduardo Rodríguez López”, entre otras. Las causas pueden ser muchas: restricciones de formato, de longitud y/o en el conjunto de caracteres permitidos, errores humanos al capturar los datos, errores que surgen integrando bases de datos diferentes o haciendo migración entre sistemas, modelos de datos mal diseñados, entre otras causas.

Por todo lo anterior, Elmagarmid *et al.* [Elmagarmid *et. al.* 2007] concluyen que la gran cantidad de formas en que un mismo atributo puede aparecer representado convierten la elección de la función de similitud más apropiada en todo un problema, en el cual se requiere aún mayor investigación. Algunos estudios comparativos han sido realizados: Christen compara la eficacia de

algunas funciones de similitud sobre nombres personales [Christen, 2006]. Yancey lo hace sobre nombres personales extraídos del censo realizado en 2000 en Estados Unidos [Yancey, 2006]. Cohen *et al.* [Cohen *et. al.*, 2003] utilizan diversos conjuntos de datos: nombres personales, de animales, de empresas, de videojuegos, de parques, entre otros. Moreau *et al.* se limitan a nombres de ubicaciones, empresas y personas [Moreau *et. al.*, 2008]. Sin embargo, ninguno de ellos compara la eficacia de las técnicas basándose en distintas situaciones problemáticas como las planteadas en el presente trabajo (introducción de errores ortográficos, uso de abreviaturas, palabras faltantes, introducción de prefijos/sufijos sin valor semántico, reordenamiento de palabras y eliminación/adición de espacios en blanco) ni propone una guía metodológica para ayudar en la selección de las técnicas más adecuadas a una situación particular de acuerdo con la naturaleza de los datos.

El resto del presente capítulo está organizado como sigue: la sección 4.1 describe las funciones de similitud sobre cadenas de texto comparadas en este trabajo. La sección 4.2 describe la métrica de evaluación utilizada. La sección 4.3 describe el diseño del experimento realizado para evaluar la eficacia de las diferentes técnicas. La sección 4.4 muestra los resultados obtenidos y en la sección 4.5 se presenta la guía metodológica para el problema de la detección de duplicados. Por último se presentan las conclusiones sobre este tema en la sección 4.6.

3.1. Funciones de similitud sobre cadenas de texto

La notación para el presente trabajo es la siguiente: Σ denota algún conjunto finito ordenado de caracteres y Σ^* el conjunto de *strings* formados por la concatenación de cero o más caracteres de Σ. A y B denotan dos cadenas de longitud *n* y *m* definidos sobre Σ^*, donde $n \geq m$. a_i representa algún carácter de A para $1 \leq i \leq n$, y b_j es análogo respecto a B.

Aunque algunas veces, en lugar de similitud, se habla de distancia entre dos cadenas, es importante entender que una magnitud puede calcularse a partir de la otra. Una distancia real $d \in [0,1]$, donde 0 indica que ambas cadenas son idénticas y 1 que no tienen ni un solo carácter en común, equivale a una similitud $s = 1 - d$, lo que quiere decir que a mayor distancia, menor es la similitud y viceversa.

Actualmente existen diversas funciones de similitud, las cuales pueden ser clasificadas en dos categorías: basadas en caracteres y basadas en *tokens* (Elmagarmid *et. al.*, 2007)

3.1.1. Funciones de similitud basadas en caracteres.

Estas funciones de similitud consideran cada cadena como una secuencia ininterrumpida de caracteres. En esta sección se cubren las siguientes:

- ✓ Distancia de edición.
- ✓ Distancia de brecha afín.
- ✓ Similitud Smith-Waterman.
- ✓ Similitud de Jaro.
- ✓ Similitud de *q-grams*.

3.1.1.1. *Distancia de edición.* La distancia de edición entre dos cadenas A y B se basa en el conjunto mínimo de operaciones de edición necesarias para transformar A en B (o viceversa). Las operaciones de edición permitidas son:

- ✓ Reemplazar un carácter de A por otro de B (o viceversa).
- ✓ Eliminar un carácter de A ó B.
- ✓ Insertar un carácter de B en A (o viceversa).

En el modelo original, cada operación de edición tiene costo unitario, siendo referido como distancia de Levenshtein [Levenshtein, 1966]. Needleman y Wunsch [Needleman y Wunsh, 1970] lo modificaron para permitir operaciones de edición con distinto costo, permitiendo modelar errores ortográficos y tipográficos comunes. Por ejemplo, es usual encontrar “n” en lugar de “m” (o viceversa) [Ramírez y López, 2006], entonces, tiene sentido asignar un costo de sustitución menor a este par de caracteres que a otros dos sin relación alguna. Lowrance y Wagner [Lowrence y Wagner, 1975] introducen un modelo que permite la trasposición de dos caracteres adyacentes como una cuarta operación de edición, usualmente referido como distancia de Damerau-Levenstein.

Los modelos anteriores tienen una desventaja: la distancia entre dos cadenas de texto carece de algún tipo de normalización, lo cual los hace imprecisos. Por ejemplo, tres errores son más significativos entre dos cadenas de longitud 4 que entre dos cadenas de longitud 20. Existen varias técnicas de normalización. Las más simples lo hacen dividiendo por la longitud de la cadena más larga [Christen, 2006] o por la suma de la longitud de ambos *strings* [Weigel y Fein, 1994]. Marzal y Vidal [Marzal y Vidal, 1993] proponen dividir por el número de operaciones de edición. Más recientemente, Yujiang y Bo [Yujian y Bo, 2007] desarrollaron la primera técnica de normalización que

satisface la desigualdad triangular[2]. Por otro lado, Ristad y Yiannilos [Ristad y Yianilos, 1998] lograron un modelo que aprende automáticamente los costos más óptimos de las operaciones de edición a partir de un conjunto de datos de entrenamiento.

La distancia de edición no normalizada puede ser calculada en $O(nm)$ mediante el algoritmo de programación dinámica propuesto por Wagner y Fisher [Wagner y Fischer, 1974]. Masek presenta un algoritmo que toma $O\left(n \cdot \max(1, m/\log n)\right)$ siempre que los costos de las operaciones de edición sean múltiplos de un único real positivo y que sea finito [Masek, 1980]. Hyyrö presenta un algoritmo basado en operaciones a nivel de bits para verificar si la distancia de Damerau-Levenshtein entre dos *strings* es menor que cierta constante d, el cual toma $O(|\Sigma| + \lceil d/w \rceil \cdot m)$, donde w es el tamaño de la palabra definida por el procesador [Hyyrö, 2002]. La técnica de normalización propuesta por Marzal y Vidal toma $O(n^2 m)$, siendo reducida a $O(nm \log m)$ [Weigel y Fein, 1994] y a $O(nm)$ mediante programación fraccionaria [Marzal y Vidal, 1993]. Eğecioğlu e Ibel proponen algoritmos paralelos eficientes [Eğecioğlu e Ibel, 1996]. El orden computacional de la técnica de normalización de Yujiang y Bo es igual al del algoritmo que se utilice para calcular la distancia no normalizada, al ser función directa de esta última.

3.1.1.2. *Distancia de brecha afín.* Como se muestra más adelante, la distancia de edición y otras funciones de similitud tienden a fallar identificando cadenas equivalentes que han sido demasiado truncadas, ya sea mediante el uso de abreviaturas o la omisión de tokens ("Jorge Eduardo Rodríguez López" vs "Jorge E Rodríguez"). La distancia de brecha afín ofrece una solución al penalizar la inserción/eliminación de k caracteres consecutivos (brecha) con bajo costo, mediante una función afín $\rho(k) = g + h \cdot (k - 1)$, donde g es el costo de iniciar una brecha, h el costo de extenderla un carácter, y $h \ll g$ [Gotoh, 1982]. Bilenko y Mooney describen un modelo para entrenar automáticamente esta función de similitud a partir de un conjunto de datos [Bilenko y Mooney, 2003].

La distancia de brecha afín no normalizada puede ser calculada en $O(nm)$ mediante el algoritmo de programación dinámica propuesto por Gotoh [Gotoh, 1982].

[2] En el sentido matemático estricto, toda medida de distancia (métrica) debe satisfacer la desigualdad triangular: la distancia directa para ir del punto *x* al *z* nunca es mayor que aquella para ir primero del punto *x* al *y* y después del *y* al *z*. Sin embargo, esta propiedad carece de importancia para procesos de detección de duplicados, y por lo tanto no será discutida.

3.1.1.3. *Similitud Smith-Waterman.* La similitud Smith-Waterman entre dos cadenas A y B es la máxima similitud entre una pareja (A′,B′), sobre todas las posibles, tal que A′ es subcadena de A y B′ es subcadena de B. Tal problema se conoce como alineamiento local. El modelo original de Smith y Waterman [Smith y Waterman, 1981] define las mismas operaciones de la distancia de edición, y además permite omitir cualquier número de caracteres al principio o final de ambas cadenas. Esto lo hace adecuado para identificar cadenas equivalentes con prefijos/sufijos que, al no tener valor semántico, deben ser descartados. Por ejemplo, "PhD Jorge Eduardo Rodríguez López" y "Jorge Eduardo Rodríguez López, Universidad Nacional de Colombia" tendrían una similitud cercana a uno, pues el prefijo "PhD" y el sufijo "Universidad Nacional de Colombia" serían descartados sin afectar el puntaje final.

Es posible normalizar la similitud de Smith-Waterman con base en la longitud de la mayor cadena, la longitud de la menor cadena o la longitud media de ambas cadenas [da Silva *et. al.*, 2007], que corresponden al coeficiente de Jaccar, Overlap y Dice respectivamente. Arslan y Eğecioğlu proponen normalizar por la suma de la longitud de las subcadenas A' y B' más similares [Arslan y Eğecioğlu, 2001]. Breimer y Goldberg describen un modelo que puede ser entrenado automáticamente a partir de un conjunto de datos [Breimer y Goldberg, 2002].

La similitud Smith-Waterman puede ser calculada en $O(nm)$ mediante el algoritmo de Smith-Waterman [Smith y Waterman, 1981]. Baeza-Yates y Gonnet presentan un algoritmo que toma $O(n)$ para verificar si la similitud Smith-Waterman entre dos cadenas es menor que cierta constante k [Baeza-Yates y Gonnet, 1992].

3.1.1.4. *Similitud de Jaro.* Jaro desarrolló una función de similitud que define la trasposición de dos caracteres como la única operación de edición permitida [Jaro, 1976]. A diferencia de la distancia de Damerau-Levenshtein, los caracteres no necesitan ser adyacentes, sino que pueden estar alejados cierta distancia d que depende de la longitud de ambas cadenas. Específicamente, dadas dos cadenas A y B es necesario:

Encontrar el número de caracteres comunes c entre A y B; son comunes todos aquellos caracteres a_i y b_j tal que $a_i = b_j$ y $i - d \leq j \leq i + d$, donde $d = min\,(n,m)/2$.

Hallar el número de sustituciones necesarias t entre los caracteres comunes de A y B; si el *i-ésimo* carácter común de A es diferente al *i-ésimo* común de B, se cuenta una sustitución (equivalente a media transposición).

Una vez calculados c y t, la similitud de Jaro entre A y B viene dada por:

$$Jaro(A,B) = \frac{1}{3} \cdot \left(\frac{c}{n} + \frac{c}{m} + \frac{c - (t/2)}{c} \right) \quad (1)$$

Winkler propone una variante que asigna puntajes de similitud mayores a cadenas que comparten algún prefijo, basándose en un estudio realizado por Pollock y Zamora sobre aproximadamente cuarenta mil palabras inglesas extraídas de textos escolares y científicos [Pollock y Zamora, 1984], que muestra que la tasa de errores tipográficos/ortográficos se incrementa monótonamente hacia la derecha [Winkler, 1990]. Cohen *et al.* proponen un modelo basado en distribuciones Gaussianas [Cohen *et. al.*, 2003]. Yancey compara la eficacia de la similitud de Jaro y algunas de sus variantes [Yancey, 2006].

La similitud de Jaro puede ser calculada en $O(n)$.

3.1.1.5. *Similitud de q-grams.* Un q-gram, también llamado n-gram, es una subcadena de longitud q [Yancey, 2006]. El principio tras esta función de similitud es que, cuando dos cadenas son muy similares, tienen muchos q-grams en común. La similitud de q-grams entre dos cadenas A y B viene dada por:

$$q - grams(A,B) = \frac{c_q}{div_q} \quad (2)$$

Donde c_q es el número de *q-grams* comunes entre A y B y div_q es un factor de normalización que puede ser:

- ✓ El número de *q-grams* en la cadena de mayor longitud (coeficiente de Jaccard).
- ✓ El número de *q-grams* en la cadena de menor longitud (coeficiente de Overlap).
- ✓ El número medio de *q-grams* en ambas cadenas (coeficiente de Dice).

Es común usar *uni-grams* (q = 1), *bi-grams* o *di-grams* (q = 2) y *tri-grams* (q = 3) [Yancey, 2006].

Es posible agregar q – 1 ocurrencias de un carácter especial (no definido en el alfabeto original) al principio y final de ambas cadenas. Esto llevará a un puntaje de similitud mayor entre cadenas que compartan algún prefijo y/o sufijo, aunque presenten diferencias hacia el medio [Yancey, 2006]. Una extensión natural a los *q-grams* consiste en utilizar la posición en que ocurren dentro de cada *string* [Sutinen y Tarhio, 1995]. Bajo este modelo, conocido como *q-grams* posicionales, se pueden considerar comunes sólo aquellos *q-*

grams que no estén separados más de cierta distancia, lo cual aumenta la sensibilidad ante el reordenamiento de tokens. Un modelo alternativo se conoce como *k-skip-q-grams*: *q-grams* que omiten *k* caracteres adyancentes. Por ejemplo, el string "Peter" contiene los *bi-grams* "Pe", "et", "te", "er" y los *1-skip-2-grams* "Pt", "ee", "tr" [Yancey, 2006]. Keskustalo *et al.* muestran que el uso de *k-skip-q-grams* mejora la identificación de *strings* equivalentes escritos en diferentes lenguajes [Keskustalo *et. al.*, 2003].

Mediante el uso de funciones hash adecuadas, la similitud de *q-grams* puede ser calculada en $O(n)$ [Ukkonen, 1997], [Cohen, 1997]. Ukkonen presenta un algoritmo alternativo basado en autómatas de sufijos que también toma $O(n)$ [Ukkonen, 1997].

3.1.2. Funciones de similitud basadas en tokens

Estas funciones de similitud consideran cada cadena como un conjunto de subcadenas separadas por caracteres especiales, como por ejemplo espacios en blanco, puntos y comas. Esto es, como un conjunto de *tokens*, y calculan la similitud entre cada pareja de *tokens* mediante alguna función de similitud basada en caracteres. En esta sección se cubren dos de las funciones basadas en *tokens* más comunes: Monge-Elkan y coseno TF-IDF.

3.1.2.1. *Similitud de Monge-Elkan.* Dadas dos cadenas A y B, sean $\alpha_1, \alpha_2 \dots \alpha_K$ y $\beta_1, \beta_2 \dots \beta_L$ sus tokens respectivamente. Para cada token α_i existe algún β_j de maxima similitud. Entonces la similitud de Monge-Elkan entre A y B es la similitud máxima promedio entre una pareja (α_i, β_j) [Monge y Elkan, 1996]. Esto es:

$$Monge - Elkan(A, B) = \frac{1}{K} \cdot \sum_{i=1}^{K} \underset{j = 1 \dots L}{max} \{sim(\alpha_i, \beta_j)\} \quad (3)$$

Donde $sim(\cdots)$ es alguna función de similitud basada en caracteres, comúnmente llamada en este contexto *función de similitud secundaria*. Gelbukh *et al.* presentan un modelo basado en la media aritmética generalizada, en lugar del promedio, el cual supera al modelo original sobre varios conjuntos de datos [Gelbukh *et. al.*, 2009].

La similitud de Monge-Elkan puede ser calculada en $O(nm)$ [Gelbukh *et. al.*, 2009].

3.1.2.2. Similitud coseno TF-IDF. Dadas dos cadenas A y B, sean $\alpha_1, \alpha_2 \dots \alpha_K$ y $\beta_1, \beta_2 \dots \beta_L$ sus tokens respectivamente, que pueden verse como dos vectores V_A y V_B de K y L componentes. Cohen propone una función que mide la

similitud entre A y B como el coseno del ángulo que forman sus respectivos vectores [Cohen, 1998]:

$$\frac{V_A \cdot V_B}{\|VA\|_2 \|VB\|_2} \quad (4)$$

Donde $\cdot$ y $\| \ \|_2$ corresponden a los operadores producto punto y norma *L2*[3] respectivamente.

A cada componente α_i (y análogamente a cada β_j) se le asigna un valor numérico $\gamma(\alpha_i)$ dado por:

$$\gamma(\alpha_i) = log\left(tf_{\alpha_i,A} + 1\right) \cdot log\left(\frac{N}{df_{\alpha_i}} + 1\right) \quad (5)$$

Donde tf_{α_i} es el número de veces que aparece α_i en A y df_{α_i} el número de cadenas, dentro de los N a detectar duplicados, en los cuales aparece por lo menos una vez α_i. Lo anterior produce altos valores de similitud para cadenas que comparten muchos *tokens* poco comunes (con alto poder discriminante).

Como ha sido definida hasta aquí, la similitud coseno TF-IDF no es eficiente bajo la presencia de variaciones a nivel de caracteres, como errores ortográficos o variaciones en el orden de los *tokens*. Por ejemplo, las cadenas "Jorge Eduardo Rodríguez López" y "Rodríguez Jorge Eduardo" tendrían similitud cero. Cohen *et al.* proponen una variante llamada SoftTF-IDF para solucionar este problema, que tiene en cuenta parejas de *tokens* (α_i, β_j) cuya similitud es mayor que cierto umbral (mediante alguna función de similitud basada en caracteres) y no necesariamente idénticos [Cohen *et. al.*, 2003].

3.2. Evaluación de funciones de similitud sobre cadenas de texto

Para medir la eficacia de una función de similitud es necesario:

1) Una colección de cadenas de prueba V.
2) Un conjunto de muestra $Q \subseteq V$; cada cadena $q \in Q$ será usada como un objeto de consulta contra V. El número de cadenas en Q se denota $|Q|$.
3) Para cada cadena $q \in Q$, conocer el conjunto de cadenas $v \in V$ relevantes a q; esto es, el conjunto de cadenas que representan la misma entidad que q. Para esto, las cadenas de V deben estar etiquetadas con algún identificador que compartan aquellas que se refieran a una misma entidad.

[3] Norma euclídea que permite calcular la magnitud de un vector bidimensional.

Las *curvas de precisión y de memoria* son usadas comúnmente para evaluar funciones de similitud sobre cadenas [Heuser, 2007]. Dichas curvas se obtienen a partir de un tipo especial de consultas conocidas como *top-k queries*, en las cuales se retornan las *k* instancias más similares a la cadena buscada. Entonces, para una función de similitud particular, su curva de precisión y de memoria mide la capacidad que tiene para mover las cadenas relevantes a la buscada dentro de los primeros *k* resultados. Sin embargo, en el proceso de detección de duplicados se utilizan *consultas de rango*, en las cuales se retornan todas las cadenas cuya similitud con la buscada es mayor que cierto umbral *t* previamente definido. En este contexto, una técnica de evaluación acertada debe tener en cuenta [Heuser, 2007]:

- ✓ Si la función de similitud consigue separar correctamente las cadenas relevantes de las irrelevantes, asignando puntajes superiores al umbral a las primeras e inferiores al umbral a las últimas.

- ✓ El grado de separación entre cadenas relevantes e irrelevantes. Una buena función de similitud no sólo debe distinguir entre unas y otras, sino ponerlas dentro de una distancia razonable de forma que creen dos conjuntos distintos claramente definidos.

- ✓ La variación del umbral *t* seleccionado, pues la distribución de valores de similitud puede variar de un conjunto de datos a otro.

Las curvas de precisión y de memoria sólo tienen en cuenta la separación de las cadenas relevantes de las irrelevantes. Por esta razón, en el presente trabajo se utiliza una métrica de evaluación propuesta en 2007 por Da Silva *et al.*, llamada *función de discernibilidad* [da Silva *et. al.*, 2007].

3.2.1. Función de discernibilidad

La discernibilidad es una métrica que ha sido utilizada en trabajos comparativos como es el caso de la herramienta SimEval [Heuser *et. al.*, 2007] para determinar cuán eficaces son las funciones de similitud identificando las entradas que realmente corresponden a un mismo objeto. Esta métrica intrínsecamente incorpora los cuatro elementos tradicionales de una tabla de contingencia o matriz de confusión de 2*2: aciertos, desaciertos, falsos positivos y falsos negativos.

La discernibilidad de una función de similitud se calcula como [da Silva *et. al.*, 2007]:

$$\frac{c_1}{c_1 + c_2}\left(t_{max}^{óptimo} - t_{min}^{óptimo}\right) + \frac{c_2}{c_1 + c_2}\left(\frac{F_{max}}{2|Q|}\right) \quad (6)$$

El rango $\left[t_{min}^{óptimo}, t_{max}^{óptimo}\right]$ determina el intervalo de umbrales que mejor separa cadenas relevantes de irrelevantes, pues actúa como indicador de la distancia entre estos (así, entre mayor sea, mejor), y puede ser calculado por cualquiera de los dos algoritmos propuestos en [da Silva *et. al.*, 2007]. El término $F_{max}/2|Q|$ indica el porcentaje de separaciones correctas logrado. Los coeficientes c_1 y c_2 permiten controlar la importancia de cada uno de los dos aspectos anteriores. Independientemente de c_1 y c_2, los valores de discernibilidad siempre caerán en el intervalo [-1,1]: -1 en el peor caso y 1 en el caso de la función de similitud ideal, que logra separar correctamente todas las cadenas relevantes de las irrelevantes a la distancia máxima posible. Para más detalles ver [da Silva *et. al.*, 2007].

3.3. Diseño del Experimento para la comparación de las técnicas

Usando la función de discernibilidad, se compararon nueve funciones de similitud sobre cadenas bajo seis situaciones problemáticas diferentes, detalladas en la Tabla 1. Cada situación corresponde a una variación normalmente encontrada en la representación textual de una misma entidad. En la tabla se coloca entre paréntesis la abreviatura como se identificará esa situación problemática en el resto del documento (Ejemplo: ABR para abreviaturas).

Para las pruebas realizadas se definió c_1 = 3 y c_2 = 7 en (6), de forma que el intervalo óptimo contribuya en 30% al valor de discernibilidad y $F_{max}/2|Q|$ en 70%. Esto por dos razones: primero, para fines prácticos, es mucho más importante que la función de similitud separe correctamente *strings* relevantes de irrelevantes, independientemente de la distancia a la que ubique unos de otros. Segundo, al dar igual importancia a ambos factores (c_1 = 1 y c_2 = 1), una función de similitud podría obtener un valor de discernibilidad considerablemente alto con un intervalo óptimo largo de umbrales que logren un porcentaje de separaciones correctas bajo, lo cual no tiene sentido.

Mediante la herramienta Web FakeNameGenerator[4] se generaron aleatoriamente conjuntos de datos compuestos por registros con atributos como nombre, apellidos, dirección, ocupación y correo electrónico. A partir de estos registros, fueron derivados otros de acuerdo con la variación textual o situación problemática a probar. Así, para la situación problemática ERR (Errores ortográficos), se generaron nuevas cadenas a las cuales se les introdujeron errores de ortografía cambiando unas letras por otras (por

[4] http://www.fakenamegenerator.com/

ejemplo *g* por *j*, *v* por *b*, *c* por *s*, entre otras). Para ABR (Abreviaturas), se generaron nuevas cadenas a las cuales se les recortaron los segundos nombres, segundos apellidos o las ocupaciones, dejando sólo la inicial o primera letra de estos. Para TFL (Tokens Faltantes), se generaron nuevas cadenas a las cuales se les omitió una o varias palabras. Para PSF (prefijos/sufijos), se generaron nuevas cadenas a las cuales se antepuso o pospuso al nombre completo de la persona, textos de diferentes longitudes como "Doctor", "Magíster", "Estudiante", "Universidad Nacional de Colombia", entre otros. Para TDR (Tokens en desorden), se cambió al nombre completo el orden de las palabras colocando primero los dos apellidos y luego los nombres y al oficio también se le cambió el orden de las palabras. Para ESP (Espacios en Blanco), se agregaron/eliminaron espacios en blanco entre las palabras.

Tabla 1. Situaciones Problemáticas para comparación de funciones de similitud

Situación Problemática	Ejemplo
Errores ortográficos (ERR)	"Jorge Eduardo Rodríguez López" vs. "Jorje Eduadro Rodrígues Lopes"
Abreviaturas: truncamiento de uno o más *tokens* (ABR)	"Jorge Eduardo Rodríguez López" vs. "Jorge E Rodríguez L"
Tokens faltantes: eliminación de uno o más tokens (TFL)	"Jorge Eduardo Rodríguez López" vs. "Jorge Rodríguez"
Prefijos/sufijos sin valor semántico: presencia de subcadenas al principio y/o al final (PSF)	"Jorge Eduardo Rodríguez López" vs. "PhD Jorge Eduardo Rodríguez López, U Nal"
Tokens en desorden (TDR)	"Jorge Eduardo Rodríguez López" vs. "Rodríguez López Jorge Eduardo"
Espacios en blanco: eliminación o adición de espacios en blanco (ESP)	"Jorge Eduardo Rodríguez López" vs. "JorgeEduardo Rodríguez López"

Se generaron diez conjuntos de datos de prueba para cada situación problemática con las respectivas variaciones textuales. Cada conjunto de datos tiene 400 cadenas de texto representando 200 entidades (dos cadenas por entidad). Entonces se comparó para cada situación problemática, la eficacia de las siguientes funciones de similitud: distancia de Levenshtein, distancia de brecha afín, similitud Smith-Waterman, similitud de Jaro y Jaro-Winkler, similitud de *bi-grams* y *tri-grams*, similitud de Monge-Elkan y similitud SoftTF-IDF.

Vale la pena recalcar, que uno de los aportes originales de este trabajo es precisamente la evaluación de las diferentes funciones de similitud a la luz de esas diferentes situaciones problemáticas definidas, utilizando varios conjuntos de datos especialmente diseñados para probar un tipo de variación en el texto. Adicionalmente, también se comparó la eficacia de las anteriores funciones de similitud sobre dos conjuntos de datos reales que combinan varias situaciones problemáticas, extraídos de la base datos Scienti[5]. En ésta se registra gran parte de la actividad investigativa en Colombia, incluyendo: grupos de investigación, investigadores, proyectos en curso y proyectos finalizados. En ocasiones el título de un mismo proyecto se registra – en texto libre – varias veces (por ejemplo, una vez por cada participante). De igual forma, una misma persona registra – de nuevo, en texto libre – varias veces su nombre (por ejemplo, una vez por cada proyecto en el que participe). Así, es muy probable que una misma entidad, sea título de proyecto o nombre de investigador, aparezca representada varias veces de distintas formas.

El primer conjunto de datos se formó con 642 nombres personales representados en 1284 cadenas de texto diferentes, extraídas y etiquetadas manualmente; cada nombre es representado de dos formas distintas. De igual forma, el segundo conjunto de datos se formó con 303 títulos de proyectos representados en 606 cadenas diferentes; cada título es representado de dos formas distintas. Las variaciones más frecuentes encontradas en ambos conjuntos de datos fueron: i) Errores ortográficos/tipográficos, ii) Abreviaturas: en el primer conjunto de datos, usualmente de un segundo nombre y/o apellido, como por ejemplo “Jorge Eduardo Rodríguez López” vs. “Jorge E. Rodríguez L.” y iii) *Tokens* faltantes: en el primer conjunto de datos, usualmente la ausencia de un segundo nombre o apellido. En el segundo conjunto de datos, la misión de una o más palabras, como por ejemplo “Antimaláricos en artritis reumatoidea” vs. “Antimaláricos en el tratamiento de la artritis reumatoidea.”

3.4. Resultados de la comparación de las funciones de similitud sobre cadenas de texto

Las Tablas 2 a 7 muestran los resultados de la discernibilidad al aplicar las diferentes funciones de similitud sobre los diez conjuntos de datos de cada situación problemática.

[5] http://thirina.colciencias.gov.co:8081/scienti/

Tabla 2. Discernibilidad situación problemática: Abreviaturas

	FUNCIÓN DE SIMILITUD								
Conjunto de datos	**Levensh Tein**	**Affine Gap**	**Smith Waterman**	**Jaro**	**Jaro Winkler**	**2-grams**	**3-grams**	**Monge Elkan**	**Soft TF-IDF**
1	0,5737	0,7180	0,6995	0,4258	0,3579	0,6070	0,6632	0,6802	0,7060
2	0,5965	0,7090	0,7025	0,4719	0,4246	0,6193	0,6662	0,6737	0,6965
3	0,5772	0,7210	0,6972	0,4860	0,3509	0,6088	0,6561	0,6737	0,6982
4	0,5930	0,7090	0,6995	0,6995	0,3807	0,6263	0,6579	0,6702	0,7000
5	0,6053	0,7150	0,7030	0,4439	0,3544	0,6175	0,6509	0,6725	0,6999
6	0,5825	0,7180	0,6930	0,4807	0,4035	0,6035	0,6609	0,6719	0,6990
7	0,5860	0,7180	0,6930	0,4211	0,3451	0,6065	0,6544	0,6754	0,6995
8	0,5877	0,7180	0,6925	0,4509	0,4263	0,6123	0,6439	0,6702	0,6982
9	0,5965	0,7240	0,7060	0,4667	0,4088	0,6333	0,6719	0,6807	0,7030
10	0,5965	0,7180	0,7060	0,4351	0,3193	0,6140	0,6609	0,6837	0,6965
Promedio	**0,5895**	**0,7168**	**0,6992**	**0,4781**	**0,3771**	**0,6149**	**0,6586**	**0,6752**	**0,6997**

Tabla 3. Discernibilidad situación problemática: Prefijos/Sufijos

	FUNCIÓN DE SIMILITUD								
Conjunto de datos	**Levensh tein**	**Affine Gap**	**Smith Waterman**	**Jaro**	**Jaro Winkler**	**2-grams**	**3-grams**	**Monge Elkan**	**Soft TFIDF**
1	0,7210	0,7360	0,7090	0,6579	0,5035	0,7210	0,7360	0,7150	0,7720
2	0,7120	0,7330	0,7300	0,6228	0,4368	0,7060	0,7240	0,7180	0,7690
3	0,7030	0,7360	0,7127	0,6596	0,4316	0,7270	0,7270	0,7090	0,7690
4	0,7060	0,7390	0,7270	0,6807	0,4596	0,7090	0,7270	0,7120	0,7750
5	0,7090	0,7330	0,7000	0,6772	0,4912	0,7210	0,7330	0,7180	0,7630
6	0,7240	0,7330	0,7000	0,6644	0,5105	0,7120	0,7300	0,7180	0,7870
7	0,7090	0,7330	0,7265	0,6614	0,4907	0,7085	0,7210	0,7030	0,7540
8	0,7120	0,7360	0,7180	0,6596	0,4789	0,7150	0,7300	0,7180	0,7810
9	0,7150	0,7330	0,7120	0,6719	0,5263	0,7120	0,7360	0,7090	0,7630
10	0,7210	0,7420	0,7240	0,6579	0,4737	0,7240	0,7390	0,7180	0,7750
Promedio	**0,7132**	**0,7354**	**0,7159**	**0,6614**	**0,4803**	**0,7155**	**0,7303**	**0,7138**	**0,7708**

Tabla 4. Discernibilidad situación problemática: Tokens en desorden

	FUNCIÓN DE SIMILITUD								
Conjunto de datos	**Levenshtein**	**Affine Gap**	**Smith Waterman**	**Jaro**	**Jaro Winkler**	**2-grams**	**3-grams**	**Monge Elkan**	**Soft TF-IDF**
1	0,6985	0,7300	0,7210	0,7360	0,7090	0,8198	0,8380	0,7370	0,7730
2	0,6860	0,7240	0,7090	0,7360	0,7150	0,8020	0,8260	0,7450	0,7660
3	0,6890	0,7240	0,7030	0,7330	0,7090	0,7960	0,8200	0,7510	0,7720
4	0,6841	0,7210	0,7090	0,7360	0,7120	0,8080	0,8290	0,7510	0,7720
5	0,6930	0,7270	0,7060	0,7330	0,7090	0,7930	0,8200	0,7510	0,7720
6	0,6860	0,7300	0,7090	0,7360	0,7150	0,8140	0,8320	0,7510	0,7750
7	0,6995	0,7300	0,7210	0,7360	0,7090	0,8200	0,8380	0,7390	0,7720
8	0,6960	0,7330	0,7300	0,7360	0,7120	0,8080	0,8350	0,7510	0,7750
9	0,6995	0,7360	0,7210	0,7360	0,7120	0,8170	0,8410	0,7510	0,7660
10	0,6960	0,7240	0,7060	0,7346	0,7120	0,8200	0,8530	0,7570	0,7780
Promedio	**0,6927**	**0,7279**	**0,7135**	**0,7353**	**0,7114**	**0,8098**	**0,8332**	**0,7484**	**0,7721**

Tabla 5. Discernibilidad situación problemática: Tokens faltantes

	FUNCIÓN DE SIMILITUD								
Conjunto de datos	**Levenshtein**	**Affine Gap**	**Smith Waterman**	**Jaro**	**Jaro Winkler**	**2-grams**	**3-grams**	**Monge Elkan**	**Soft TF-IDF**
1	0,7060	0,7390	0,7690	0,6228	0,4333	0,7150	0,7420	0,7120	0,7150
2	0,7150	0,7360	0,7540	0,6263	0,4193	0,7060	0,7270	0,7060	0,7060
3	0,7300	0,7330	0,7450	0,6193	0,4684	0,7180	0,7450	0,7090	0,7150
4	0,7210	0,7330	0,7570	0,5930	0,4684	0,7360	0,7480	0,7090	0,7090
5	0,7145	0,7270	0,7210	0,6193	0,4544	0,7145	0,7150	0,7060	0,7030
6	0,7180	0,7360	0,7540	0,6298	0,4930	0,7120	0,7420	0,7060	0,7150
7	0,7210	0,7330	0,7810	0,5719	0,4649	0,7180	0,7390	0,7012	0,7090
8	0,7240	0,7360	0,7720	0,6211	0,4930	0,7120	0,7300	0,7060	0,7180
9	0,7300	0,7360	0,7870	0,6105	0,4714	0,7360	0,7630	0,7090	0,7150
10	0,7360	0,7360	0,7630	0,6140	0,4579	0,7420	0,7720	0,7210	0,7210
Promedio	**0,7215**	**0,7345**	**0,7603**	**0,6128**	**0,4624**	**0,7209**	**0,7423**	**0,7085**	**0,7126**

Tabla 6. Discernibilidad situación problemática: Espacios en blanco

Conjunto de datos	Levenshtein	Affine Gap	Smith Waterman	Jaro	Jaro Winkler	2-grams	3-grams	Monge Elkan	Soft TFIDF
1	0,7600	0,7480	0,8140	0,7180	0,6965	0,7600	0,7840	0,7270	0,7300
2	0,7540	0,7420	0,7990	0,7180	0,7025	0,7390	0,7690	0,7240	0,7360
3	0,7540	0,7420	0,7930	0,7210	0,6995	0,7480	0,7810	0,7240	0,7270
4	0,7570	0,7420	0,7990	0,7150	0,7000	0,7570	0,7780	0,7240	0,7300
5	0,7330	0,7390	0,7630	0,7180	0,6965	0,7300	0,7570	0,7240	0,7270
6	0,7540	0,7390	0,7780	0,7210	0,7000	0,7570	0,7780	0,7270	0,7300
7	0,7660	0,7450	0,7990	0,7120	0,6930	0,7480	0,7780	0,7270	0,7390
8	0,7540	0,7420	0,8020	0,7210	0,7000	0,7570	0,7810	0,7300	0,7390
9	0,7600	0,7480	0,8020	0,7180	0,7030	0,7600	0,7870	0,7240	0,7330
10	0,7660	0,7480	0,8140	0,7150	0,7030	0,7660	0,7930	0,7270	0,7330
Promedio	**0,7558**	**0,7435**	**0,7963**	**0,7177**	**0,6994**	**0,7522**	**0,7786**	**0,7258**	**0,7324**

Tabla 7. Discernibilidad situación problemática: Errores Ortográficos y tipográficos

Conjunto de datos	Levenshtein	Affine Gap	Smith Waterman	Jaro	Jaro Winkler	2-grams	3-grams	Monge Elkan	Soft TFIDF
1	0,7631	0,7062	0,7484	0,7151	0,7062	0,7121	0,7123	0,6952	0,6911
2	0,7123	0,6991	0,6964	0,6852	0,6823	0,7002	0,6942	0,6895	0,6825
3	0,7570	0,7120	0,7270	0,6860	0,6719	0,6965	0,6965	0,6960	0,7060
4	0,8110	0,7450	0,8020	0,7090	0,6995	0,7780	0,7840	0,7270	0,6942
5	0,7930	0,7330	0,8020	0,7030	0,7000	0,7900	0,7930	0,7330	0,7750
6	0,7510	0,6995	0,7030	0,6955	0,6930	0,7240	0,7060	0,6895	0,6772
7	0,7570	0,7060	0,7180	0,6930	0,6995	0,6995	0,7145	0,7000	0,6920
8	0,7572	0,7120	0,7275	0,6862	0,6719	0,6965	0,6968	0,6961	0,7060
9	0,8111	0,7450	0,8020	0,7091	0,6984	0,7782	0,7842	0,7274	0,6942
10	0,7929	0,7331	0,8018	0,7034	0,7017	0,7903	0,7932	0,7333	0,7744
Promedio	**0,7706**	**0,7706**	**0,7191**	**0,7528**	**0,6986**	**0,6924**	**0,7365**	**0,7375**	**0,7087**

Para las seis situaciones problemáticas, las funciones de similitud arrojaron resultados variables de la discernibilidad. Para poder llegar a conclusiones válidas, se quiso realizar un análisis de varianza (ANOVA) con el fin de determinar si las diferencias encontradas con las distintas técnicas eran estadísticamente significativas, según el índice de discernibilidad, usando las distintas técnicas, pero luego de realizar la prueba de Kolmogorov-Smirnov se determinó que el supuesto de normalidad no se cumple. Por ello, se aplicó la prueba no paramétrica de Kruskal-Wallis, la cual no requiere que se cumplan los supuestos de normalidad y homoscedasticidad [Álvarez, 2007]. Esta prueba permite verificar la hipótesis nula de igualdad de las medianas del grado de

discernibilidad para las nueve funciones de similitud. A continuación se incluyen los resultados arrojados por el paquete estadístico StatGraphics versión 5.0 para la situación problemática ABR.

```
Contraste de Kruskal-Wallis para Discernibilidad Abreviaturas según Función de Similitud

Función de SimilitudTamaño muestral     Rango Promedio
------------------------------------------------------------
Bigrams             10                  34,4
Brecha Afín         10                  85,5
Jaro                10                  20,25
Jaro Winkler        10                  5,8
Levenshtein         10                  24,6
Monge Elkan         10                  54,25
Smith Waterman      10                  69,85
SoftTFIDF           10                  70,1
Trigrams            10                  44,75
------------------------------------------------------------
Estadístico = 82,7995   P-valor = 0,0
```

La prueba de Kruskal-Wallis arrojó como resultado para todas las situaciones problemáticas un valor *p* inferior a 0,05, pudiéndose afirmar que hay diferencias estadísticamente significativas entre las medianas, con un nivel de confianza del 95%. La columna de la derecha corresponde al rango promedio y al ordenar por su valor puede establecerse que tan exitosas son las funciones de similitud en cuanto a la discernibilidad. Se construyeron Gráficos de Cajas y Bigotes, como ayuda visual en cuanto a las medianas significativamente diferentes. La figura 1, presenta los diagramas de las seis situaciones problemáticas.

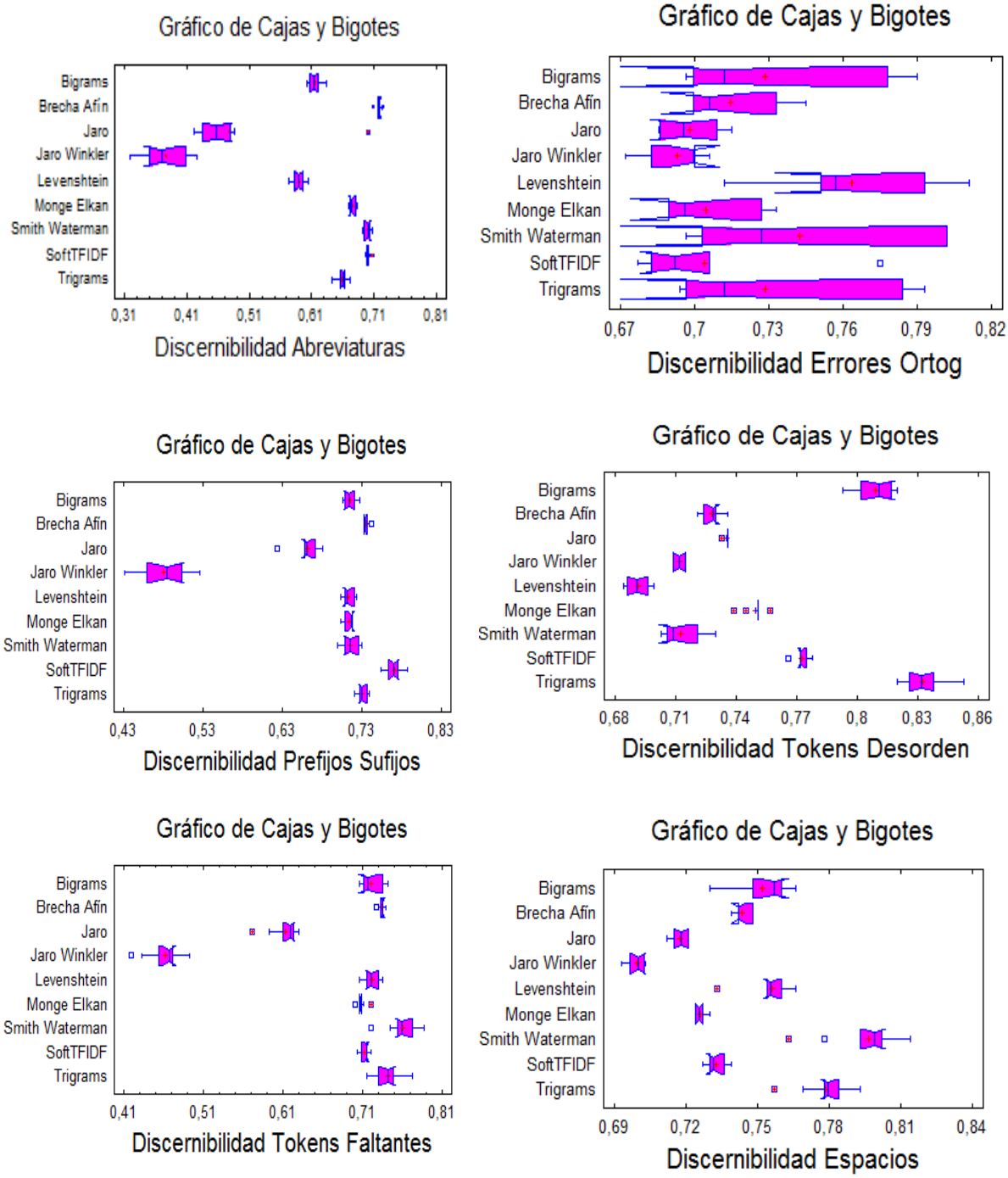
El análisis de las gráficas, especialmente del punto central de cada caja, permite identificar a las funciones de similitud de mejores resultados en cada caso. La tabla 8 presenta las posiciones obtenidas por las funciones según su eficacia para cada situación problemática.

Tabla 8. Resultados eficacia por situación problemática

Situación Problemática	**Posición**								
	1	2	3	4	5	6	7	8	9
Errores ortográficos	LE	SW	2G	3G	BA	ME	JA	SW	JW
Abreviaturas	BA	ST	SW	ME	3G	2G	LE	JA	JW
Tokens Faltantes	SW	3G	BA	LE	2G	SW	ME	JA	JW
Prefijos / Sufijos	ST	BA	3G	SW	2G	ME	LE	JA	JW
Tokens en desorden	3G	2G	ST	ME	JA	BA	JW	SW	LE
Espacios en blanco	SW	3G	LE	2G	BA	SW	ME	JA	JW

BA: Brecha Afín ST: Soft TF-IDF SW: Smith-Waterman LE: Levenshtein 2G: Bi-grams 3G: Tri-grams ME: Monge-Elkan JA: Jaro JW: Jaro Winkler

Figura 1. Gráficos de Cajas y Bigotes para situaciones problemáticas.



La tabla 9, en cada fila presenta las posiciones obtenidas pero vistas por función de similitud, lo cual permite observar que tan buena es una función en las diferentes situaciones problemáticas. Las dos últimas columnas corresponden a la media aritmética y la desviación estándar de las posiciones ocupadas por cada una de las funciones de similitud. Nótese como un valor promedio bajo acompañado de una desviación estándar baja, es indicativo de que la función obtuvo buenos resultados, no sólo para una función problemática, sino que tuvo un buen comportamiento general. Obsérvese como la función Tri-grams obtuvo la mejor posición promedio (2.8) y su desviación estándar es moderada (1.47), lo que significa que para todas las funciones obtuvo posiciones no muy alejadas del promedio. Tri-grams, es seguida por Smith Waterman con un promedio de 3.2 una desviación estándar de 2.64, esto es, un promedio mayor y valores menos uniformes. Asimismo, la función Jaro-Winkler, obtuvo el promedio más alto (8.7) con una desviación estándar baja (0.82), lo cual significa que logró resultados uniformemente desfavorables para todas las situaciones problemáticas.

Tabla 9. Resultados Eficacia por Función de Similitud

	Situación Problemática							
Función Similitud	ERR	ABR	TFL	PSF	TDR	ESP	Prom	Desv
LevenshTein	1	7	4	7	9	3	5.2	2.99
Brecha Afín	5	1	3	2	6	5	3.7	1.97
Bi-grams	3	6	5	5	2	4	4.2	1.47
Tri-grams	4	5	2	3	1	2	2.8	1.47
Jaro	7	8	8	8	5	8	7.3	1.21
Jaro Winkler	9	9	9	9	7	9	8.7	0.82
Smith Waterman	2	3	1	4	8	1	3.2	2.64
Monge Elkan	6	4	7	6	4	7	5.7	1.37
Soft TF-IDF	8	2	6	1	3	6	4.3	2.73

La tabla 10 presenta los resultados de la discernibilidad obtenida por las diferentes funciones de similitud sobre los conjuntos de datos reales extraídos de la base de datos Scienti, en los cuales los principales problemas observados son errores ortográficos/tipográficos, abreviaturas y *tokens* faltantes. En esta ocasión, puede verse que las diferencias en la discernibilidad de las funciones de similitud son menos marcadas, esto es, la eficacia de las funciones es relativamente similar, pudiéndose pensar que en presencia de varias situaciones problemáticas en forma simultánea la eficacia de las funciones tiende a confundirse. Esto es, en esta situación es prácticamente indiferente

cual función se utilice y la selección de la técnica puede obedecer más al criterio de eficiencia computacional.

Tabla 10. Resultados de la discernibilidad para los conjuntos de datos extraídos de Scienti.

	Nombres Personales	Títulos de Proyectos
Levenshtein	0,608	0,706
Brecha afín	0,666	0,698
Smith-Waterman	0,619	0,693
Jaro	0,570	0,677
Jaro-Winkler	0,567	0,659
Bi-grams	0,614	0,659
Tri-grams	0,606	0,695
Monge-Elkan	0,667	0,695
SoftTF-IDF	0,656	0,692

3.5. Guía Metodológica para la selección de técnicas para la detección de duplicados.

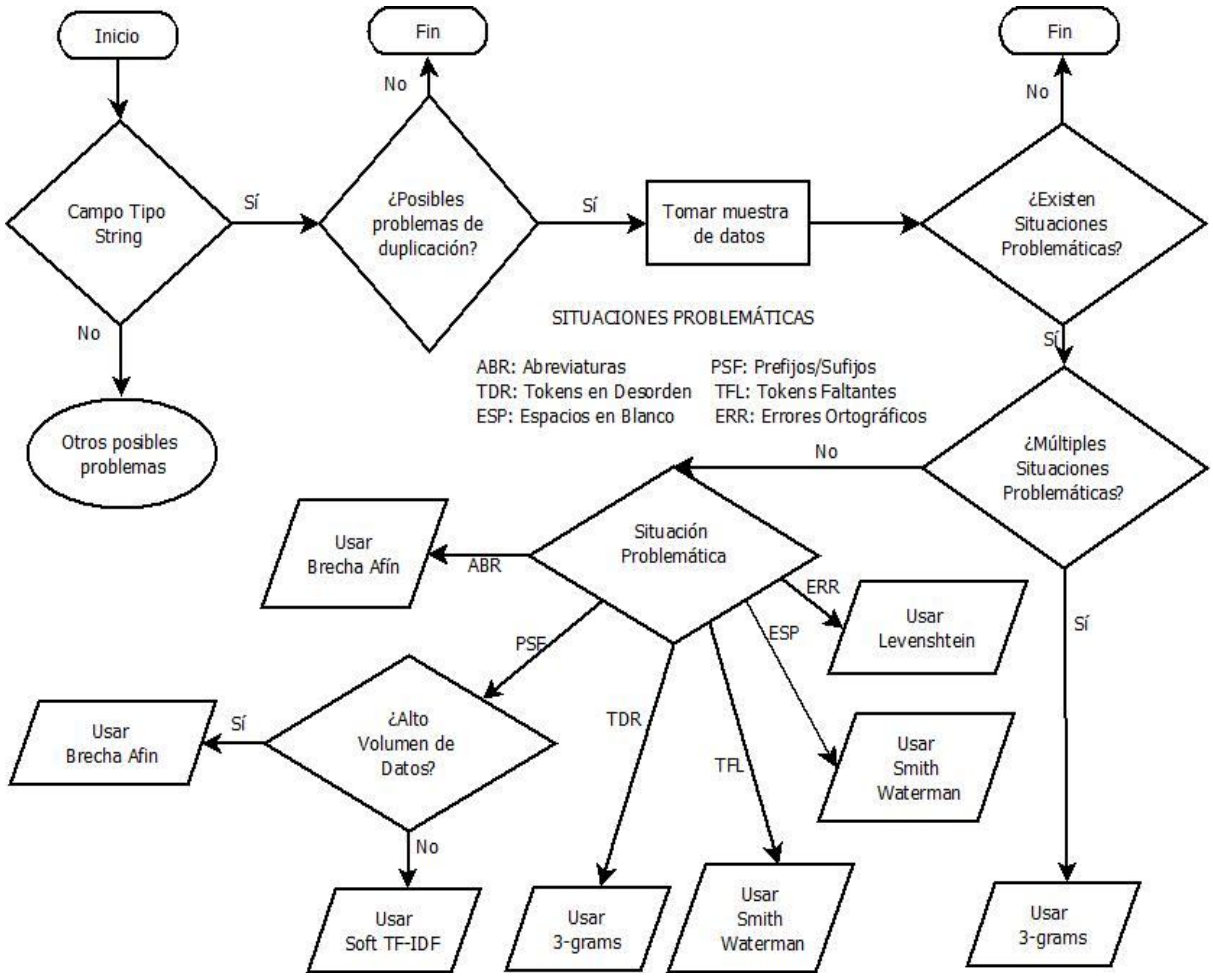
El principal aporte de esta tesis de maestría, es proveer guías que orienten a los analistas de datos en la selección de las técnicas más apropiadas para la situación particular que pueda presentar un cierto conjunto de datos. La figura 2 presenta la guía para el problema de la detección de duplicados, mediante un diagrama de flujo de datos.

El diagrama comienza indagando si el tipo de datos de una columna dada a la cual se desee hacer limpieza, es de tipo *String.* Aunque la duplicación no es un problema exclusivo de este tipo de datos, las técnicas están diseñadas para ser aplicadas a cadenas de texto. Otros tipos de datos a los cuales se desee hacer análisis de duplicación, requerirán ser convertidos previamente y por tanto se vuelve a la condición inicial. Para campos que tengan otros tipos de datos, se requeriría buscar otros tipos de problemas (por ejemplo: valores atípicos para campos numéricos).

Para campos tipo texto, se interroga sobre la existencia de posibles problemas de duplicación. Aunque aparentemente es una pregunta difícil de responder por parte de un usuario, realmente no lo es tanto. Para un usuario conocedor de los datos y que entienda el concepto de detección de duplicados, no es difícil prever si sus datos son susceptibles a esta situación. Recuérdese que se habla de detección de duplicados cuando el contenido de un campo o de un

registro completo, aparece dos o más veces (duplicado) con diferencias textuales en sus valores y no se tiene un identificador único. Si por ejemplo, en una tabla se almacenan datos de proyectos de investigación, en un proyecto pueden participar varios investigadores y no existe un identificador único como un código que se le asigne previamente a cada proyecto, es fácil que cada investigador entre el título del proyecto con alguna variación en el texto, haciendo que pueda considerarse como un proyecto distinto. Otro ejemplo, podría ser aquella situación en la cual se almacenan pedidos de clientes en una hoja de cálculo, sin contar con un identificador único para cada cliente sino haciéndolo a través de sus nombres. Como un cliente puede colocar varios pedidos, es factible que los nombres no sean ingresados exactamente igual.

Figura 2. Diagrama de flujo para guiar la selección de técnicas para detección de duplicados.



En caso de existir posibles problemas de duplicados, se sugiere tomar una muestra de datos para tomar decisiones con base en ella. Esto es necesario si se tiene un alto volumen de datos, pues de lo contrario se analiza todo el conjunto de datos. Además, la muestra debe ser representativa de la totalidad de los datos.

Sobre la muestra de datos, debe examinarse si se detecta alguna(s) de las seis situaciones problemáticas, es decir, si en los datos se visualizan errores ortográficos, abreviaturas, *tokens* faltantes o en desorden, presencia de prefijos y sufijos o espacios en blanco adicionales o suprimidos. En caso de que detecte una o más de estas situaciones problemáticas, debe establecerse si se perciben varias de ellas simultáneamente o si hay un alto predominio de sólo una de ellas. En caso de predominar una sola situación problemática, se recomienda la técnica de mayor eficacia según la tabla 8, excepto para el caso de PSF en el que la función recomendada dependerá del volumen de datos (la técnica Soft TF-Idf es muy pesada computacionalmente y por tanto se recomienda sólo para un volumen bajo de datos). En caso de observarse en los datos varias situaciones problemáticas simultáneamente, se recomienda la técnica tri-grams ya que es de bajo costo computacional y obtuvo la mejor posición promedio y desviación estándar moderada de todas las funciones (ver tabla 9).

3.6. Conclusiones y Trabajo Futuro sobre la Detección de Duplicados

Existen diversas funciones de similitud sobre cadenas de texto y algunas son más eficaces detectando ciertas situaciones problemáticas o variaciones textuales que otras. Decidir cuál función utilizar en un caso particular, es un asunto complejo, el cual puede ser facilitado a los usuarios mediante una guía metodológica que recomiende la función adecuada para su situación.

Este objetivo se logra en el presente trabajo, mediante el cumplimiento de los objetivos específicos planteados, así:

- ✓ Se logró identificar nueve técnicas para detección de duplicados (distancia de Levenshtein, distancia de brecha afín, similitud Smith-Waterman, similitud de Jaro y Jaro-Winkler, similitud de *bi-grams* y *tri-grams*, similitud de Monge-Elkan y similitud SoftTF-IDF).

- ✓ Se identificaron las características de cada técnica determinando sus propiedades, funcionalidad, fortalezas y debilidades.

- ✓ Se compararon las técnicas determinando las funciones más eficaces en cada caso (Levenshtein para Errores Ortográficos, Brecha Afín para Abreviaturas, Smith-Waterman para *Tokens* Faltantes, Soft TFIDF para Prefijos y Sufijos, Tri-grams para *Tokens* en desorden y Smith-Waterman para Espacios en blanco). Para esto fue necesario:

    - o Proponer situaciones problemáticas que permitieran determinar la eficacia de las funciones (errores ortográficos, abreviaturas, *tokens* faltantes, prefijos y sufijos, *tokens* en desorden y espacios en blanco).
    - o Identificar una métrica para realizar la comparación (discernibilidad).
    - o Construir conjuntos de datos especialmente diseñados para las seis situaciones problemáticas planteadas.
    - o Calcular la discernibilidad arrojada por cada técnica en cada situación problemática.
    - o Hacer el tratamiento estadístico correspondiente que permitió obtener conclusiones válidas.

- ✓ Se diseñó un diagrama que sirve como guía al usuario para la selección de la técnica más adecuada según la naturaleza de los datos a depurar.

Como trabajo futuro, se plantea incorporar a la guía otrás técnicas no analizadas.

4. CORRECCIÓN DE VALORES FALTANTES

Según Oliveira *et. al.*, los datos faltantes hacen referencia a la ausencia de un valor para un atributo requerido, esto es, sólo puede hablarse de datos faltantes cuando se trata de la ausencia de un valor en un atributo obligatorio y por tanto no todo campo vacío es realmente un problema de calidad en los datos [Oliveira *et. al.*, 2005].

La ausencia de datos necesarios es una situación frecuente. Según Juster y Smith "la ausencia de información y la presencia de datos errados son un mal endémico en las ciencias sociales y el análisis económico" [Juster y Smith, 1998]. Farhangfar *et.al.*, afirman que muchas de las bases de datos industriales y de investigación están plagadas por un problema inevitable de datos incompletos (valores ausentes o faltantes) [Farhangfar *et. al.*, 2007]. Entre las razones para esto, se encuentran procedimientos imperfectos de captura de datos en forma manual, mediciones incorrectas, errores en los equipos y migraciones entre diferentes aplicaciones. En muchas áreas de aplicación, no es raro encontrar bases de datos que tienen 50% o más de sus datos faltantes. En [Lakshminarayan *et. al.*, 1999] se documenta un caso de una base de datos de mantenimiento industrial mantenida por Honeywell con más del 50% de los datos faltantes. En [Kurgan *et. al.*, 2005] se presenta una base de datos médica de pacientes con fibrosis cística con más de 60% de sus datos ausentes.

El conocimiento para tratar con datos incompletos se ha incrementado en campos como mercadeo [Kaufman, 1998], educación [Poirier y Rudd, 1983], economía [Johnson, 1989], psicometría [Browne, 1983], medicina [Berk, 1987], enfermería [Musil *et. al.*, 2002] y muchos otros.

Los datos faltantes son un problema real, pero no siempre se aprecia en su justa medida. Medina y Galván advierten sobre la falta de consciencia por parte de los usuarios y aún de muchos investigadores, sobre las implicaciones estadísticas que conlleva trabajar con datos faltantes o aplicar procedimientos de imputación o sustitución de información deficientes: "La aplicación de procedimientos inapropiados de sustitución de información introduce sesgos y reduce el poder explicativo de los métodos estadísticos, le resta eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones del estudio" [Medina y Galván, 2007]. En el mismo sentido Acock, afirma que los procedimientos de imputación que se utilizan con mayor frecuencia limitan o sobredimensionan el poder explicativo de los modelos y generan estimadores sesgados que distorsionan las relaciones de causalidad entre las variables, generan subestimación en la varianza y alteran el valor de los coeficientes de correlación [Acock, 2005].

Para tratar con datos ausentes, existen varias alternativas. Tan *et. al.* consideran como estrategias posibles trabajar con los datos completos, trabajar con los datos disponibles o reconstruir las observaciones a través de la imputación[Tan *et. al.*, 2006] [Medina y Galván, 2007].

A trabajar con los datos completos se le conoce como *Listwise Deletion* (LD) y debido a la facilidad de llevar a cabo esta estrategia, es habitual que los paquetes estadísticos trabajen —por defecto— sólo con información completa (*listwise*), a pesar de que se reconoce que esta práctica no es la más apropiada [Kalton y Kasprzyk, 1982] [UCLA, 2009]. LD significa trabajar únicamente con las observaciones que disponen de información completa para todos los atributos o variables y tiene serias desventajas, como descartar una cantidad considerable de información, lo cual en cantidades pequeñas de datos es más grave aún. Adicionalmente, si las observaciones completas no son una submuestra al azar de los datos originales, LD puede introducir sesgos en los coeficientes de asociación y de correlación y por lo tanto, realizar un modelo de predicción bajo estas circunstancias puede ser engañoso [Myrtveit *et. al.*, 2001].

Otra alternativa es trabajar con información disponible (*Available-case* (AC)). El procedimiento AC, en contraste con LD, utiliza distintos tamaños de muestra, por lo que también se le conoce como *pairwise deletion* (PD) o *pairwise inclusion*. La tabla 11 corresponde a un ejemplo de Medina y Galván e ilustra la diferencia de este enfoque con LD [Medina y Galván, 2007].

Tabla 11. Análisis con los datos disponibles (*Pairwise Deletion*)

Folio	Sexo	Edad	Escolaridad	Salario	Ocupación	Ponderación
1	Mujer	40	16	4500	2	50
2	Hombre	35	15	?	1	75
3	Mujer	65	?	1200	1	100
4	Hombre	23	12	2200	2	80
5	Hombre	25	?	?	3	250
6	Mujer	38	15	1800	4	140
…						

Fuente: [Medina y Galván, 2007].

En la tabla 11 se observa información completa para distintos registros de las variables salario y escolaridad, por lo que es posible calcular la correlación entre ambas utilizando los folios 1, 4 y 6, en tanto que la relación entre el salario y la ocupación se podría determinar con los datos de los registros 1, 3, 4 y 6. Sin embargo, debido a la diferencia en el tamaño de muestra, no es posible comparar los valores de los coeficientes obtenidos por ambos procedimientos. Este método hace uso de toda la información disponible sin efectuar ningún tipo de corrección en los factores de expansión. Las observaciones que no tienen datos se eliminan, y los cálculos se realizan con

diferentes tamaños de muestra lo que limita la comparación de resultados [Medina y Galván, 2007]. Debido a los inconvenientes de LD y PD, surge el interés por métodos alternativos para datos ausentes como la imputación, que consiste en llenar los valores faltantes con un valor estimado mediante alguna técnica específica.

Aunque la imputación tiene sus detractores por considerar que se están "inventando datos", es ampliamente reconocida entre la comunidad estadística como un método, que bien aplicado, puede mejorar la calidad de los datos. En la literatura se han reportado diferentes técnicas de imputación, pero como lo dicen Useche y Mesa "se debe evitar la no respuesta en la medida posible, para usar imputación sólo cuando sea absolutamente necesario, pues nunca unos datos imputados serán mejores que unos datos reales" [Useche y Mesa, 2006].

Autores como Medina y Galván [Medina y Galván, 2007] y Cañizares *et. al.* [Cañizares *et. al.*, 2004], coinciden en que antes de analizar la técnica de imputación a aplicar sobre un conjunto de datos, es necesario identificar primero el mecanismo o patrón que describe la distribución de los datos faltantes y para ello utilizan la clasificación hecha por Little y Rubin la cual se basa en la aleatoriedad con que se distribuyen los valores faltantes [Little y Rubin, 1987]. Estos autores definen tres tipos de patrones: datos ausentes completamente al azar (*Missing Completely At Random,* MCAR), datos ausentes al azar (*Missing At Random,* MAR) y datos ausentes no al azar (*Missing Not At Random,* MNAR) [Medina y Galván, 2007]. Dada las implicaciones del tema sobre la guía metodológica objetivo de esta tesis de maestría, a continuación se explican en detalle los tres patrones:

Los datos siguen un patrón MCAR cuando los objetos –tuplas o registros, para el caso de una tabla en una base de datos relacional- con los datos completos son similares a los de los datos incompletos; es decir, los objetos con datos incompletos constituyen una muestra aleatoria simple de todos los sujetos que conforman la muestra. Pensando los datos como una gran matriz, los valores ausentes están distribuidos aleatoriamente a través de la matriz. Sirva de ejemplo, una encuesta nacional en la cual se necesitan estudios costosos, como los electrocardiogramas; podría entonces seleccionarse una submuestra mediante muestreo aleatorio simple de los encuestados, para que se aplique este examen [Medina y Galván, 2007].

El patrón MAR se presenta cuando los objetos con información completa difieren del resto. Los patrones de los datos faltantes se pueden predecir a partir de la información contenida en otras variables –atributos o campos- y no de la variable que está incompleta. Un ejemplo de MAR lo presentan Useche y Mesa [Useche y Mesa, 2006]: En un estudio de depresión maternal, 10% o más de las madres puede negarse a responder preguntas acerca de su nivel de depresión. Supóngase que el estudio incluye el estado de pobreza, el cual toma los valores *1* para *pobreza* y *0* para *no pobreza*. El puntaje de las madres en

cuanto a la depresión es MAR, si los valores faltantes de la depresión no dependen de su nivel de depresión. Si la probabilidad de negarse a responder está relacionada con el estado de pobreza mas no con la depresión dentro de cada nivel del estado de pobreza, entonces los valores ausentes son MAR. El asunto no es si el estado de pobreza puede predecir la depresión maternal, sino si el estado de pobreza es un mecanismo para explicar si una madre reportará o no su nivel de depresión (patrón de ausencia). Bajo este patrón la distribución depende de los datos pero no depende de los datos ausentes por sí mismos y es asumida por la mayoría de los métodos existentes para imputación de datos ausentes [Little y Rubin, 1987]. En el caso de MCAR, la suposición es que las distribuciones de los datos ausentes y completos son las mismas, mientras que para MAR ellas son diferentes y los datos ausentes pueden predecirse usando los datos completos [Shafer, 1997].

En el caso MNAR, el patrón de los datos ausentes no es aleatorio y no se puede predecir a partir de la información contenida en otras variables. Bajo este patrón, contrario al MAR, el proceso de ausencia de los datos sólo se explica por los datos que están ausentes (p. ej., un ensayo sobre la pérdida de peso en que un participante abandona el estudio debido a preocupaciones por su pérdida de peso). La distribución depende de los datos ausentes y es raramente usada en la práctica.

La formulación matemática de los patrones es la siguiente [Little y Rubin, 2002]:

- ✓ MCAR: Si $f(M|Y, \Phi) = f(M|\Phi)$ para todo Y, Φ. Si la probabilidad de que el valor de una variable Y_j sea observado para un individuo i no depende ni del valor de esa variable, Y_{ij}, ni del valor de las demás variables consideradas, Y_{ik} $k \neq j$.

- ✓ MAR: Si $f(M|Y, \Phi) = f(M|Y_{obs}, \Phi)$ para todo Y_{aus}, Φ. Si la probabilidad de que el valor de una variable Y_j sea observado para un individuo i no depende del valor de esa variable, Y_{ij}, pero tal vez del que toma alguna otra variable observada, Y_{ik} $k \neq j$.

- ✓ MNAR: Si el mecanismo de ausencia depende de Y_i. Si la probabilidad de que un valor Y_{ij} sea observado depende del propio valor Y_{ij}.

Dada la importancia del patrón de ausencia, en la selección de las técnicas para imputación, se presenta otro ejemplo. Las tablas 12 a 15 hacen parte de un ejemplo presentado por Magnani, el cual permite entender mejor las diferencias entre los tres patrones [Magnani, 2004]. La tabla 12 representa

una lista de pacientes, con su edad y el resultado de una prueba médica. *Edad* es la variable independiente mientras *Resultado* es la variable dependiente.

Tabla 12. Datos para ejemplo sobre MCAR, MAR y NMAR.

Identificación	Edad	Resultado
Pac1	23	1453
Pac2	23	1354
Pac3	23	2134
Pac4	23	2043
Pac5	75	1324
Pac6	75	1324
Pac7	75	2054
Pac8	75	2056

Fuente: [Magnani, 2004]

La tabla 13 corresponde al patrón MCAR. En ella se supone que dado el costo de la prueba médica sólo se aplica a algunos pacientes elegidos aleatoriamente y por lo tanto los médicos pueden considerar sólo aquellos registros con resultado en la prueba. Nótese que *P(Resultado ausente | Edad = 23)* = *P(Resultado ausente | Edad = 75)* y *P(Resultado ausente | Resultado)* = *P(Resultado ausente)*.

Tabla 13. Datos para ejemplo sobre MCAR.

Identificación	Edad	Resultado
Pac1	23	1453
Pac2	23	*Null*
Pac3	23	2134
Pac4	23	*Null*
Pac5	75	1324
Pac6	75	*Null*
Pac7	75	2054
Pac8	75	*Null*

Fuente: [Magnani, 2004]

La tabla 14 corresponde al patrón MAR. En ella, la prueba es aplicada principalmente a personas viejas y en consecuencia, el hecho de que un *TestResult* esté ausente depende de la variable *Age*. En particular, *P(TestResult ausente | Age = 23) = 0.5*, mientras *P(TestResult ausente | Age = 75) = 0.0.*

Tabla 14. Datos para ejemplo sobre MAR.

Identificación	Edad	Resultado
Pac1	23	1453
Pac2	23	*Null*
Pac3	23	2134
Pac4	23	*Null*
Pac5	75	1324
Pac6	75	1324
Pac7	75	2054
Pac8	75	2056

Fuente: [Magnani, 2004]

La tabla 15 corresponde al patrón NMAR. En este caso la ausencia de un *TestResult* depende del valor ausente. Más formalmente, *P(TestResult ausente | TestResult < 2000) = 0.0*, mientras *P(TestResult ausente | TestResult > 2000) = 1.0*

Verbeke y Molenberghs indican cómo identificar el patrón que describe la distribución de los valores faltantes [Verbeke y Molenberghs, 2000].

Tabla 15. Datos para ejemplo sobre NMAR.

Identificación	Edad	Resultado
Pac1	23	1453
Pac2	23	1354
Pac3	23	*null*
Pac4	23	*Null*
Pac5	75	1324
Pac6	75	1324
Pac7	75	*Null*
Pac8	75	*Null*

Fuente: [Magnani, 2004]

Existen múltiples técnicas para imputación de valores, las cuales pueden ser subdivididas en métodos de imputación simple e imputación múltiple. Para el caso de los métodos de imputación simple, un valor faltante es imputado por un solo valor, mientras que en el caso de los métodos de imputación múltiple, se calculan varias opciones, usualmente ordenadas por probabilidades [Rubin, 1977]. Rubin define a los métodos de imputación múltiple como un proceso donde varias bases de datos completas son creadas mediante la imputación de diferentes valores para reflejar la incertidumbre acerca de los valores correctos a imputar. Luego, cada una de las bases de datos es analizada mediante

procedimientos estándar específicos manipulando los datos completos para, por último, combinar los análisis de las bases de datos individuales en un resultado final. Según Metha *et. al.*, el objetivo primario de la imputación múltiples es crear un conjunto de datos imputado que mantenga la variabilidad de la población total mientras preserva las relaciones con otras variables, así, las características importantes del conjunto de datos como un todo (medias, varianzas, parámetros de regresión) son preservados [Metha *et. al.*, 2004]. Como es de esperarse, estos métodos en general son de mayor complejidad asintótica algorítmica [Rubin, 1996].

Algunas de las técnicas de imputación son Imputación usando la media (IM), Imputación usando la Mediana, Imputación *Hot Deck*, Imputación por regresión, EM (*Expectation Maximization*), SRPI (*Similar Response Pattern Imputation*), FIML (*Full Information Maximum Likelihood*), RBHDI (*Resemblance-Based Hot-Deck Imputation*), ISRI (*Iterative Stochastic Regression Imputation*), kNNSi (*k-Nearest Neighbour Single Imputation*), entre otras.

Se han realizado diversos trabajos comparativos entre técnicas para valores faltantes. En el ámbito de la ingeniería de software, los trabajos de El-Emam y Birk [El-Emam y Birk, 2000] y Strike *et. al.* [Strike *et. al.*, 2001] comparan *Hot Deck* y otras técnicas de imputación incluyendo imputación múltiple. También en ingeniería de software, Myrtveit *et. al.* comparan LD, IM, SRPI y FIML usando datos de 176 proyectos [Myrtveit *et. al.*, 2001]. Brown, evalúa la eficacia de cinco técnicas (LD, PD, IM, *Hot Deck* y SRPI) en el contexto del modelamiento de ecuaciones estructurales encontrando que SRPI suministraba el menor sesgo [Brown, 1994]. Gold y Bentler compararon los métodos de imputación RBHDI, ISRI y dos variaciones de FIML [Gold y Bentler, 2000]. Browne, estudió LD, PD, IM y FIML mediante simulaciones de MonteCarlo, encontrando a FIML superior a las otras tres técnicas [Browne, 1983]. En [Lakshminarayan *et. al.*, 1999] se comparan dos métodos de imputación basados en algoritmos de aprendizaje de máquina [Lakshminarayan *et. al.*, 1999]. En [Cartwright *et. al.*, 2003] se analiza el desempeño de kNNSI e imputación de la media usando dos pequeños conjuntos de datos industriales. En [Song y Shepperd, 2004] se evalúa kNNSI e imputación de la media para diferentes patrones y mecanismos de datos ausentes. Olinsky *et. al.* compararon la eficacia para estimar modelos de ecuaciones estructurales con datos incompletos de las técnicas EM, FIML, IM, MI e imputación por regresión [Olinsky et. al, 2002]. Sin embargo, ninguno de estos trabajos, suministra una guía metodológica para seleccionar alguna de las técnicas evaluadas bajo una situación particular. Esto es, de acuerdo con la naturaleza de los datos en cuestión. La idea principal es dotar al analista de datos, quien no necesariamente conoce las técnicas existentes para limpieza de datos ni tiene mayores conocimientos estadísticos, de una guía que lo oriente, logrando así acercar el conocimiento científico al usuario común. Este capítulo se centra sólo en cuatro de ellas, tratando de determinar las condiciones para que su

aplicación sea correcta y poder así hacer recomendaciones al analista de los datos.

Las técnicas de imputación examinadas en este trabajo son:

- ✓ Imputación usando la Media.
- ✓ Imputación usando la Mediana.
- ✓ Imputación *Hot Deck.*
- ✓ Imputación por Regresión simple.

El resto del presente capítulo está organizado como sigue: la sección 5.1 describe las técnicas de imputación comparadas en este trabajo. La sección 5.2 describe las métricas de evaluación utilizadas. La sección 5.3 describe el diseño del experimento realizado para evaluar la eficacia de las diferentes técnicas. La sección 5.4 muestra los resultados obtenidos y en la sección 5.5 se presenta la guía metodológica para el problema de los valores ausentes, faltantes o nulos. Por último se presentan las conclusiones sobre este tema en la sección 5.6.

4.1. Técnicas de imputación

4.1.1. Imputación usando la media

En la imputación usando la media, la media aritmética de los valores de una variable que contiene datos faltantes es usada para sustituir los valores faltantes [Farhangfar *et. al.*, 2007].

Según Myrtveit *et. al.*, la imputación de la media (IM) es probablemente la técnica más ampliamente usada y la motivación para seleccionarla es su rapidez computacional [Myrtveit *et. al.*, 2001]. Puede usarse con variables tanto de tipo discreto como continuas que cumplan un patrón MCAR [Little y Rubin, 1987, 1990].

Anderson *et. al.* afirman que en el caso de una distribución normal, la media muestral provee un estimado óptimo del valor más probable [Anderson *et. al.*, 1983]. Si bien uno puede imputar todos los valores ausentes X_i, la varianza de X se contraerá debido a que todos los valores X_i adicionados no contribuirán en nada a la varianza. El uso de IM, afectará la correlación entre la variable imputada y cualquiera otra, reduciendo su variabilidad. Esto es, la sustitución de la media en una variable, puede llevar a perjudicar estimaciones de los efectos de otra o todas las variables en un análisis de regresión, porque el

perjuicio en una correlación puede afectar los pesos de todas las variables. Adicionalmente, si se imputa un gran número de valores usando la media, la distribución de frecuencias de la variable imputada puede ser engañosa debido a demasiados valores localizados centralmente creando una distribución más alargada o leptocúrtica [Rovine y Delaney, 1990].

En razón de lo anterior, Myrtveit *et. al.* recomiendan usar IM sólo cuando el porcentaje de valores faltantes no exceda de 5 a 10 por ciento para lograr algún grado de confianza. Esto debería ser verificado por simulación. Aunque IM es un método simple, y generalmente no recomendado, bajo estas condiciones es preferible a LD ya que puede reducir enormemente la pérdida de información sin introducir un sesgo significativo en los datos [Myrtveit *et. al.*, 2001]. Para Acock este es el peor de los procedimientos de imputación, y por tanto no recomienda su uso. Bajo este procedimiento de imputación, el valor medio de la variable se preserva, pero otros estadísticos que definen la forma de la distribución —varianza, covarianza, quartiles, sesgo, curtosis, entre otros, pueden ser afectados [Acock, 2005]. Brick *et. al.*, agregan que como resultado, los intervalos de confianza son demasiado cortos y tienen niveles nominales más bajos y las pruebas de hipótesis tienen niveles de significancia que son mayores a los niveles nominales [Brick *et. al.*, 2005].

IM y otras técnicas similares están motivadas por el deseo de llenar los valores en una matriz de datos, permitiendo así que la matriz de datos resultante sea usada en cualquier análisis de datos subsiguiente. Por ejemplo si el objetivo no es construir un modelo de predicción por regresión, sino más bien un requerimiento tipo CART (*Classification And Regression Trees*) u otro como por ejemplo análisis de clusters, deben llenarse los valores ausentes, o alternativamente remover las observaciones incompletas y por lo tanto, debe recurrirse a LD, IM o SRPI [Myrtveit *et. al.*, 2001].

Schafer y Graham, no descartan completamente a la IM y otras técnicas de imputación simple. Según ellos, hay situaciones en las que la imputación simple “es razonable” y mejor que LD. Colocan como ejemplo un conjunto de datos con 25 variables en las cuales se tiene el 3% de todos los valores de datos faltantes. Si los valores ausentes están propagados uniformemente a través de la matriz de datos, entonces LD descarta más de la mitad de los participantes ($1 - 0.97^{25} = 0.53$). De otra parte, imputar una vez una distribución condicional permite el uso de todos los participantes con sólo un impacto negativo menor en las medidas de estimación e incertidumbre [Schafer y Graham, 2002].

4.1.2. Imputación usando la Mediana

Según Acuña y Rodríguez, dado que la media es afectada por la presencia de valores extremos, parece natural usar la mediana en vez de la media con el fin de asegurar robustez. En este caso el valor faltante de una característica dada es reemplazado por la mediana de todos los valores conocidos de ese atributo. Este método es también una opción recomendada cuando la distribución de los valores de una característica es sesgada [Acuña y Rodríguez, 2009].

Obviamente técnicas como la imputación de la media y la mediana, sólo son aplicables a variables cuantitativas y no pueden usarse con valores faltantes en una característica categórica, en cuyo caso puede usarse la imputación de la moda. Estos métodos de imputación son aplicados separadamente en cada característica que contiene valores faltantes. Nótese que la estructura de correlación de los datos no está siendo considerada en los métodos anteriores. La existencia de otras características con información similar (alta correlación), o poder de predicción similar puede hacer la imputación del valor faltante inútil, o aun perjudicial [Acuña y Rodríguez, 2009].

En el mismo sentido, Mcknight *et. al.* afirman "la media es el valor más probable de las observaciones cuando los datos se distribuyen normalmente y por lo tanto sirve como estimación para los valores faltantes. Sin embargo, la media puede ser una caracterización inapropiada de los datos que no están distribuidos normalmente, especialmente cuando la distribución se desvía fuertemente de la Gaussiana normal. Las distribuciones que son asimétricas, así como aquellas que son planas (platicúrticas) o con picos (leptocúrticas) están en riesgo de ser pobremente representadas por una media. Por lo tanto, una medida alternativa de tendencia central representa mejor la distribución subyacente y por tanto una mejor estimación para los valores faltantes. La mediana, en particular, frecuentemente funciona bien como una medida de tendencia central cuando las distribuciones se desvían considerablemente de la distribución normal estándar. El procedimiento para sustituir la mediana para los valores faltantes para una variable particular sigue la misma lógica y protocolo que la sustitución de la media" [Mcknight *et. al.*, 2007].

4.1.3. Imputación Hot Deck

Con el propósito de preservar la distribución de probabilidad de las variables con datos incompletos, los estadísticos de encuestas desarrollaron el procedimiento de imputación no paramétrico denominado *Hot Deck* [Nisselson *et. al.*, 1983].

El método tiene como objetivo llenar los registros vacíos (receptores) con información de campos con información completa (donantes) y los datos

faltantes se reemplazan a partir de una selección aleatoria de los valores observados, lo cual no introduce sesgos en la varianza del estimador.

El algoritmo consiste en ubicar registros completos e incompletos, identificar características comunes de donantes y receptores y decidir los valores que se utilizarán para imputar los datos omitidos. Para la aplicación del procedimiento es fundamental generar agrupaciones que garanticen que la imputación se llevará a cabo entre observaciones con características comunes y la selección de los donantes se realiza de forma aleatoria evitando que se introduzcan sesgos en el estimador de la varianza [Medina y Galván, 2007].

Hot Deck identifica las omisiones y sustituye el valor faltante por el ingreso de algún registro "similar", utilizando para ello un conjunto de covariables correlacionadas con la variable de interés logrando preservar mejor la distribución de probabilidad de las variables imputadas. La aplicación del método, requiere adoptar algún criterio que permita identificar cual de los valores observados será utilizado en la imputación [Medina y Galván, 2007].

Goicoechea, se refiere a los procedimientos *Hot Deck*, como procedimientos de duplicación. Cuando falta información en un registro se duplica un valor ya existente en la muestra para reemplazarlo. Todas las unidades muestrales se clasifican en grupos disjuntos de forma que sean lo más homogéneas posible dentro de los grupos. A cada valor que falte, se le asigna un valor del mismo grupo. Se está suponiendo que dentro de cada grupo la no-respuesta sigue la misma distribución que los que responden [Goicoechea, 2002]. En este mismo sentido, Lavrakas anota que los métodos de imputación *Hot Deck* asumen que el patrón o mecanismo de ausencia de los datos es MAR dentro de cada grupo o clase de imputación. Esto es, condiciona que quienes no responden no sean diferentes a quienes responden para las variables auxiliares que construyen las clases de imputación [Lavrakas, 2008].

Ávila advierte sobre un posible peligro en usar el procedimiento *Hot Deck*: la duplicación del mismo valor reportado muchas veces. Este peligro ocurre cuando en los grupos de clasificación hay muchos valores faltantes y pocos valores registrados. El procedimiento *Hot Deck* resulta mejor cuando se trabaja con tamaños de muestra grandes para así poder seleccionar valores que reemplacen a las unidades faltantes [Ávila, 2002]. En el mismo sentido, Goicoechea resume sus inconvenientes: "estos métodos tienen algunas desventajas, ya que distorsionan la relación con el resto de las variables, carecen de un mecanismo de probabilidad y requieren tomar decisiones subjetivas que afectan a la calidad de los datos, lo que imposibilita calcular su confianza. Otros de los inconvenientes son: 1. que las clases han de ser definidas con base en un número reducido de variables, con la finalidad de asegurar que habrá suficientes observaciones completas en todas las clases y 2. la posibilidad de usar varias veces a una misma unidad que ha respondido" [Goicoechea, 2002]. Así mismo, también resalta sus ventajas: "El método *Hot-*

Deck tienen ciertas características interesantes a destacar: 1. los procedimientos conducen a una post-estratificación sencilla, 2. no presentan problemas a la hora de encajar conjuntos de datos y 3. no se necesitan supuestos fuertes para estimar los valores individuales de las respuestas que falten. Otra ventaja de este método es la conservación de la distribución de la variable" [Goicoechea, 2002].

Wang presenta el siguiente ejemplo de imputación *Hot Deck* [Wang, 2003]:

La tabla 16 presenta un conjunto de datos con valores faltantes. Nótese que el caso tres tiene un dato faltante en el atributo *Item 4*. Usando técnicas de tipo *Hot Deck*, cada uno de los otros casos con los datos completos es examinado y el valor del caso más similar es sustituido por el valor faltante. En este ejemplo, el caso uno, dos y cuatro son examinados. El caso cuatro es fácilmente eliminado, ya que no tiene nada en común con el caso 3. Los casos uno y dos tienen similitudes con el caso 3. El caso uno tiene un ítem en común mientras que el caso dos tiene dos ítems en común. Por tanto el caso dos es más similar al caso 3.

Tabla 16. Conjunto de datos incompletos Ejemplo *Hot Deck*

Case	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	5	10	3	**?**
4	2	5	10	2

Fuente: [Wang, 2003]

Una vez el caso más similar es identificado, la imputación *Hot Deck* sustituye el valor del caso más completo por el valor faltante. Ya que el segundo caso contiene el valor 13 en el ítem cuatro, el valor de 13 reemplaza el valor faltante en el caso tres (Tabla 17).

Tabla 17. Conjunto de datos imputados Ejemplo *Hot Deck*

Case	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	5	10	3	**13**
4	2	5	10	2

Fuente: [Wang, 2003]

Existen diferentes técnicas de imputación *Hot Deck:*

✓ **Imputación *Hot Deck*: Muestreo aleatorio simple** [Juárez, 2003]

Los donantes se extraen de manera aleatoria. Dado un esquema de muestreo equiprobable, la media se puede estimar como la media de los receptores y los donantes.

✓ **Imputación *Hot Deck*: Por clase** [Juárez, 2003]

El donante se escoge al azar de la clase a la que pertenece el receptor. Los valores faltantes dentro de cada celda se reemplazan por los valores registrados de la misma celda. La oficina de censos de los Estados unidos emplea este método para imputar ingresos en el suplemento de ingresos de la encuesta de la población actual (CPS) [Hanson, 1978], basada en variables observadas (edad, raza, sexo, relación familiar, hijos, estado civil, ocupación, escolaridad, tipo de residencia) de individuos semejantes, de tal manera que su clasificación crea una gran matriz.

✓ **Imputación *Hot Deck*: Secuencial** [Useche y Mesa, 2006]

Cada caso es procesado secuencialmente. Si el primer registro tiene un dato faltante, este es reemplazado por un valor inicial para imputar, pudiendo ser obtenido de información externa. Si el valor no está perdido, este será el valor inicial y es usado para imputar el subsiguiente dato faltante. Entre las desventajas se encuentra que cuando el primer registro está perdido, se necesita de un valor inicial, (generalmente obtenido de manera aleatoria), además cuando se necesitan imputar muchos registros se tiende a emplear el mismo registro donante, llevando esto a su vez la pérdida de precisión en las estimaciones.

✓ **Imputación *Hot Deck*: Vecino más cercano** [Juárez, 2003]

Es un procedimiento no paramétrico basado en la suposición de que los individuos cercanos en un mismo espacio tienen características similares.

Requiere de la definición de medida de distancia (generalmente euclidiana). El uso de la distancia euclidiana tiene el inconveniente de tratar todas las variables de la misma forma. Esto implica estandarizar las variables antes de calcular la distancia. Reemplaza los valores ausentes en una observación por aquellos de otra(s) observación(es) de alguna forma cercana a ella, de acuerdo con una idea predefinida de cercanía en el espacio de las variables comunes X.

Un elemento importante a considerar es el número de vecinos. Si el número de vecinos es pequeño, la estimación se hará sobre una muestra pequeña y por lo tanto el efecto será una mayor varianza en la estimación. Por otro lado, si la

imputación se hace a partir de un número grande de vecinos, el efecto puede ser la introducción de sesgo en la estimación por información de individuos alejados.

4.1.4. Imputación por Regresión

En los métodos basados en regresión, los valores ausentes para un registro dado son imputados por un modelo de regresión basado en los valores completos de los atributos para ese registro. Este método requiere múltiples ecuaciones de regresión, cada una para un conjunto diferente de atributos completos, lo cual puede conducir a altos costos computacionales [Farhangfar *et. al.*, 2007]. Existen diferentes modelos de regresión como lineal, logística, polinómica, entre otros [Lakshminarayan *et. al.*, 1999]. La regresión logística aplica MLE (*Maximum Likelihood Estimation*) después de transformar el atributo ausente en una variable Logit. Usualmente, el modelo de regresión logística es aplicada para atributos binarios, la regresión polinómica para atributos discretos y la regresión lineal para atributos continuos [Grahramani y Jordan, 1997]. En este mismo sentido Cañizares *et. al.*, afirman: "Cuando se imputa utilizando modelos de regresión, hay que tener en cuenta el tipo de variable que tiene la información incompleta. Si el valor que ha de imputarse es un número (p. ej., la edad, el salario o los valores de presión arterial), se puede emplear la regresión múltiple. En el caso que sea una variable categórica, como el sexo, el estatus socioeconómico o la práctica de ejercicio físico en el tiempo libre, podría emplearse la regresión logística y hacer la imputación según la probabilidad que el modelo de regresión estimado otorgue a cada categoría para el sujeto en cuestión" [Cañizares *et. al.*, 2004].

Estos procedimientos, al igual que el análisis de los casos completos, tienen la ventaja de que se trabaja con una base de datos completa, que se puede analizar empleando los procedimientos y paquetes estadísticos estándares. Sin embargo, la ventaja sobre el análisis de casos completos está en el hecho de que no hay pérdida de información, puesto que se trabaja con todas las unidades que fueron estudiadas [Cañizares *et. al.*, 2004].

Según Goicoechea [Goicoechea, 2002], los procedimientos de imputación asignan a los campos a imputar valores en función del modelo:

$$y_{vi} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \quad (7)$$

Donde y_{vi} es la variable dependiente a imputar y las variables X_j son las explicativas de la dependiente, también llamadas regresoras, que pueden ser tanto cualitativas como cuantitativas, variables altamente correlacionadas con la dependiente. Las variables cualitativas se incluyen en el modelo mediante variables ficticias o dummy. En este tipo de modelos se supone aleatoriedad

MAR, donde ε es el término aleatorio. A partir de este modelo se pueden generar distintos métodos de imputación dependiendo de: 1. Subconjunto de registros a los que se aplique el modelo. 2. Tipo de regresores 3. Los supuestos sobre la distribución y los parámetros del término aleatorio ε .

Según Medina y Galván, no se sugiere aplicar este método cuando el análisis secundario de datos involucra técnicas de análisis de covarianza o de correlación, ya que sobreestima la asociación entre variables y en modelos de regresión múltiple puede sobredimensionar el valor del coeficiente de determinación R^2. Si el método se aplica por estrato (subgrupos), es necesario garantizar suficientes grados de libertad (observaciones completas por subgrupo)[6]. En este caso, a pesar de que el sesgo del estimador disminuye el modelo asignará el mismo valor a un grupo de observaciones, lo cual afecta el estimador de la varianza, el coeficiente de correlación de las covariables y la variable imputada, y en los modelos de regresión multivariada el R^2 es sesgado. Una variante a este procedimiento es la imputación por "regresión estocástica", en donde los datos faltantes se obtienen con un modelo de regresión más un valor aleatorio asociado al término de error. Este procedimiento garantiza variabilidad en los valores imputados y contribuye a reducir el sesgo en la varianza y en el coeficiente determinación del modelo [Medina y Galván, 2007].

Varios paquetes estadísticos ofrecen imputación por regresión. En el caso del programa SAS este método de imputación se implementa mediante un procedimiento denominado PROC MI. Se aplica este método bajo el supuesto que los datos son perdidos al azar (MAR), o sea que la probabilidad de que una observación sea faltante depende de los valores observados pero no de los faltantes [Badler *et. al.*, 2005].

4.2. Métricas de Evaluación para Técnicas de Imputación

De acuerdo con lo expresado por los diferentes autores, en el sentido de que un buen método de imputación debe preservar las estimaciones de los parámetros que definen la forma de la distribución, se utilizó para la comparación de las cinco técnicas de imputación bajo estudio, la media aritmética, la desviación típica, la asimetría, la curtosis y el coeficiente de correlación entre las variables. Lectores no familiarizados con los conceptos básicos sobre estos estadísticos, pueden consultar [Pita y Pértega, 1997], [Webster, 2000], [Montgomery y Runger, 2003] y [Martínez, 2005].

[6] La teoría estadística establece como mínimo 30 observaciones por celda para que de acuerdo con el teorema del límite central se pueda asumir que, en el límite, la variable observada se asemeja a una distribución normal.

4.3. Diseño del Experimento para Comparación de Técnicas de Imputación

Para comparar las cuatro técnicas de imputación, se utilizó un archivo con datos extraídos del censo de los Estados Unidos, el cual ha sido ampliamente usado en otros trabajos de investigación, en libros y en guías de usuario de paquetes de Estadística. Este archivo, junto con otros, es proporcionado por el Laboratorio de Tecnología de la Información (ITL) del Instituto Nacional de Estándares y Tecnología (NIST) del gobierno de Estados Unidos, con el propósito de permitir a los investigadores analizar o verificar el comportamiento de diversas técnicas estadísticas o de aprendizaje de máquinas [ITL, 2006]. Consta de 32561 registros compuestos por los campos edad, salario, escolaridad, años_estudio, estado_civil, raza, género, entre otros, correspondientes a personas censadas entre 17 y 90 años años.

Con el fin de determinar el efecto de las técnicas de imputación sobre los datos, se calcularon los estadísticos (media, desviación estándar, coeficientes de asimetría, curtosis y correlación) para las siguientes situaciones:

- ✓ Los datos completos
- ✓ Luego de eliminar al azar el 8% de los valores para los atributos (edad y años_estudio).
- ✓ Luego de imputar por las diferentes técnicas el 8% de datos faltantes.
- ✓ Luego de eliminar al azar el 15% de los valores para los atributos (edad y años_estudio).
- ✓ Luego de imputar por las diferentes técnicas el 15% de datos faltantes.

Esto permitió ver el comportamiento de los estadísticos ante los dos niveles de datos faltantes para cada una de las técnicas de imputación y sacar conclusiones.

Las imputaciones más especializadas como *Hot Deck* y Regresión, se realizaron mediante el software estadístico SOLAS, el cual es una herramienta que ofrece seis técnicas de imputación y su propio lenguaje de programación[7].

[7] http://www.statsol.ie/index.php?pageID=5

4.4. Resultados de la Comparación de Técnicas de Imputación

A continuación se presentan los resultados arrojados por el experimento llevado a cabo. El análisis correspondiente se presenta en el numeral 5.4.1.
Las tablas 18 y 19 presentan los estadísticos para los datos de la muestra del censo de los Estados Unidos para las variables *edad* y *años_estudio* respectivamente. En ellas están los valores de la media, desviación estándar, coeficiente de asimetría y curtosis en las diferentes situaciones (datos completos, datos incompletos luego de eliminar al azar el 8% de los valores, datos completos luego de imputados usando la media, la mediana, la moda, *Hot Deck* y Regresión).

Tabla 18. Estadísticos Censo USA antes y después de imputar el 8% de los datos de la variable Edad

Situación	N	Media	Desviación Estándar	Asimetría	Curtosis
Datos completos	32,561	38.58	13.640	0.559	-0.166
Datos sin 8%	29,719	38.63	13.670	0.558	-0.165
Datos imputados (media)	32,561	38.63	13.057	0.584	0.106
Datos imputados (mediana)	32,561	38.48	13.065	0.615	0.124
Datos imputados (*Hot Deck*)	32,561	38.57	13.576	0.561	-0.171
Datos imputados (regresión)	32,561	38.63	13.060	0.577	0.113

Tabla 19. Estadísticos Censo USA antes y después de imputar el 8% de los datos de la variable Años_Estudio.

Situación	N	Media	Desviación Estándar	Asimetría	Curtosis
Datos completos	32,561	10.08	2.573	-0.312	0.623
Datos sin 8%	29,719	10.08	2.572	-0.312	0.620
Datos imputados (media)	32,561	10.08	2.462	-0.318	0.946
Datos imputados (mediana)	32,561	10.07	2.462	-0.318	0.946
Datos imputados (*Hot Deck*)	32,561	10.10	2.571	-0.313	0.619
Datos imputados (regresión)	32,561	10.08	2.465	-0.317	0.716

Las tablas 20 y 21 presentan los estadísticos para los datos de la muestra del censo de los Estados Unidos para las variables *edad* y *años_estudio* respectivamente. En ellas están los valores de la media, desviación estándar, coeficiente de asimetría y curtosis en las diferentes situaciones (datos completos, datos incompletos luego de eliminar al azar el 15% de los valores,

datos completos luego de imputados usando la media, la mediana, la moda, *Hot Deck* y Regresión).

Las figuras 3 y 4, presentan las distribuciones de probabilidad para las variables *edad* y *años_estudio* respectivamente, para las diferentes situaciones imputando el 8% de los valores.

Tabla 20. Estadísticos Censo USA antes y después de imputar el 15% de los datos de la variable Edad

Situación	N	Media	Desviación Estándar	Asimetría	Curtosis
Datos completos	32,561	38.58	13.640	0.559	-0.166
Datos sin 15%	29,719	38.66	13.687	0.558	-0.164
Datos imputados (media)	32,561	38.66	12.552	0.608	0.372
Datos imputados (mediana)	32,561	38.39	12.566	0.669	0.410
Datos imputados (*Hot Deck*)	32,561	38.63	13.701	0.564	-0.162
Datos imputados (regresión)	32,561	38.66	12.564	0.569	0.155

Tabla 21. Estadísticos Censo USA antes y después de imputar el 15% de los datos de la variable Años_Estudio

Situación	N	Media	Desviación Estándar	Asimetría	Curtosis
Datos completos	32,561	10.08	2.573	-0.312	0.623
Datos sin 15%	29,719	10.08	2.567	-0.313	0.629
Datos imputados (media)	32,561	10.08	2.365	-0.340	1.278
Datos imputados (mediana)	32,561	10.07	2.365	-0.325	1.270
Datos imputados (*Hot Deck*)	32,561	10.07	2.577	-0.319	0.632
Datos imputados (regresión)	32,561	10.08	2.369	-0.320	0.719

Las figuras 5 y 6, presentan las distribuciones de probabilidad para las variables *edad* y *años_estudio* respectivamente, para las diferentes situaciones imputando el 15% de los valores.

Figura 3. Distribuciones de probabilidad de la variable Edad. Imputaciones para 8% de los datos.

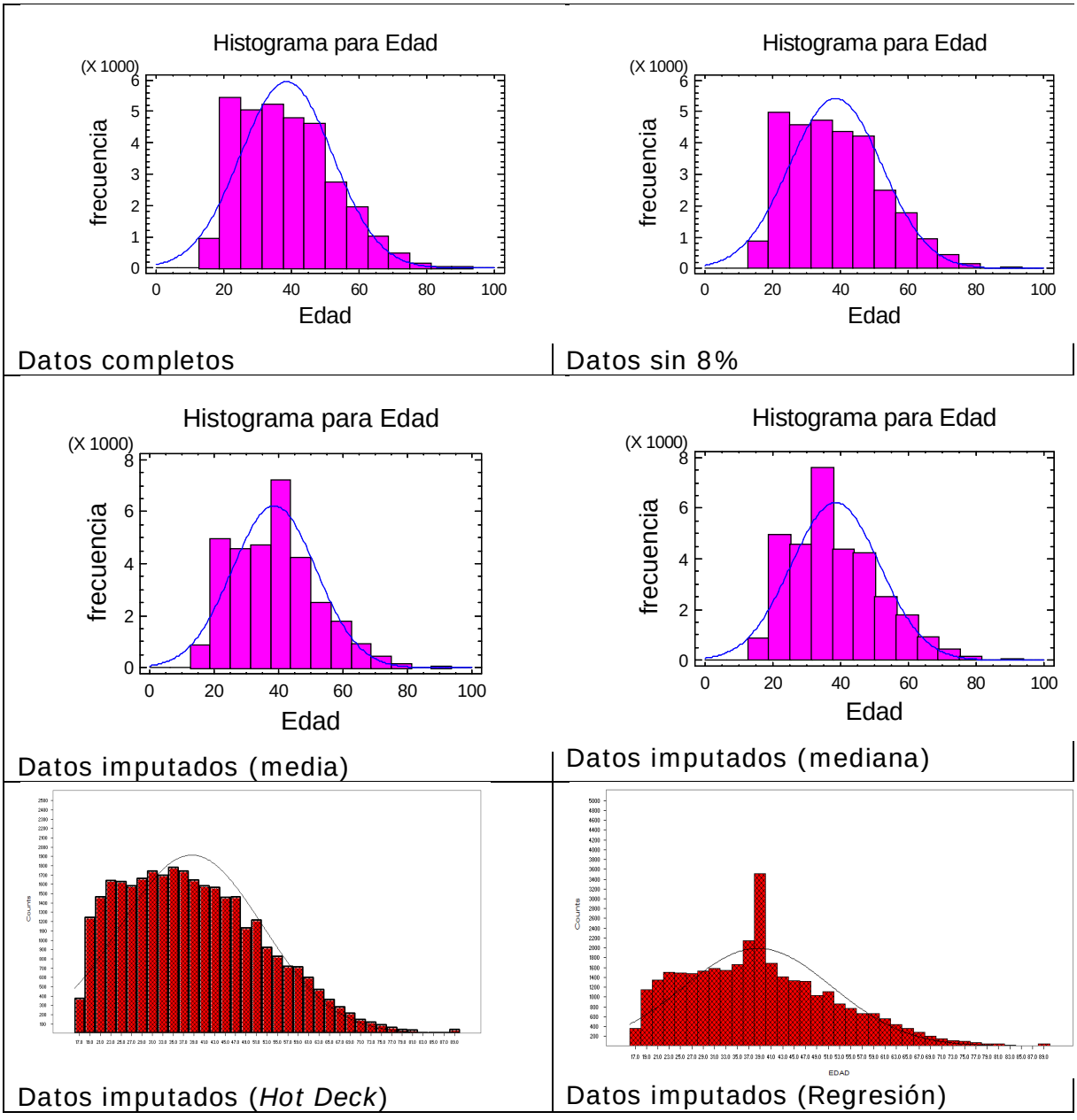


Figura 4. Distribuciones de probabilidad de la variable Años_Estudio. Imputaciones para 8% de los datos.

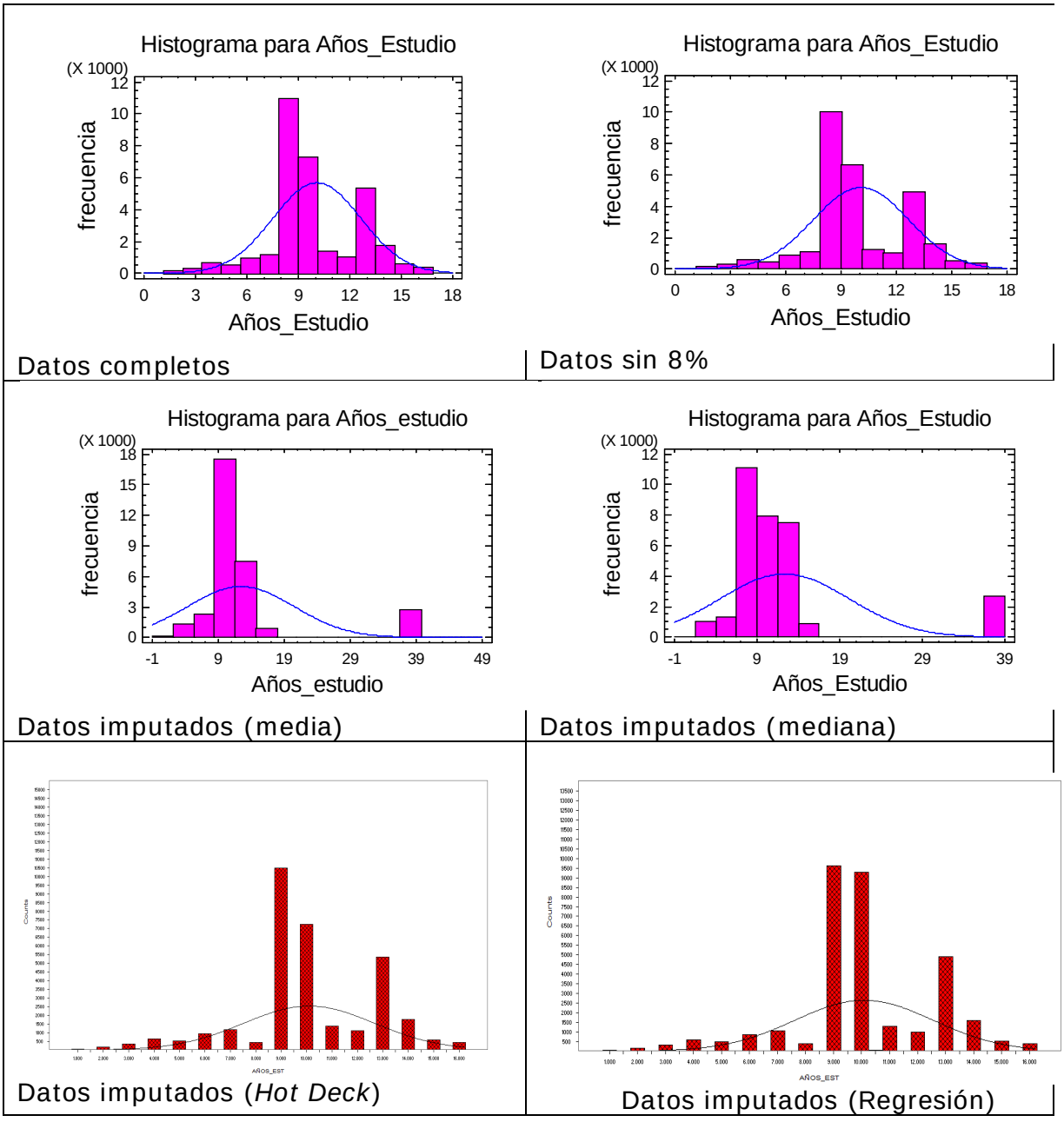


Figura 5. Distribuciones de probabilidad de la variable Edad. Imputaciones para 15% de los datos.

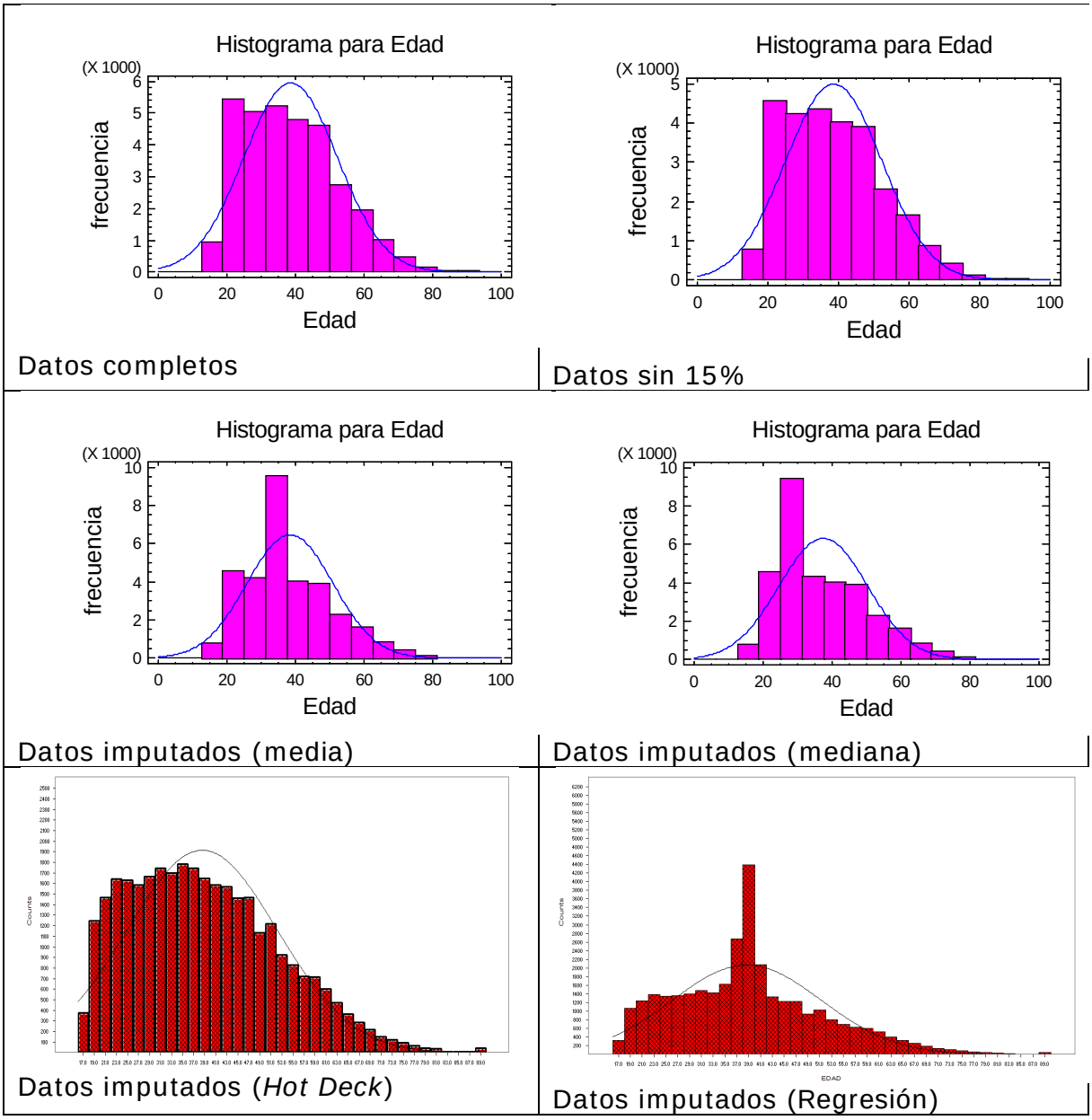
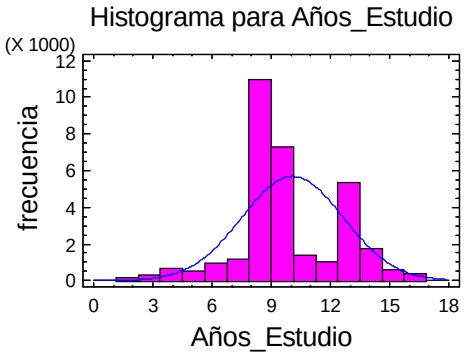
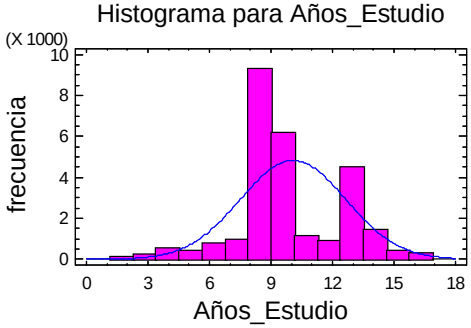


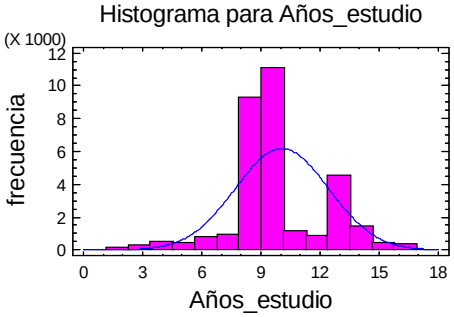
Figura 6. Distribuciones de probabilidad de la variable Años_Estudio. Imputaciones para 15% de los datos.



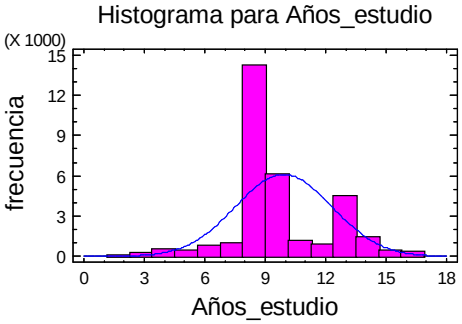
Datos completos



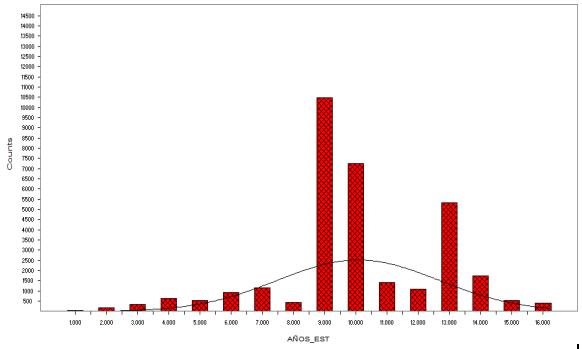
Datos sin 15%



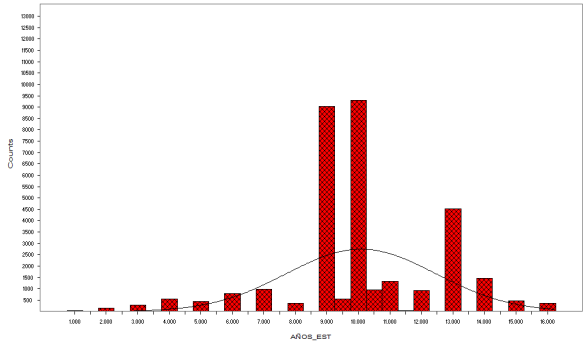
Datos imputados (media)



Datos imputados (mediana)



Datos imputados (*Hot Deck*)



Datos imputados (Regresión)

Conocer la distribución de los datos, es importante para poder aplicar algunas de las técnicas de imputación. A continuación se presentan los resultados de realizar pruebas de bondad de ajuste a la distribución normal de los datos

originales para la variable *Edad*. Para ello se utilizó el paquete estadístico *StatGraphics* versión 5.

```
Chi-cuadrado = 4186,26 con 13 g.l.   P-Valor = 0,0

Estadístico DMAS de Kolmogorov = 0,0630614
Estadístico DMENOS de Kolmogorov = 0,0568038
Estadístico DN global de Kolmogorov = 0,0630614
P-Valor aproximado = 0,0

Estadístico EDF        Valor       Forma Modificada    P-Valor
------------------------------------------------------------------------
Kolmogorov-Smirnov D   0,0630614       11,3789         <0.01*
Anderson-Darling A^2   238,088         238,094         0,0000*
------------------------------------------------------------------------
*Indica que el p-valor se ha comparado con las tablas de valores críticos
especialmente construido para el ajuste de la distribución actualmente seleccionada.
Otros p-valores están basados en tablas generales y pueden ser muy conservadores.
```

Dado que el *p-valor* o valor *p* de la prueba (la probabilidad de obtener un resultado al menos tan extremo como el que realmente se ha obtenido) más pequeño de las pruebas realizadas es inferior a 0.01, se puede rechazar que la variable *Edad* sigue una distribución normal con un nivel de confianza del 99%.

A continuación se presentan los resultados de realizar pruebas de bondad de ajuste a la distribución normal de los datos originales para la variable *Años_estudio*.

```
Chi-cuadrado = 69511,0 con 14 g.l.   P-Valor = 0,0

Estadístico DMAS de Kolmogorov = 0,189548
Estadístico DMENOS de Kolmogorov = 0,206605
Estadístico DN global de Kolmogorov = 0,206605
P-Valor aproximado = 0,0

Estadístico EDF        Valor       Forma Modificada    P-Valor
------------------------------------------------------------------------
Kolmogorov-Smirnov D   0,206605        37,2801         <0.01*
Anderson-Darling A^2   1104,35         1104,38         ****************
------------------------------------------------------------------------
*Indica que el p-valor se ha comparado con las tablas de valores críticos
especialmente construido para el ajuste de la distribución actualmente seleccionada.
Otros p-valores están basados en tablas generales y pueden ser muy conservadores.
```

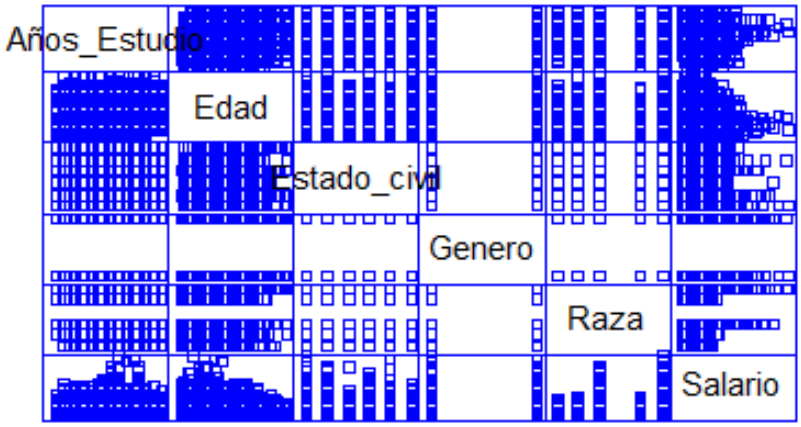
En forma similar a la variable *Edad*, dado que el *p-valor* más pequeño de las pruebas realizadas es inferior a 0.01, se puede rechazar que la variable *Años_estudio* sigue una distribución normal con un nivel de confianza del 99%.

De acuerdo con lo anterior, los valores originales de las dos variables, *Edad* y *Años_estudio*, no siguen una distribución normal, y por eso no se consideró necesario realizar pruebas de bondad de ajuste a la distribución normal luego de ser imputados mediante las diferentes técnicas. Esto sería importante si las

variables originalmente hubieran tenido distribuciones normales, ya que debería verificarse si luego de realizar las imputaciones dichas distribuciones hubieran cambiado.

Por último, antes de hacer análisis a los resultados, se requería conocer las relaciones existentes entre las variables. Para esto se calculó el coeficiente de correlación entre las variables. Debido a que las variables no se distribuyen normalmente, no pudo calcularse el tradicional coeficiente de correlación de Pearson y en su lugar se calculó el coeficiente de correlación no paramétrico de Spearman, el cual no requiere normalidad y además al trabajar por rangos es menos sensible a los posibles valores atípicos [Guilford y Fruchter 1984]. La figura 7 permite visualizar la relación entre las variables.

Figura 7. Relación entre las variables



La tabla 22 presenta los resultados del coeficiente de correlación de Spearman entre las variables.

Los resultados presentados en la figura 7 y los bajos valores del coeficiente de correlación de Spearman, permiten concluir que no existe relación lineal ni de otro tipo entre las variables. Por tanto, en este caso, no es importante el análisis de la relación entre las variables después de realizar las imputaciones mediante las diferentes técnicas.

Tabla 22. Correlaciones por Rangos Spearman para las variables

	Años_ Estudio	Edad	Estado _civil	Género	Raza	Salario
Años _Estudio		0,0663	-0.0645	0.063	0.0459	-0,0357
		0,0000	0,0000	0,2569	0,0000	0,000
Edad	0,0663		-0,3755	0,1004	0,0282	-0,0781
	0,0000		0,0000	0,0000	0,0000	0,000
Estado_ civil	-0,0645	-0,3755		-0,1550	-0,0868	0,0351
	0,0000	0,0000		0,0000	0,0000	0,000
Género	0,0063	0,1004	-0,1550		0,1000	0,0251
	0,2569	0,0000	0,0000	0,0000		0,000
Raza	0,0459	0,0282	-0,0868	0,1000		-0,0360
	0,0000	0,0000	0,0000	0,0000	0,0000	0,000

Los valores resaltados corresponden al *p-valor* para cada combinación de variables.

4.4.1. Análisis de los resultados del experimento para Valores Faltantes.

El análisis de los gráficos construidos y los estadísticos, muestra lo siguiente:

Ninguna de las técnicas de imputación comparadas afectó significativamente la media, ni siquiera con el aumento del porcentaje de valores faltantes. Esta situación, obvia para la imputación con la media, no lo es tanto para las otras técnicas. La imputación con la mediana, no la afectó ya que los valores de la media y la mediana, por ser medidas de tendencia central, están muy cercanas. La técnica *Hot Deck,* en este sentido hizo un buen trabajo eligiendo donantes que no afectaron esta medida, debido a que se contó con un volumen de datos relativamente alto y por tanto hubo buena disponibilidad de donantes. La técnica de Regresión, también logró estimar valores que no afectaran la media.

De las técnicas comparadas, la que menos subvaloró la varianza y por tanto la desviación estándar, fue la imputación *Hot Deck*. Esto sucedió a pesar del incremento en el porcentaje de valores faltantes. Las técnicas restantes redujeron en mayor proporción la varianza y la situación empeoró al crecer la cantidad de valores faltantes.

Ninguna de las técnicas de imputación afectó significativamente la asimetría, ni siquiera con el aumento del porcentaje de valores faltantes. La imputación de la media, al reemplazar por la media no alteró la distribución de los datos con respecto al punto central y por tanto no afectó esta medida. De nuevo, la imputación de la mediana, afectó muy poco la simetría ya que los valores de la media y la mediana son muy cercanos. Como en las medidas anteriores, *Hot Deck* eligió donantes que no afectaron esta medida. Regresión, también logró estimar valores que no la afectaron.

La curtosis fue la medida más sensible y más afectada por la imputación con las distintas técnicas. Las imputaciones con la media y la mediana, al reemplazar por un valor único volvieron la distribución mucho más alargada (leptocúrtica). La imputación por regresión, aunque no reemplaza por un valor único, tampoco logró buenos resultados. También para esta medida, la imputación *Hot Deck* logró los mejores resultados respetando el grado original de concentración de los valores en la región central de la distribución.

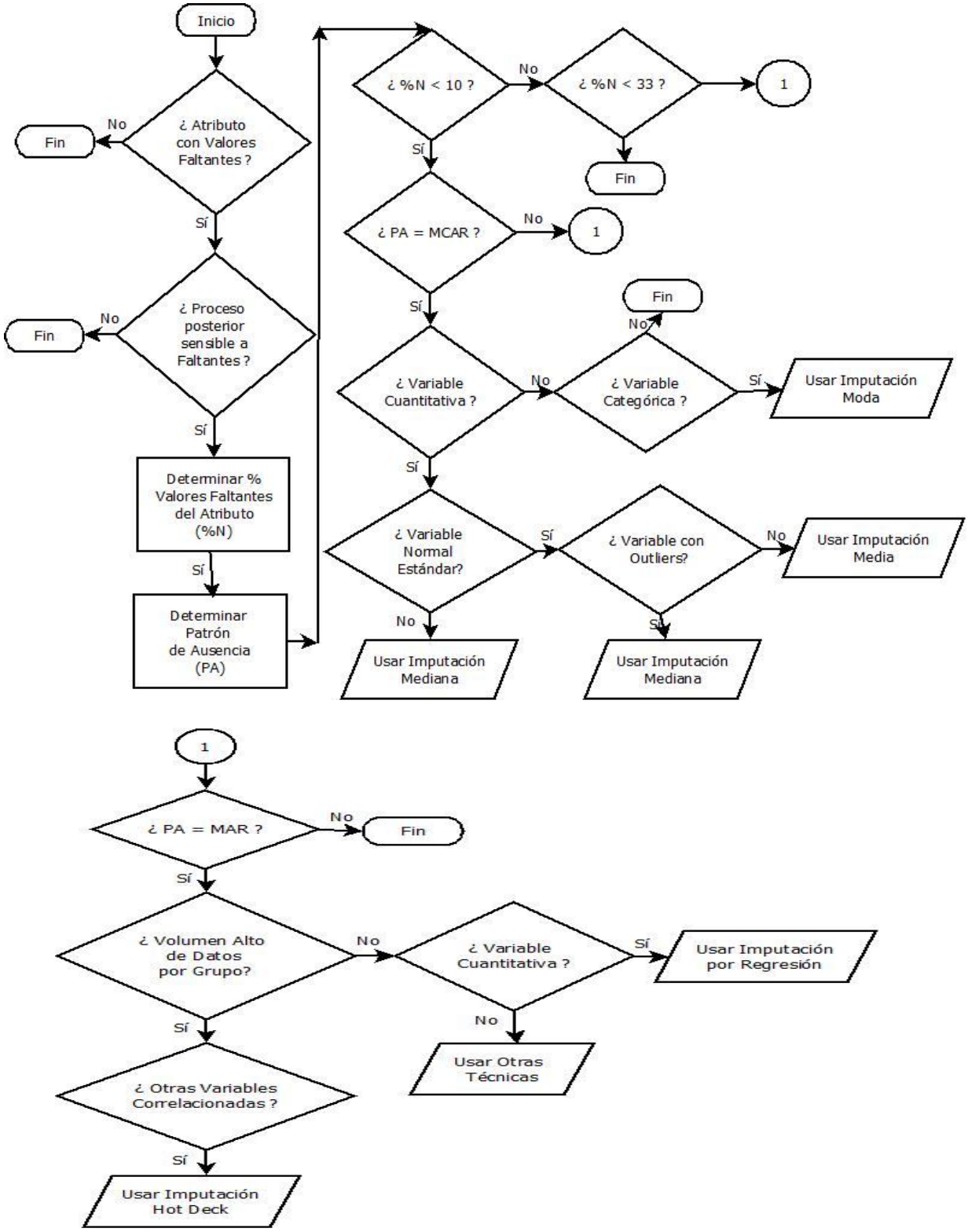
En este caso, las imputaciones no afectaron las relaciones de causalidad (correlación) entre las variables dado que desde el principio no había ninguna tendencia claramente definida.

En conclusión, en este experimento, la imputación *Hot Deck* arrojó los mejores resultados al respetar en mejor grado la distribución original de los datos. Estos resultados son consistentes con la teoría, dadas las características de los datos utilizados.

4.5. Guía Metodológica para la Selección de las Técnicas para Valores Faltantes

Al igual que para la detección de duplicados, la guía tiene la forma de un diagrama de flujo de datos. Para construirla, se tomó en cuenta la revisión literaria del numeral 5.1 y los resultados del experimento. La figura 8 corresponde al diagrama guía para la selección de las técnicas para Valores Faltantes.

Figura 8. Diagrama para la selección de técnicas para Valores Faltantes



El proceso descrito en la Figura 8 comienza indagando si el atributo a limpiar (la guía debe seguirse por cada atributo que requiera limpiarse), contiene valores faltantes. Acá debe tenerse presente que no todo valor nulo es un faltante pues de acuerdo con la naturaleza del dato, es posible que éste no sea obligatorio y por tanto no sea un error el que no tenga valor. Luego, se pregunta si el proceso a realizar con los datos es sensible a los datos faltantes. Se debe tener claro lo que se pretende hacer con los datos. Si por ejemplo, se piensa realizar un proceso de minería utilizando árboles de decisión, estos pueden trabajar con un cierto nivel de ruido y datos faltantes [Aluja, 2000], lo que no sucede con las técnicas de agrupamiento [Wagstaff, 2004].

Después de realizar lo anterior, se debe determinar el porcentaje de valores faltantes que contiene el atributo que se está examinando. Esta labor se puede realizar utilizando herramientas de perfilamiento de datos (*Data Profiling*) disponibles tanto en forma comercial como libre. De igual forma, determinar el patrón de ausencia de los datos puede realizarse con la ayuda de herramientas de software o con un análisis detallado de los datos previo entendimiento de los patrones de ausencia MCAR, MAR y MNAR explicados en este documento.

La imputación usando la media se recomienda sólo cuando el porcentaje de faltantes es bajo (menor a 10%), el patrón de ausencia es MCAR, se trata de una variable numérica, la variable tiene una distribución normal estándar (simétrica y mesocúrtica) y no hay presencia de valores atípicos (*outliers*). Determinar si una distribución es normal estándar, puede hacerse con la ayuda de programas estadísticos. La presencia de atípicos puede detectarse siguiendo la guía del capítulo 6.

Si el porcentaje de faltantes es bajo (menor a 10%), el patrón de ausencia es MCAR, se trata de una variable numérica pero no se está en presencia de una distribución normal estándar y/o hay presencia de valores atípicos, la técnica recomendada es la imputación usando la mediana. Para variables categóricas como género o estado civil, se recomienda la imputación usando la moda.

Hot Deck, es la técnica de imputación que la guía recomienda para mejorar la calidad de los datos si el patrón de ausencia es MAR, se cuenta con un volumen alto de datos por grupo y existen otras variables correlacionadas. Un volumen alto de datos por grupo se refiere a que existan suficientes datos completos con las mismas características (por ejemplo igual género, años de estudio y estado civil) para servir como donantes. Para determinar la correlación entre variables también se puede utilizar un paquete estadístico. Si no se tiene un volumen alto de datos por grupo y se trata de una variable cuantitativa se recomienda imputación por regresión. Si la variable no es cuantitativa deberá acudirse a otra técnica no examinada en este trabajo.

Por último, para porcentajes de valores faltantes entre 10% y 33%, se descartan las técnicas de imputación de la media, mediana y moda recomendables sólo para bajos niveles de faltantes. En esta situación, se consideran sólo las técnicas *Hot Deck* y Regresión bajo las mismas consideraciones anteriores. Para porcentajes de valores faltantes superiores a 33%, no se recomienda hacer imputación ya que se "fabricarán" demasiados valores. Este límite se establece de acuerdo con lo expresado por Laaksonen quien considera como tasa de no-respuesta elevada cuando dicha tasa supera un tercio del total [Laaksonen, 2000].

4.6. Conclusiones y Trabajo Futuro sobre Técnicas para corrección de Valores Faltantes

Ignorar los datos faltantes es considerado un enfoque impreciso, pero utilizar técnicas de imputación inadecuadas puede traer graves consecuencias sobre la calidad de los datos y sobre los procesos realizados posteriormente. Existen diversas técnicas para tratar con los valores faltantes, pero deben aplicarse con suma prudencia. Decidir cuál técnica utilizar en un caso particular, es un asunto complejo, el cual puede ser facilitado a los usuarios mediante una guía metodológica que recomiende la técnica adecuada para la situación.

Este objetivo se logra en este trabajo, mediante el cumplimiento de los objetivos espécificos planteados, así:

- ✓ Se logró identificar cinco técnicas para corrección de valores faltantes (imputación usando la media, la mediana y la moda, *Hot Deck* e imputación por regresión)

- ✓ Se identificaron las características de cada técnica determinando sus propiedades, funcionalidad, fortalezas y debilidades.

- ✓ Se compararon las diferentes técnicas. Para esto fue necesario:

    - o Plantear diferentes escenarios para la aplicación de las técnicas (datos completos, datos sin el 8% de los valores, datos imputando el 8% de datos faltantes, datos sin el 15% de los valores, datos imputando el 15% de datos faltantes).
    - o Identificar métricas para realizar la comparación (la media aritmética, la desviación típica, la asimetría, la curtosis y el coeficiente de correlación entre las variables).
    - o Obtener un conjunto de datos de prueba (censo USA).
    - o Calcular las medidas estadísticas para cada técnica en cada escenario.
    - o Hacer el tratamiento estadístico correspondiente que permitió obtener conclusiones válidas.

- ✓ Se diseñó un diagrama que sirve como guía al usuario para la selección de la técnica más adecuada según la naturaleza de los datos a depurar.

Como trabajo futuro, se plantea incorporar a la guía otrás técnicas no analizadas.

5. DETECCIÓN DE VALORES ATÍPICOS

Según Chandola *et. al.*, con el nombre de detección de *outliers* se conoce el problema de encontrar patrones en datos que no se ajustan al comportamiento esperado. Estos patrones son a menudo referidos como *outliers,* anomalías, observaciones discordantes, excepciones, fallas, defectos, aberraciones, ruido, errores, daños, sorpresas, novedades, peculiaridades, contaminantes, valores atípicos o valores extremos en diferentes dominios de aplicación [Chandola *et. al.*, 2007]. En este documento, se usará indistintamente valores atípicos y valores extremos.

La detección de valores atípicos ha sido ampliamente investigada y tiene considerable uso en una extensa variedad de dominios de aplicación tales como detección de fraudes con tarjetas de crédito, teléfonos móviles, seguros o asistencia en salud, detección de intrusos en sistemas de cómputo, detección de fallas en sistemas críticos, entre otros [Chandola *et. al.*, 2007]. En el mismo sentido, Hodge y Austin suministran una completa lista que incluye otros dominios como aprobación de préstamos para detectar clientes potencialmente problemáticos, desempeño de redes para detectar cuellos de botella, diagnóstico de fallas en equipos o instrumentos, detección de fallas en manufactura, análisis de imágenes satelitales, detección de novedades en imágenes, predicción meteorológica, monitoreo de condiciones médicas, proyección financiera, investigación farmacéutica, en la toma de decisiones, limpieza de datos, sistemas de información geográfica, detección de novedades en textos, detección de entradas no esperadas en bases de datos y votaciones irregulares [Hodge y Austin, 2005]. Han y Kamber, documentan la detección de fraudes con tarjetas de crédito, al identificar compras por un valor muy superior a los gastos efectuados rutinariamente en una misma cuenta. Igualmente pueden ser detectados con respecto a la ubicación, al tipo o a la frecuencia de compra [Han y Kamber, 2006]. Ertoz *et. al.*, desarrollaron técnicas para la detección de intrusos en la red [Ertoz *et. al.*, 2003].

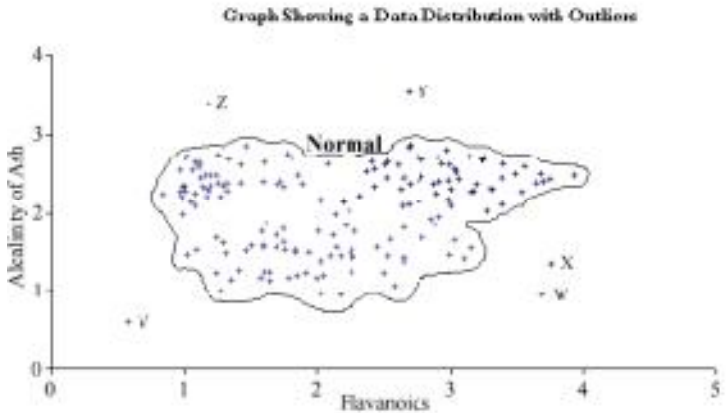
Diversos investigadores advierten sobre el daño potencial que pueden ocasionar los valores extremos al conducir a tasas de error infladas y a una significativa distorsión de los estimadores estadísticos tanto en pruebas paramétricas como no paramétricas [Zimmerman, 1994, 1995, 1998] [Rasmussen, 1998] [Schwager y Margolin, 1982]. Sin embargo, los investigadores raramente reportan en sus investigaciones la verificación de valores extremos en sus datos (según un estudio de Osborne *et. al.*, sólo el 8% de las veces) [Osborne *et. al.*, 2001]. De hecho, muchos investigadores descartan arbitrariamente observaciones que ellos califican como sospechosas o notablemente desviadas del modelo hipotético, pero esto no es considerado un enfoque adecuado. Desde años atrás, se vienen desarrollando técnicas

estadísticas para identificar observaciones candidatas para eliminación o substitución. La investigación activa sobre métodos apropiados para detectar valores atípicos ha continuado hasta el presente, pudiéndose ahora escoger entre un gran número de técnicas [Iglewicz y Hoaglin, 1993].

Barnett y Lewis definen como *Outlier* "a una observación o conjunto de observaciones las cuales parecen ser inconsistentes con el resto del conjunto de datos" [Barnett y Lewis, 1984]. Hawkins presenta una definición similar, al definir un *outlier* como "una observación que se desvía mucho de otras observaciones y despierta sospechas de ser generada por un mecanismo diferente" [Hawkins, 1980]. Beckman y Cook, se refieren a los *Outliers* ya sea como observaciones discordantes o como contaminantes. Una observación discordante es cualquier observación sorpresiva o discrepante para el investigador. Un contaminante es cualquier observación que no hace parte de la distribución objetivo [Beckamn y Cook, 1983].

Según Chandola, los datos extremos se convierten en ruido al no ser de interés para el analista y actúan como un obstáculo para el análisis de los datos [Chandola *et. al.*, 2007]. En la figura 9, hay cinco valores atípicos rotulados como X, Y, Z, V y W, los cuales están claramente aislados y son inconsistentes con el grupo principal de puntos [Hodge y Austin, 2004].

Figura 9. Ejemplo de valores atípicos en dos dimensiones.



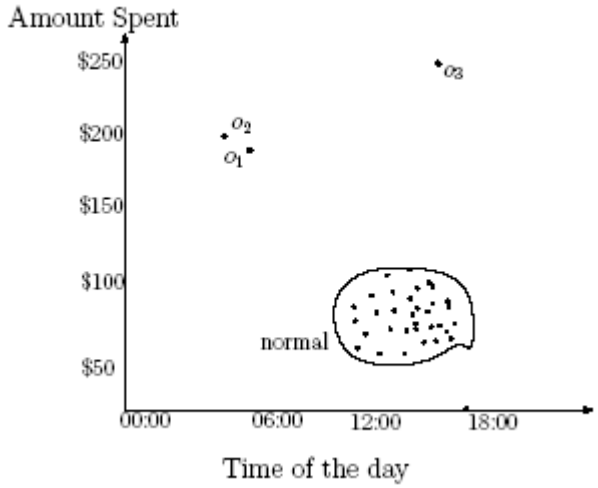
Fuente: [Hodge y Austin, 2004]

La detección de valores extremos puede ser fácil con el apoyo de una gráfica como la anterior que muestre la dispersión de los puntos, pero si se quiere realizar automáticamente la búsqueda y almacenamiento de estos valores, se necesita alguna técnica matemática para hallarlos.

Peat y Barton, establecen la existencia de dos tipos de valores atípicos: Univariantes y Multivariantes. Un valor atípico univariante es un punto de datos que es muy diferente al resto para una sola variable, como puede ser una cantidad de goles en un partido de futbol igual a 20. Un valor atípico multivariante, es un caso que es un valor extremo para una combinación de variables. Por ejemplo, un niño de 8 años de edad cuya estatura sea de 155 cms y pese 45 kg es muy inusual y sería un atípico multivariante. Los atípicos multivariantes son más difíciles de identificar y pueden ser detectados usando estadísticos como la distancia de Cook o la distancia de Mahalanobis [Peat y Barton, 2005].

Chandola *et al*. clasifican los valores atípicos en tres tipos basados en su composición y su relación con el resto de los datos [Chandola *et. al.*, 2007]. Los de tipo I son aquellos que corresponden a instancias individuales de los datos. Por ejemplo, las transacciones realizadas en una tarjeta de crédito según la hora y la cantidad gastada. En la figura 10 se muestran los datos en dos dimensiones, donde la superficie curva representa la región normal y las tres transacciones o_1. o_2 *y* o_3 se encuentran fuera de los límites de la superficie y por lo tanto son valores extremos de tipo I. Los datos normales se tomaron entre las 11 AM y 6 PM y su rango se ubica entre $60 y $100, mientras que los datos extremos pertenecen a un momento anormal y cantidades muy superiores.

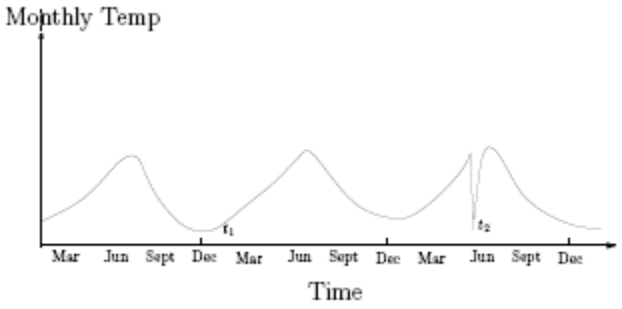
Figura 10. Ejemplo de valores atípicos Tipo I.



Fuente: [Chandola *et. al.*, 2007]

Los valores atípicos de tipo II, al igual que los de tipo I, son datos individuales, pero se diferencian en que los de tipo II se definen con respecto a un contexto específico. La figura 11 muestra una secuencia de temperaturas en una zona determinada durante un período de tiempo. Una temperatura de 35ºF puede ser normal durante el invierno (en el momento T_1), pero el mismo valor durante el verano sería una anomalía (en el momento T_2).

Figura 11. Ejemplo de valores atípicos Tipo II.



Fuente: [Chandola *et. al.*, 2007]

Los valores atípicos de tipo III constituyen un subconjunto de datos que se encuentra por fuera del conjunto total de datos. Estos son significativos solamente cuando se trata de datos espaciales o de naturaleza secuencial. Por ejemplo, una persona que realiza compras con tarjeta de crédito en diferentes sitios. Primero cancela con tarjeta en una bomba de gasolina, luego realiza compras en una tienda y en un almacén de ropa. Una nueva secuencia de transacciones con tarjeta de crédito que implica la compra en una estación de gasolina seguida de tres compras similares en el mismo lugar, el mismo día, indican la posibilidad de un robo de tarjeta. Esta secuencia de operaciones es un tipo III. Este tipo de valores atípicos corresponden a subgrafos o subsecuencias anómalas ocurridos en los datos. La figura 12 ilustra un ejemplo que muestra la salida de un electrocardiograma humano, en el cual la línea extendida corresponde a un valor atípico, pues a pesar de que existen valores igualmente bajos, no lo son por un período de tiempo tan prolongado.

Los valores extremos de tipo I pueden ser detectados en cualquier tipo de dato, mientras que los de tipo II y tipo III requieren la presencia de estructura secuencial o espacial en los datos.

Figura 12. Ejemplo de valores atípicos Tipo III.

Fuente: [Chandola *et. al.*, 2007]

Hodge y Austin, en un completo resumen sobre valores atípicos, revisan los diferentes métodos y técnicas para su detección [Hodge y Austin, 2004]. Dan cuenta de métodos estadísticos [Rousseeuw y Leroy, 1987], [Barnett y Lewis, 1994], [Beckman y Cook, 1983], [Hawkins, 1980], [Tan *et. al.*, 2006] [Angiulli *et. al.*, 2006] con técnicas como ESD (*Extreme Studentized Deviate*) o prueba de Grubbs [Grubbs, 1969] al igual que el conocido diagrama de Cajas y Bigotes (*BoxPlot*) [Laurikkala *et. al.*, 2000], métodos basados en proximidad con técnicas como k-NN (*k-nearest neighbour*) y la técnica de conectividad gráfica [Shekhar *et. al.*, 2001], métodos paramétricos con técnicas como MVE (*Minimum Volume Ellipsoid estimation*) y CP (*Convex Peeling*) [Rousseeuw y Leroy, 1996], métodos no paramétricos con técnicas como MO (*Machinery Operation*) [Dasgupta y Forrest, 1996], métodos semiparamétricos como GMM (*Gaussian Mixture Models*) [Roberts y Tarassenko, 1995], regresión, PCA, máquinas de soporte vectorial [DeCoste y Levine, 2000], redes neuronales [Markou y Singh, 2003], reglas de asociación [Narita y Kitagawa, 2008], entre otras técnicas.

También para los valores atípicos, la selección de la técnica adecuada para un proceso particular, es fundamental. Así lo ratifican Chandola *et. al.*: "El desafío clave para la detección de valores atípicos consiste en utilizar la técnica más adecuada en función de sus características, con el fin que del algoritmo se obtengan óptimos resultados en términos de precisión así como de eficiencia computacional. Sin embargo, la detección y corrección de aquellos datos inconsistentes no es tarea fácil debido a que la definición de normalidad puede no ser muy clara en muchos casos, siendo a veces el dato erróneo muy similar al dato corriente. Para ello es necesario utilizar técnicas que permitan identificar estos valores extremos de acuerdo con las características, requisitos, limitaciones y la naturaleza de los datos" [Chandola *et. al.*, 2007].

Al elegir la técnica para detectar valores atípicos, debe tenerse en cuenta el tipo de valores atípicos que ésta es capaz de identificar (univariante/multivariante, tipo I/II/III). Otro aspecto a cuidar en la selección de técnicas para detección de valores atípicos es que el procedimiento no esté afectado por ninguno de los dos posibles efectos relacionados, conocidos como efecto enmascaramiento (*masking effect*) y efecto inundación (*swamping effect*). Se trata de efectos opuestos. Ambos tienen lugar por la distorsión que

la presencia de un grupo de valores extremos causa en la media y la matriz de covarianzas, haciendo que datos con valores bastante alejados del núcleo principal de la nube de puntos no aparezcan como atípicos (*masking*) o, por el contrario, que parezcan atípicos puntos que no lo son, simplemente porque están alejados de la media calculada (*swamping*).

Se han realizado algunos trabajos comparativos que pueden servir para determinar las técnicas adecuadas para una determinada situación. Matsumoto *et. al.*, evalúan experimentalmente los métodos MOA (*Mahalanobis Outlier Analysis*), LOFM (*Local Outliers Factor Method*) y RBM (*Rule Based Modeling*), aplicados a modelos de propensión a fallas [Matsumoto *et. al.*, 2007]. Bakar *et. al.,* comparan los métodos EMM (*Extensible Markov Model*), LOF (*Local Outlier Factor*) y LCS-Mine, examinando la exactitud de la detección de valores atípicos y la complejidad computacional de los algoritmos, determinando que EMM logra mejores resultados tanto en tiempo como en exactitud. Bakar *et. al.*, describen el desempeño de las técnicas gráficos de control, regresión lineal y distancia de Manhattan en minería de datos, concluyendo que la distancia de Manhatan se comporta mejor que las otras dos técnicas *[Bakar et. al.*, 2006]. Ampanthong y Suwattee, comparan ocho técnicas encontrando que la distancia de Mahalanobis identifica la presencia de valores atípicos más a menudo que las otras técnicas para todo tipo de tamaños de datos con diferentes porcentajes de *outliers* tanto en las variables regresoras como dependientes [Ampanthong y Suwattee, 2009]. Al igual que con las técnicas existentes para otros problemas de los datos, los estudios comparativos abarcan unas cuantas técnicas y concluyen sobre su desempeño, pero no se constituyen en una guía real para un usuario común que deba enfrentar un problema de limpieza de datos. La idea principal de este trabajo, es dotar al analista de datos, quien no necesariamente conoce las técnicas existentes para limpieza de datos ni tiene mayores conocimientos estadísticos, de una guía que lo oriente, logrando así acercar el conocimiento científico al usuario común. Este capítulo se centra en cinco técnicas orientadas a valores atípicos, tratando de determinar las condiciones para que su aplicación sea correcta y poder así hacer recomendaciones al analista de los datos.

Las técnicas para detección de valores atípicos examinadas en este trabajo corresponden a *valores atípicos* tipo I. Son las siguientes:

- ✓ Prueba de Grubbs.
- ✓ Prueba de Dixon.
- ✓ Prueba de Tukey.
- ✓ MOA.
- ✓ Regresión Lineal Simple.

El resto del presente capítulo está organizado como sigue: la sección 6.1 describe las técnicas para detección de atípicos comparadas en este trabajo. La

sección 6.2 describe las métricas de evaluación utilizadas. La sección 6.3 describe el diseño del experimento realizado para evaluar la eficacia de las diferentes técnicas. La sección 6.4 muestra los resultados obtenidos y en la sección 6.5 se presenta la guía metodológica para el problema de los valores atípicos. Por último se presentan las conclusiones sobre este tema en la sección 6.6.

5.1. Técnicas para detección de valores atípicos

5.1.1. Prueba de Grubbs

Este método fue planteado por Frank E. Grubbs desde el año 1969 [Grubbs, 1969] y también es conocido como el método ESD (*Extreme Studentized Deviate*). La prueba de Grubbs se utiliza para detectar valores atípicos en un conjunto de datos univariante y se basa en el supuesto de normalidad. Es decir, primero debe verificarse que sus datos pueden aproximarse razonablemente a una distribución normal antes de aplicar la prueba. Es especialmente fácil de seguir y sirve para detectar un valor atípico a la vez [Iglewicz y Hoaglin, 1993].

Para aplicar la prueba es importante tener claros los conceptos de valor crítico y nivel de significancia. Lectores no familiarizados con ellos pueden acudir a [Triola *et. al.*, 2004].

El procedimiento de la prueba de Grubbs es el siguiente [Taylor y Cihon, 2004]:

Paso 1: Ordenar los datos ascendentemente $X_1 < X_2 < X_3 < .....X_n$

Paso 2: Decidir si X_1 o X_n es un valor sospechoso.

Paso 3: Calcular el promedio $\overline{X}$ y la desviación estándar S del conjunto de datos.

Paso 4: Se calcula T si se considera sospechoso el primer valor o el último valor.

Si X_1 es sospechoso $T = \frac{\bar{x} - x_1}{s}$ (8)

Si X_n es sospechoso $T = \frac{x_n - \bar{x}}{s}$ (9)

Paso 5: Escoger el nivel de confianza para la prueba y calcular T y compararlo con el valor correspondiente de acuerdo con una tabla de valores críticos. La tabla está disponible en [Taylor y Cihon, 2004]. Si el valor de T es mayor que el valor crítico, se dice que el dato es un valor extremo.

Iglewicz y Hoaglin, presentan el siguiente ejemplo sobre la prueba de Grubbs [Iglewicz y Hoaglin, 1993]:

El siguiente conjunto de datos tiene una media de $\overline{X}$ = 3.540, una desviación estándar S = 2.489 y un nivel de significancia α de 0.05.

2.1 2.6 2.4 2.5 2.3 2.1 2.3 2.6 8.2 8.3

Para este caso se tiene como sospechoso el último valor del conjunto de datos. Al aplicar la ecuación (9) se obtiene:

$$T = \frac{8.3 - 3.540}{2.489} = 1.913$$

Según la tabla de valores críticos, con un nivel de significancia $\alpha = 0.05$ y n = 10, el valor crítico es 2.176, concluyéndose que el dato 8.3 no es *un valor atípico* debido a que 1.913 < 2.176, pero si se observan los datos claramente se podría decir que hay un error ya que el dato es un valor inusual con respecto a los demás.

El valor X_{n-1} del conjunto de datos, 8.2, esconde el efecto de 8.3. Este fenómeno, es llamado *enmascaramiento*, y afecta un número de pruebas populares para identificar valores atípicos. El enmascaramiento ocurre cuando observaciones discordantes cancelan el efecto de observaciones extremas y evita el procedimiento de detección de atípicos, de declarar a cualquiera de las observaciones como valores atípicos. El enmascaramiento no es un problema cuando la prueba de identificación de valores atípicos es basada en estimaciones con alto desglose de puntos. En cualquier evento, un buen procedimiento de identificación de valores atípicos debería tener pequeños problemas con enmascaramiento.

Para ilustrar que T trabaja bien en la detección de valores atípicos, se debe remover el 8.3 del conjunto de datos y realizar el proceso nuevamente. Con una media $\overline{X}$ = 3.011 y S = 1.954, T es igual a 2.65. Con un nivel de significancia del 5% el valor crítico es igual a 2.110. El valor de T en este caso si es mayor que el valor crítico, y se concluye que el dato es atípico al igual que valores mayores a este, en éste, caso 8.3.

La prueba de *Grubbs* se puede modificar para un número máximo pre especificado de valores atípicos. Primero se calcula $R_1 = (\max_n \left| X_n - \overline{X} \right|) \Big/ S$. Luego se calcula R_2 de la misma manera que R_1 y la muestra se reduce a $n-1$ observaciones. Se continúa este proceso hasta que el número máximo de *outliers* definido anteriormente sean calculados. Luego se encuentra el máximo tal que $R_n > \lambda_n$ utilizando los valores de la tabla con un $\lambda = 1$ y $\alpha = 0.05$. Las observaciones removidas serán declaradas como atípicas.

Es muy importante suponer correctamente el número de atípicos (γ). En caso de duda, se debe elegir un γ grande porque uno muy pequeño podría dar como resultados valores extremos falsos. Esto puede surgir cuando la muestra contiene más de γ *outliers* o enmascaramiento cuando γ es muy pequeño. Al seleccionar un valor de γ muy grande es necesario realizar muchos cálculos, pero tiene un mínimo efecto sobre la posibilidad de identificar falsos valores atípicos.

Esta técnica es muy fácil de usar y funciona bien bajo una variedad de condiciones incluyendo tamaños de muestra muy grandes, recordando que los datos deben provenir de una distribución normal.

5.1.2. Prueba de Dixon

La prueba de Dixon permite determinar si un valor sospechoso de un conjunto de datos es un *outlier*. El método define la relación entre la diferencia del mínimo/máximo valor y su vecino más cercano y la diferencia entre el máximo y el mínimo valor aplicado [Li y Edwards, 2001].

Los datos deben provenir de una distribución normal. Si se sospecha que una población lognormal subyace en la muestra, la prueba puede ser aplicada al logaritmo de los datos. Antes de realizar el procedimiento es importante definir las hipótesis (si el valor sospechoso se encuentra al inicio o al final del conjunto de datos) y determinar la distribución de la que provienen los datos (normal o lognormal) [Davis y McCuen, 2005].

Taylor y Cihon, explican el proceso para llevar a cabo la prueba de Dixon. Se debe seguir los siguientes pasos [Taylor y Cihon, 2004]:

Paso 1: Ordenar los valores de la muestra en forma ascendente, siendo x_1 el valor más pequeño y x_n el mayor valor: $x_1 < x_2 < x_3 < .....x_n$

Paso 2: Calcular el valor de Dixon dependiendo del tamaño de la muestra según la tabla 23.

Donde las relaciones son las indicadas en la tabla 24.

Tabla 23. Prueba de Dixon de acuerdo con el tamaño del conjunto De datos

Número de datos	Relación a calcular
n = 3 a 7	r_{10}
n = 8 a 10	r_{11}
n =11 a 13	r_{21}
n = 14 a 24	r_{22}

Fuente: [Taylor y Cihon, 2004]

Tabla 24. Relaciones Pueba de Dixon

R	Si x_n es sospechoso	Si x_1 es sospechoso
r_{10}	$\frac{(x_n - x_{n-1})}{(x_n - x_1)}$	$\frac{(x_2 - x_1)}{(x_n - x_1)}$
r_{11}	$\frac{(x_n - x_{n-1})}{(x_n - x_2)}$	$\frac{(x_2 - x_1)}{(x_{n-1} - x_1)}$
r_{21}	$\frac{(x_n - x_{n-2})}{(x_n - x_2)}$	$\frac{(x_3 - x_1)}{(x_{n-1} - x_1)}$
r_{22}	$\frac{(x_n - x_{n-2})}{(x_n - x_3)}$	$\frac{(x_3 - x_1)}{(x_{n-2} - x_1)}$

Fuente: [Taylor y Cihon, 2004]

Buscar el valor crítico de *r* de acuerdo con el nivel de significancia en la tabla para valores críticos para la prueba de Dixon [Taylor y Cihon, 2004].

Si el valor de *r* calculado es mayor que el valor crítico de la tabla se concluye que es un valor atípico.

En el caso de la prueba de Dixon con más de un valor extremo sospechoso, el valor más extremo tiende a ser *enmascarado* por la presencia de otros valores. El enmascaramiento ocurre cuando dos o más valores atípicos tienen valores similares. En un conjunto de datos, si los valores más pequeños o más grandes

son casi iguales, una prueba de *outlier* para el valor más extremo de los dos no es estadísticamente significativa. Esto es especialmente cierto en el caso de los tamaños de las muestras de menos de diez, cuando el numerador de la relación es la diferencia entre los dos valores más extremos.

La prueba de Dixon es usualmente utilizada para un grupo pequeño de datos (entre 3 y 30 datos) y dispone de un valor crítico con tres puntos decimales, lo cual limita seriamente la aplicación de la prueba en muchos campos de las ciencias e ingenierías. Sin embargo un trabajo realizado por Verma y Quiroz, introdujo nuevas tablas de valores críticos más precisos y exactos con cuatro puntos decimales y se extiende hasta 100 el tamaño de la muestra [Verma y Quiroz, 2006]. Davis y Mccuen, extienden la aplicación del test de Dixon hasta 200 observaciones donde el valor de la prueba denotado como R depende del tamaño de la muestra y el valor crítico denotado como R_c se calcula por medio de polinomios. Para los valores mayores de 26 es necesario calcular la desviación estándar y la media de la muestra. La Tabla 25 muestra los valores de R y R_c [Davis y McCueen, 2005].

Zhang *et. al.*, presenta un algoritmo para la prueba de Dixon en [Zhang, 2008].

Tabla 25. Valores críticos para la prueba de Dixon extendida a 200 observaciones.

Sample Size	Low Outlier Test Statistic	High Outlier Test Statistic	Polynomial for Critical value, R_c
3 to 7	$R = \frac{X_2 - X_1}{X_n - X_1}$	$R = \frac{X_n - X_{n-1}}{X_n - X_1}$	$R_c = 1.975 - 0.4994n + 0.5895n^2 + 0.0025n^3$
8 to 10	$R = \frac{X_2 - X_1}{X_{n-1} - X_1}$	$R = \frac{X_n - X_{n-1}}{X_n - X_2}$	$R_c = 1.23 - 0.125n + 0.005n^2$
11 to 13	$R = \frac{X_3 - X_1}{X_{n-1} - X_1}$	$R = \frac{X_n - X_{n-2}}{X_n - X_2}$	$R_c = 0.90 - 0.03n$
14 to 25	$R = \frac{X_3 - X_1}{X_{n-2} - X_1}$	$R = \frac{X_n - X_{n-2}}{X_n - X_3}$	$R_c = 0.9975 - 0.04268n + 0.000764n^2$
26 to 200	$R = \frac{X_n - \overline{X}}{S_X}$	$R = \frac{\overline{X} - X_1}{S_X}$	$R_c = 2.2795 + 0.025012n - 0.00018427n^2 + 4.61106x10^{-7}n^3$

Fuente: [Davis y McCueen, 2005]

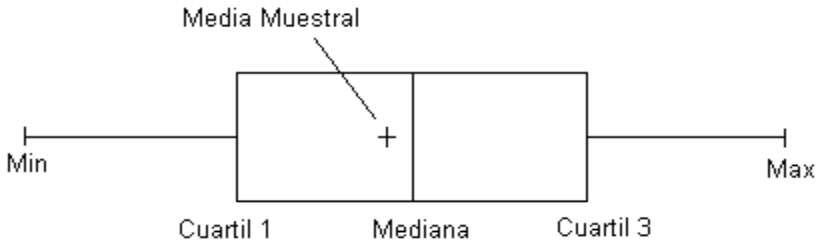
La prueba de *Dixon* es muy fácil de utilizar, pero el resultado depende fuertemente de escoger correctamente el número exacto y ubicación de todos los valores sospechosos. Por esto y ser una prueba muy susceptible al ocultamiento o enmascaramiento, se recomienda utilizar la prueba de *Dixon* sólo para pequeñas muestras cuando sólo uno o dos valores son considerados como atípicos [Iglewicz y Hoaglin, 1993]. Adicionalmente debe cumplirse que la muestra de datos proviene de una distribución normal o lognormal.

5.1.3. Prueba de Tukey

El diagrama conocido como *diagrama de cajas y bigotes* (*Box and Whiskers Plot* o simplemente *BoxPlot)* es un gráfico representativo de las distribuciones de un conjunto de datos creado por Tukey en 1977, en cuya construcción se usan cinco medidas descriptivas de los mismos: mediana, primer cuartil (Q1), tercer cuartil (Q3), valor máximo y valor mínimo [Tukey, 1977]. Está compuesto por un rectángulo o caja la cual se construye con ayuda del primer y tercer cuartil y representa el 50% de los datos que particularmente están ubicados en la zona central de la distribución, la mediana es la línea que atraviesa la caja, y dos brazos o bigotes son las líneas que se extienden desde la caja hasta los valores más altos y más bajos. En algunos casos, dentro de la caja suele trazarse una cruz para representar el promedio de los datos [Palomino, 2004].

En la figura 13 se presenta un diagrama de cajas y bigotes.

Figura 13. Diagrama de Cajas y bigotes.



Fuente: [Palomino, 2004]

Esta presentación visual asocia las cinco medidas que suelen trabajarse de forma individual y puede ser graficada de manera vertical u horizontal.

Presenta al mismo tiempo, información sobre la tendencia central, dispersión y simetría de los datos de estudio. Además, permite identificar con claridad y de forma individual, observaciones que se alejan de manera poco usual del resto de los datos, esto es, sirve para detectar los valores atípicos. Por su facilidad de construcción e interpretación, permite también comparar a la vez varios grupos de datos sin perder información ni saturarse de ella.

Usando los mismos cálculos necesarios para construir el diagrama de cajas y bigotes, puede hacerse detección automática de los valores atípicos presentes en un conjunto de datos. El método es el siguiente: se encuentra la mediana de todos los datos, luego se halla tanto la mediana de los valores iguales o inferiores a la mediana como de los superiores. Este será un valor de datos o será la mitad de entre dos valores de datos dependiendo de si la cantidad de los datos es par o impar [CQU, 1997]. Con un conjunto de datos impar, se incluye la mediana en cada una de las dos mitades del conjunto de datos y luego se encuentra el medio de cada mitad. Esto da como resultado el primer y tercer cuartil. Si el conjunto de datos tiene un número par de valores, los datos se dividen en dos mitades, y se encuentra el medio de cada mitad.

En [CQU, 1997], se presenta el siguiente ejemplo utilizando un pequeño conjunto de datos, que contiene un número impar de valores:

35 47 48 50 51 53 54 70 75

Primero, se dividen los datos en dos mitades, cada una incluyendo la mediana:

35 47 48 50 51 y 51 53 54 70 75

Se encuentra la mediana de cada mitad. En este ejemplo, para el primer cuartil es 48 y para el tercer cuartil es 54. Por lo tanto, el rango intercuartíl IQR es 54-48 = 6.

A continuación se ilustra el procedimiento para un número par de valores adicionando a la serie anterior el valor 60.

35 47 48 50 51 53 54 60 70 75

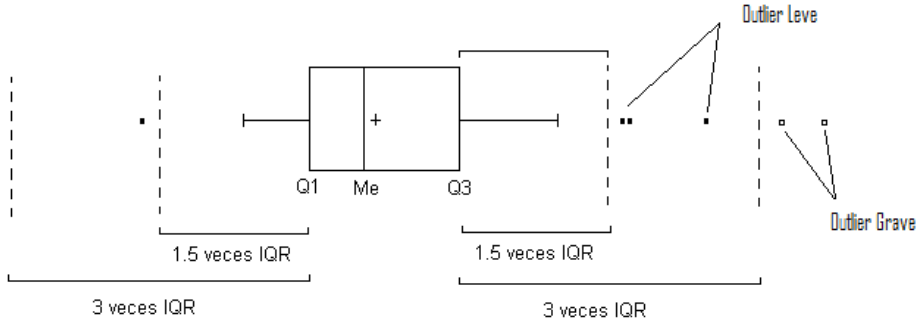
Primero se calcula la mediana entre 51 y 53 = 52 y luego se divide los datos en dos mitades:

35 47 48 50 51 y 53 54 60 70 75

Ahora, encontrar el medio de cada mitad. El primer cuartil es 48 y el tercer cuartil es 60. De ahí el rango intercuartil IQR es 60-48 = 12.

Para la detección de los valores atípicos, la longitud máxima de cada uno de los bigotes es de K = 1,5 veces el rango intercuartil (IQR) es decir 1.5x(Q3-Q1) por encima y por debajo de los cuartiles. Las observaciones fuera de los bigotes son dibujadas separadamente y etiquetadas como valores atípicos. El método de Tukey utiliza un K=3 adicionalmente del K =1.5, las observaciones que están entre 1.5 y 3 veces el rango intercuartil reciben el nombre de atípicos leves. Las observaciones que están más allá de 3 veces el rango intercuartil se conocen como valores atípicos extremos. En la figura 14 se muestra un diagrama de cajas y bigotes con valores atípicos leves y graves.

Figura 14. Diagrama de caja con valores atípicos leves y graves



Fuente: [Palomino, 2004]

A continuación se presenta un ejemplo que ilustra la detección de atípicos mediante la prueba de Tukey [CQU, 1997].

Ejemplo: Se utiliza el primer conjunto de números del ejemplo anterior usando el método de Tukey determinado por Q1 y Q3.

35 47 48 50 51 53 54 70 75

El IQR es de 6. Ahora 1,5 veces 6 es igual a 9. Este es el máximo tamaño del bigote. Restar 9 del primer cuartil: 48 - 9 = 39. Note que 35 es un valor extremo, y el bigote debe ser dibujado en 47, el cual es el menor valor que no es un atípico. Luego multiplicamos IQR por k =3 y obtenemos como resultado 18, y al restar 18 al primer cuartil 48-18=30 podemos concluir que no hay valores atípicos graves debido a que no hay datos menores que 30 pero si un valor atípico leve.

Luego se adiciona 9 al tercer cuartil 54+9=63. Cualquier valor mayor a 63 es un *outlier*, es decir, en este caso 70 y 75 son valores atípicos. Se dibuja el bigote para el mayor número del conjunto de datos que no sea un valor atípico, en este caso 54. Si se adiciona 18 al tercer cuartil 54+18=72 se concluye que 70 es un valor atípico leve y 75 un valor atípico grave. El diagrama de cajas y bigotes correspondiente se muestra en la figura 15.

Aunque existen diversas variaciones del diagrama de cajas y bigotes original, no se discutirán en este trabajo, ya que el foco de interés no es tanto la parte gráfica sino la forma automática de detectar los valores atípicos. Los interesados pueden encontrar información en [McGill *et. al.*, 1978] y [Hintze y Nelson, 1998].

Figura 15. Diagrama de cajas y bigotes para el ejemplo

Fuente: [CQU, 1987]

Según Iglewicz y Hoaglin, aunque la prueba de Tukey no es la forma más eficiente para etiquetar valores atípicos, ésta cuenta con valiosos elementos. Como consecuencia, el investigador recibe una alerta temprana para hacer frente a los valores fuera y tratar de explicar por qué se produjeron [Iglewicz y Hoaglin, 1993].

Hofmann *et. al.*, destacan como ventajas del método de Tukey: facilidad de cálculo, capacidad para manejar gran número de valores y no dependencia de un parámetro de suavización. Asimismo, indican como debilidad que puede identificar como atípicos una gran cantidad de datos (0.4 + 0.007n), lo cual para un conjunto de n=100,000 datos podría arrojar aproximadamente 700 valores atípicos [Hofmann *et. al.*, 2006]. Kampstra, refiriéndose al método de Tukey original indica que “la detección de valores atípicos es bastante arbitraria, especialmente en caso de distribuciones subyacentes no-normales. Incluso para las distribuciones normales, el número de valores extremos detectados crecerá si el número de observaciones crece, lo cual hace que los valores extremos individuales sean indetectables” [Kampstra, 2008]. Huber y Vandervieren, explican la razón para que muchos puntos sean clasificados como valores atípicos cuando los datos son sesgados. Esto es debido a que la

regla para detección de valores atípicos está basada únicamente en medidas de localización y escala, y los valores de corte son derivados de una distribución normal [Huber y Vandervieren, 2008]. En este punto merece mención especial el caso cuando el rango intercuartil (IQR) es cero, ya que bajo esta situación, esta prueba catalogará como atípicos a los valores diferentes de cero. Tómese como ejemplo, el atributo *Inversión en Investigación* que hace parte de la Encuesta Anual Manufacturera [EAM, 2008] realizada por el gobierno colombiano, en donde la gran mayoría de empresas reportan un porcentaje de inversión igual a cero. En este caso el IQR es igual a cero y por tanto aquellas empresas que inviertan un valor diferente de cero serán atípicas.

5.1.4. Análisis de Valores Atípicos de Mahalanobis

El Análisis de Valores atípicos de Mahalanobis (*Mahalanobis Outlier Analysis – MOA*), es un método basado en una distancia, llamada distancia de Mahalanobis (DM). Esta distancia es calculada con base en la varianza de cada punto. Ésta describe la distancia entre cada punto de datos y el centro de masa. Cuando un punto se encuentra en el centro de masa, la distancia de Mahalanobis es cero y cuando un punto de datos se encuentra distante del centro de masa, la distancia es mayor a cero. Por lo tanto, los puntos de datos que se encuentran lejos del centro de masa se consideran valores atípicos [Matsumoto *et. al.*, 2007].

La DM es un enfoque multivariante y es calculado para cada observación en el conjunto de datos. Entonces a cada observación se le da un peso como inverso de la distancia de Mahalanobis. Las observaciones con valores extremos obtienen menores pesos. Finalmente una regresión ponderada se ejecuta para minimizar el efecto de los valores extremos [Tiwary et. al., 2007].

La DM es diferente de la distancia euclidiana por lo siguiente [Tiwary et. al., 2007]:

Está basada en correlaciones entre variables por lo cual pueden ser identificados y analizados diferentes patrones.
Es invariante a la escala, es decir, no depende de la escala de las mediciones.
Toma en cuenta las correlaciones del conjunto de datos.

La DM se calcula de la siguiente forma [Maesschlck et. al., 2000]:

$$\mathrm{MD}_i^o = \sqrt{(x_i - \bar{x})\mathrm{C}_x^{-1}(x_i - \bar{x})^T} \quad \text{for } i = 1 \text{ to } n, \tag{10}$$

Donde C_x es la matriz de covarianza. La distancia Mahalanobis sigue una distribución chi-cuadrado con grados de libertad igual al número de variables incluidas en el cálculo [Filzmoser, 2004]. En [Maesschlck et. al., 2000] se puede encontrar un ejemplo completo sobre el cálculo de la DM y también sobre la distancia euclidiana.

Según Maesschalck *et. al*., la DM toma en cuenta la correlación en los datos, dado que ésta es calculada usando la inversa de la matriz de covarianza del conjunto de datos de interés. Sin embargo, el cálculo de la matriz de covarianza puede causar problemas. Cuando los datos investigados son medidos sobre un gran número de variables, ellos pueden contener información redundante o correlacionada. Esto conduce a una matriz de covarianza que no puede ser invertida. Una segunda limitación para el cálculo de la matriz de covarianza es que el número de objetos en el conjunto de datos tiene que ser más grande que el número de variables, requiriéndose en muchos casos reducción de características [Maesschlck et. al., 2000]. Adicionalmente, el uso de la distancia clásica de Mahalanobis para la detección de atípicos ha sido criticado por estar afectado por el efecto enmascaramiento [Rousseeuw y Van Driessen, 1999] [Becker y Gather, 1999].

Para la detección de atípicos multivariantes Rousseeuuw y Van Zomeren proponen el uso de un test de discordancia usando lo que denominan "distancia robusta" [Rousseeuuw y Van Zomeren, 1990]. Se trata de las distancias de Mahalanobis de todos los puntos respecto al estimador robusto[8] MCD (*Minimum Covariance Determinant*). El método MCD consiste, para un número determinado de datos en la muestra, en buscar la matriz de covarianza con mínimo determinante para diferentes muestras de dicho tamaño. La idea subyacente es que el determinante de la matriz de covarianzas está inversamente relacionado con la intensidad de las correlaciones. Al estar la distancia referida al estimador robusto de medias y covarianzas, no está afectada por el efecto enmascaramiento [Morillas y Díaz, 2007].

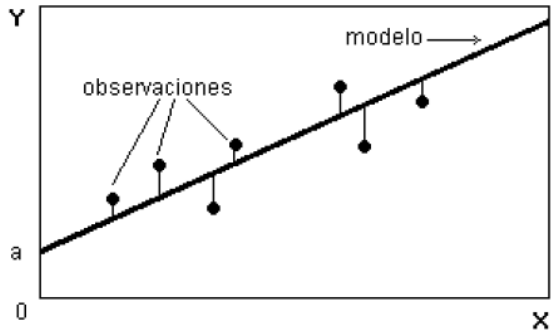
5.1.5. Detección de Valores Atípicos mediante Regresión Simple

El análisis de regresión es una importante herramienta estadística que se aplica en la mayoría de las ciencias. De muchas posibles técnicas de regresión, el método de mínimos cuadrados (LS) ha sido generalmente la más adoptada por tradición y facilidad de cálculo. Este método a través de unos cálculos,

8 Los Estimadores Robustos, conocidos también como Estimadores no Paramétricos o Estimadores Libres de Distribución o simplemente Robustos, son estimadores libres de la asunción de una forma de distribución de la población de la cual se extrae la muestra.

aproxima un conjunto de datos a un modelo, el cual puede ser lineal, cuadrado, exponencial, entre otros. Es decir, es una técnica de optimización, que intenta encontrar una función que se aproxime lo mejor posible a los datos. La diferencia entre el valor observado y el valor obtenido del modelo de regresión se denominan residuos o suma de cuadrados y el objetivo es tratar de minimizar este valor y así obtener el mejor ajuste. La figura 16 ilustra el método de mínimos cuadrados [Rousseeuw y Leroy, 1996].

Figura 16. Regresión por Mínimos cuadrados.



Fuente: [Rousseeuw y Leroy, 1996]

En la regresión lineal o simple se parte de un modelo lineal, donde existe una relación de la variable x también llamada variable independiente hacia la variable y denominada variable dependiente. La ecuación que relaciona estas dos variables es:

$$y_i = a + bx_i + e_i \quad \text{para} \quad i = 1,2,.....,n \qquad (11)$$

Donde a es el valor de la ordenada donde la línea de regresión se interceca con el eje y, b es el coeficiente de la pendiente de la línea recta y e es el error que se comete al ajustar los datos donde se supone que tiene valor esperado cero y desviación estándar común. Es deseable que los valores de 'y' ajustados al modelo, sean lo más parecidos posible a los valores observados. Una medida de lo parecido que son, es el coeficiente de correlación R^2 la cual se define como el cuadrado del coeficiente de correlación entre los valores de 'y' observados y los valores de 'y' ajustados. El rango de R^2 es entre 0 y 1, el valor entre más se acerque a 1 quiere decir que tiene un mejor ajuste [Edwards, 1976].

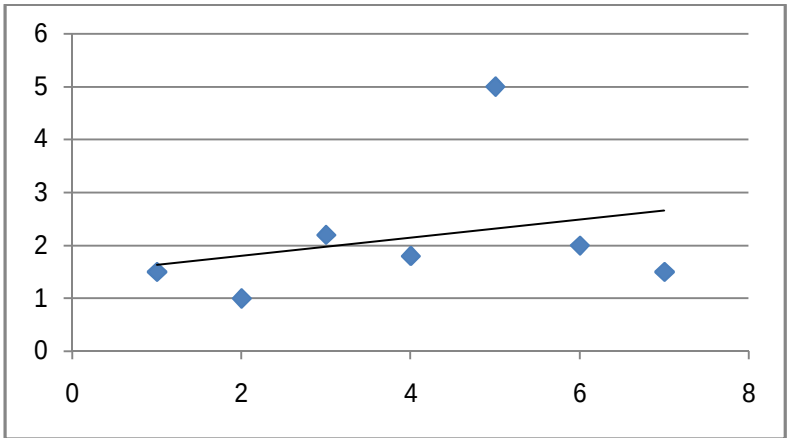
Los valores de a y b se determinan mediante las fórmulas:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad (12)$$

$$a = \bar{y} - b\bar{x} \qquad (13)$$

Un modelo de regresión permite detectar valores atípicos al considerar a los datos alejados del modelo como tales. Esto es, los casos que no siguen el modelo como el resto de los datos pueden representar datos erróneos, o pueden indicar un pobre ajuste de la línea de regresión. La figura 17 ilustra esta situación.

Figura 17. Detección de atípicos mediante regresión.



Fuente: [Rousseeuw y Leroy, 1996]

Antes de tratar de ajustar un modelo lineal a los datos observados, primero se debe determinar si existe una relación o no entre las variables de interés. Esto no implica necesariamente que una variable sea causa de la otra pero existe cierta asociación significativa entre las dos variables. Una gráfica de dispersión puede ser una herramienta útil para determinar la fuerza de la relación entre las dos variables. Si no parece haber ninguna asociación entre la variable predictora o independiente y la variable de respuesta o dependiente (es decir, la dispersión no indican ningún tendencia de aumento o disminución), ajustar un modelo de regresión lineal a los datos probablemente no va a proporcionar un modelo útil [YALE, 1998].

Una vez que un modelo de regresión ha sido ajustado a un grupo de datos, el examen de los residuos (la desviación de la línea ajustada a los valores observados) permite al modelador investigar la validez de que existe una relación lineal. El trazado de los residuos en el eje *y* en contra de la variable independiente en el eje *x* revela cualquier posible relación no lineal entre las variables, o puede alertar al modelador para investigar las variables que acechan. En este caso los residuos determinan la presencia de valores atípicos [YALE, 1998].

Los procedimientos de regresión lineal por mínimos cuadrados (LS), son sensibles a ciertos tipos de valores atípicos, inclusive si se trata de uno solo de estos valores. Según Rousseeuw y Leroy, se pueden presentar valores atípicos tanto en el eje *y* como en el eje *x*. En el eje *x* hay más posibilidades de que algo salga mal, su efecto en el estimador de mínimos cuadrados es muy significativo debido a su gran impacto en la pendiente [Rousseeuw y Leroy, 1996]. Para solucionar este problema, se han desarrollado nuevas técnicas estadísticas que no se ven fácilmente afectadas por los valores atípicos. Estos son los métodos robustos, que siguen siendo una técnica de confianza, incluso en una gran cantidad de datos [Rousseeuw y Leroy, 1996]. La regresión lineal robusta en vez de utilizar LS, utiliza el método *least median of squares* (LMS) definido por Rousseeuw [Rousseeuw, 1984]. Se reemplaza la suma de mínimos cuadrados por la mediana, que es un estimador robusto tanto para valores extremos en el eje *x* como en el eje *y*, y es resistente a situaciones multivariantes. El objetivo principal es ajustar la mayoría de los datos y luego los valores atípicos pueden ser identificados como los puntos que permanecen lejos de la regresión tanto para el caso de residuos positivos como negativos [Rousseeuw y Leroy, 1996].

5.2. Métrica de Evaluación para Técnicas de detección de valores atípicos.

Autores como Van den Broeck recomiendan contar con alto grado de conocimiento de los datos antes de realizar procesos de limpieza de datos [Van den Broeck, 2005]. Debe reconocerse que esto no se cumple a cabalidad en este caso ya que los datos de trabajo no son propios sino obtenidos de una fuente externa. Esta situación dificulta el análisis de los resultados. Así, es difícil juzgar la eficacia de las técnicas en cuanto a la identificación de los valores atípicos, por cuanto se requeriría alto conocimiento de los datos para decidir si las técnicas están acertando en su labor. Sin embargo, mediante gráficas de dispersión e histogramas de frecuencias, es posible establecer si una observación se asemeja o aleja de las demás, es decir, es posible intuir si se trata de un valor atípico el cual debe evaluarse con mayor detenimiento.

Con base en esto, para evaluar la eficacia de las diferentes técnicas para detección de valores atípicos, se utilizó como criterio la comparación con los atípicos detectados mediante inspección visual realizada en gráficas de dispersión entre las variables e histogramas de frecuencias de los datos.

5.3. Diseño del experimento para evaluar las diferentes técnicas.

Para comparar las cinco técnicas para detección de valores atípicos, se utilizó el mismo archivo con datos extraídos del censo de los Estados Unidos, utilizado para los valores faltantes.

Con el fin de determinar la efectividad de las diferentes técnicas para detección de valores atípicos, éstas se aplicaron a los atributos *Edad*, *Salario* y *Años_Estudio* utilizando para ello paquetes estadísticos. Cada técnica arrojó una determinada cantidad de valores atípicos.

5.4. Resultados del Experimento para Evaluación de Técnicas para detección de valores atípicos.

La figura 18 presenta gráficas de dispersión de la variable *Edad*, las cuales posibilitan observar la ubicación y distribución de los puntos y por tanto permiten identificar visualmente los valores atípicos arrojados por las diferentes técnicas. En los resultados no se incluye la prueba de Dixon, ya que esta sólo puede aplicarse para pequeños conjuntos de datos y por tanto no aplica para el caso seleccionado. La figura 19 presenta las gráficas de dispersión correspondientes a la variable *Salario* y la figura 20 a la variable *Años_estudio.*

La tabla 26 presenta la cantidad de valores atípicos detectada por cada una de las técnicas para las variables *Edad, Salario* y *Años_estudio*.

Tabla 26. Cantidad de valores atípicos detectados por las Técnicas

Variable	Edad			Salario			Años_estudio		
Técnica	**AL**	**AS**	**Valores**	**AL**	**AS**	**Valores**	**AL**	**AS**	**Valores**
Tukey	142		>78	839	151	>594727	1198		<5
Grubbs	43		=90		413	>490332		1610	<5, =16
Mahalanobis		61	>83		262	>537222		51	=1
Regresión		241	>75		393	>496414		219	<3

AL: Atípico Leve AG:Atípico Significativo

Figura 18. Gráficas de dispersión Edad vs Años_estudio

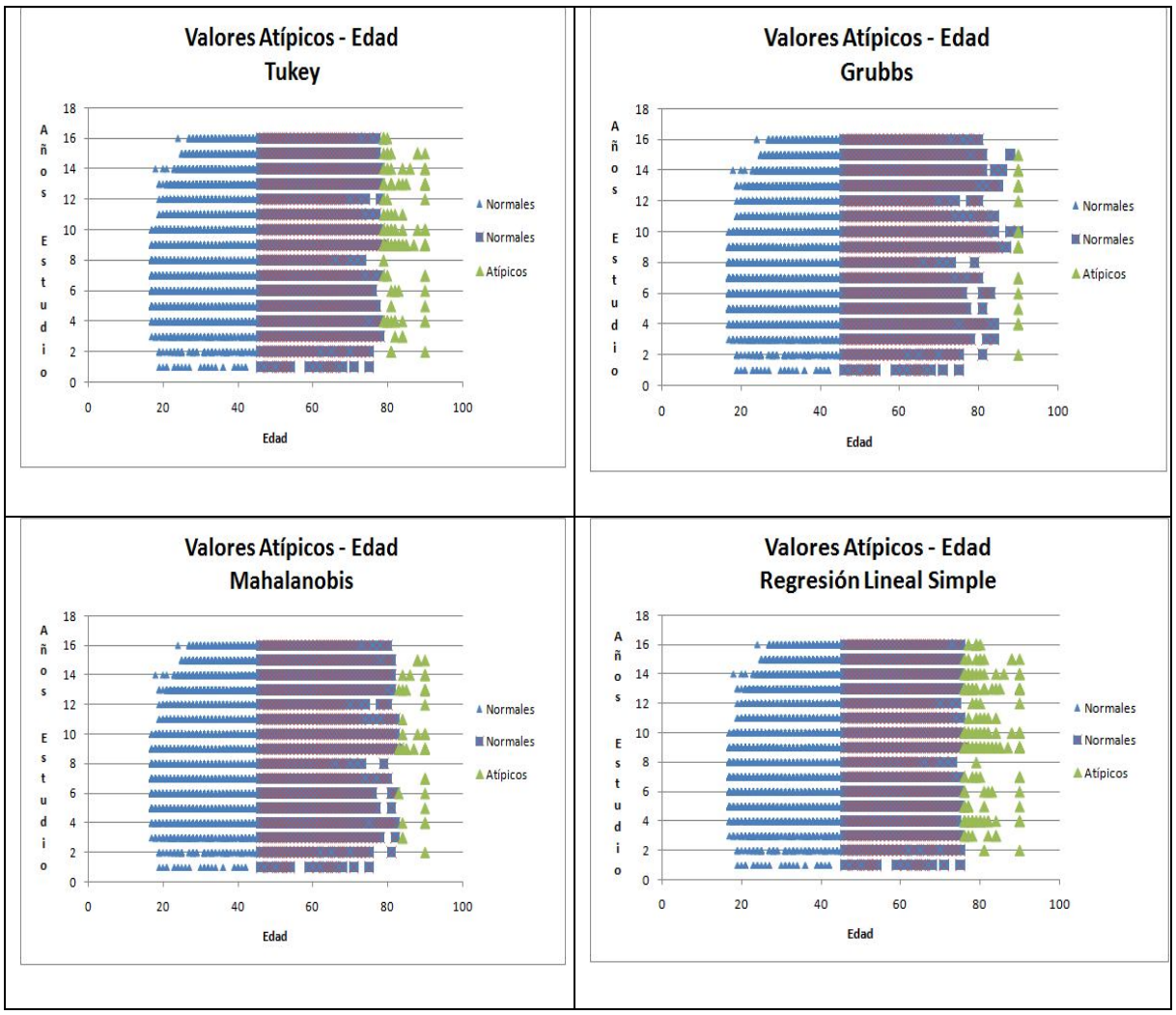


Figura 19. Gráfico de dispersión variable Salario vs Edad

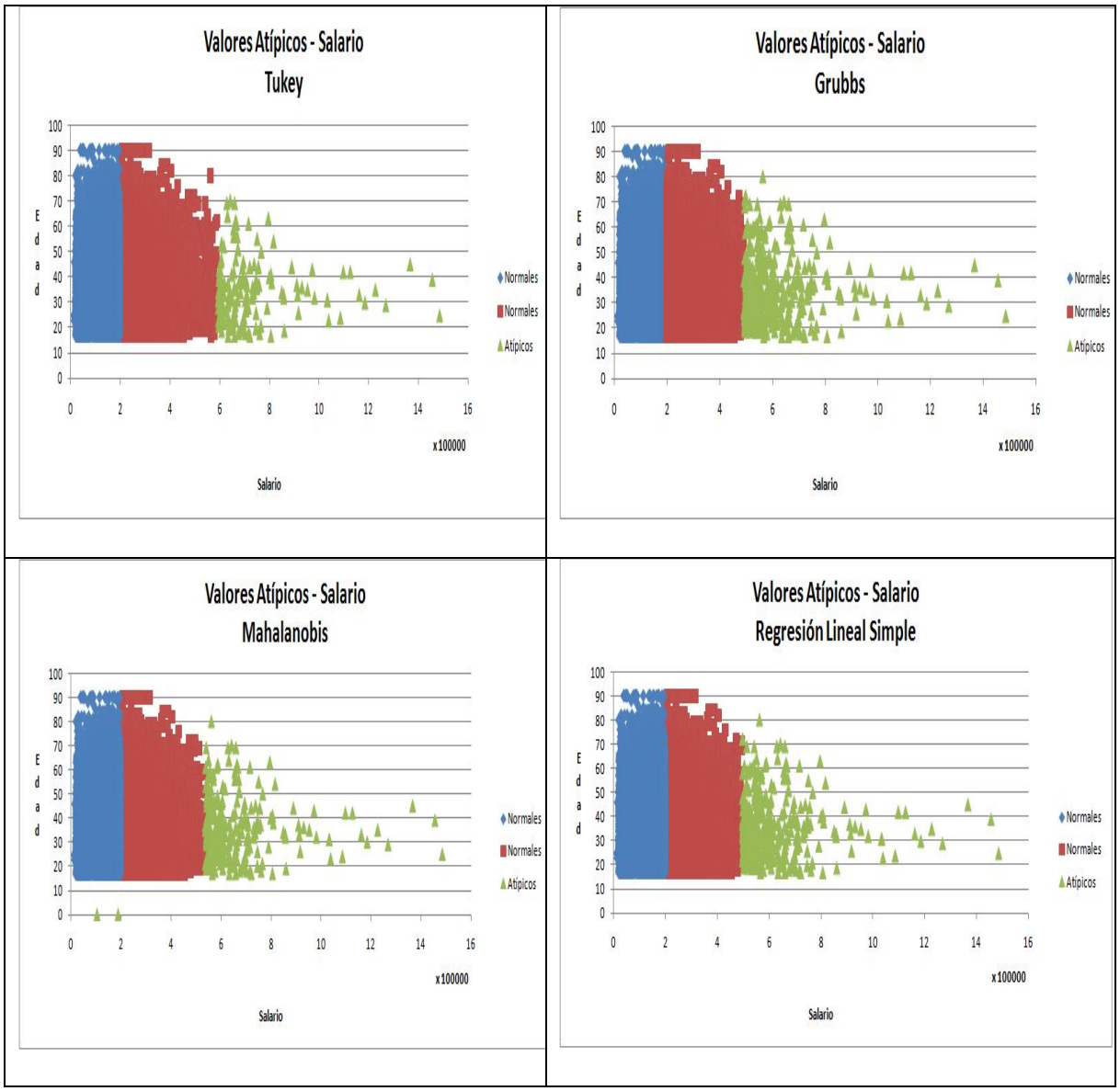
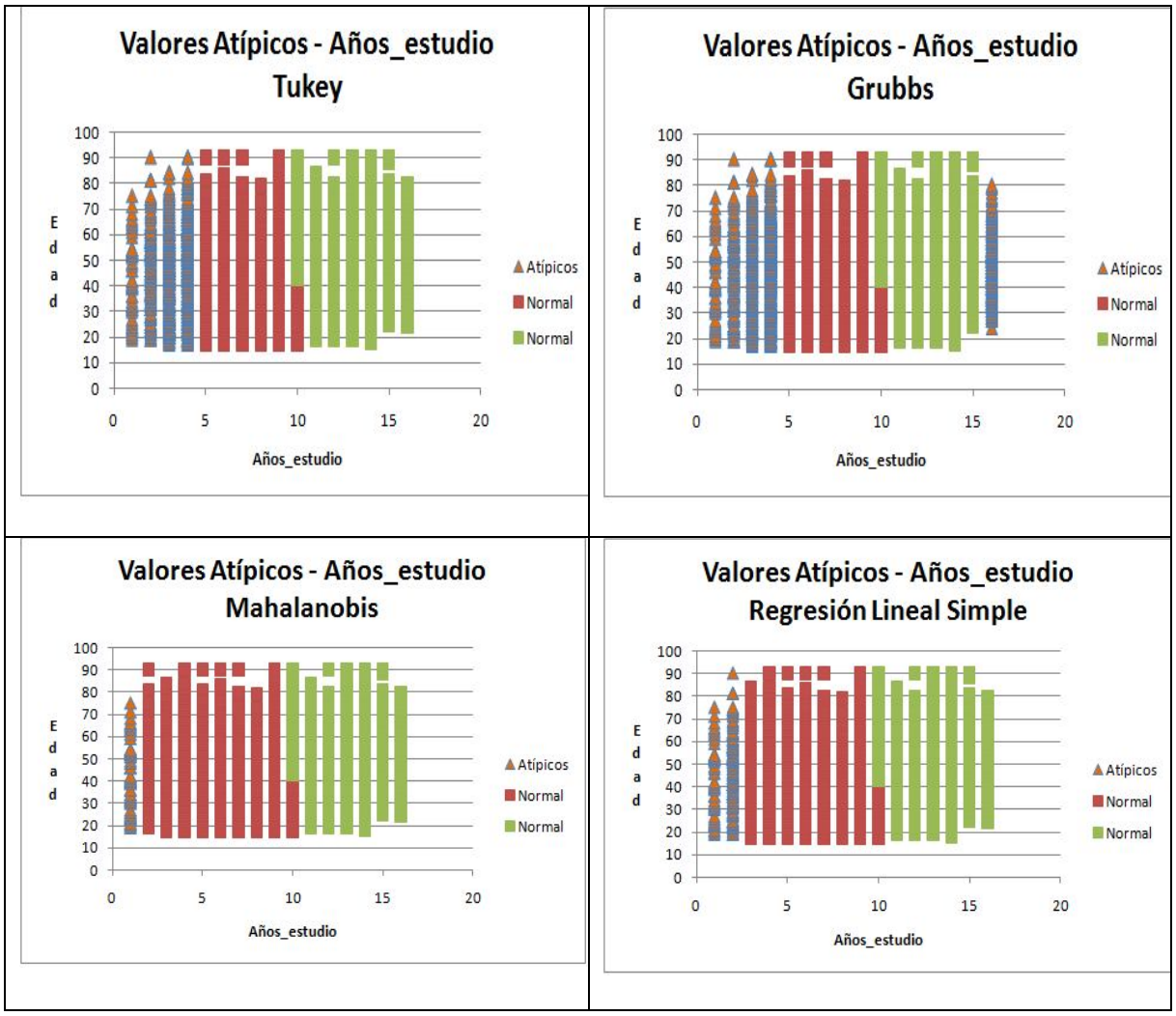
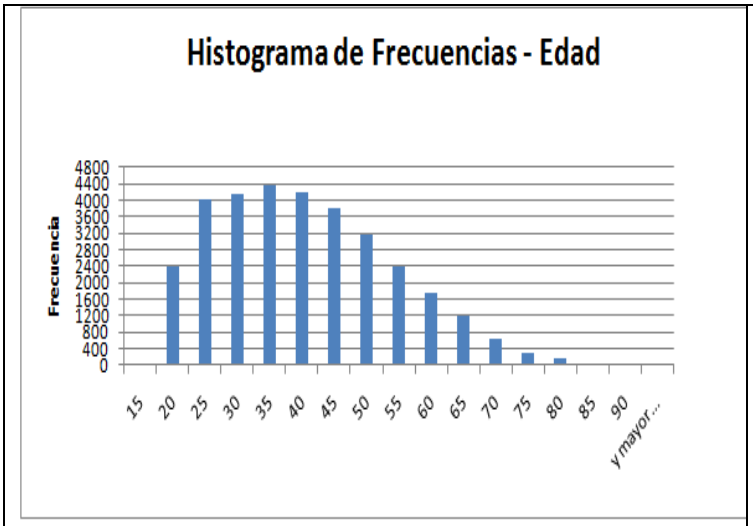


Figura 20. Gráfico de dispersión variable Años_estudio vs Edad

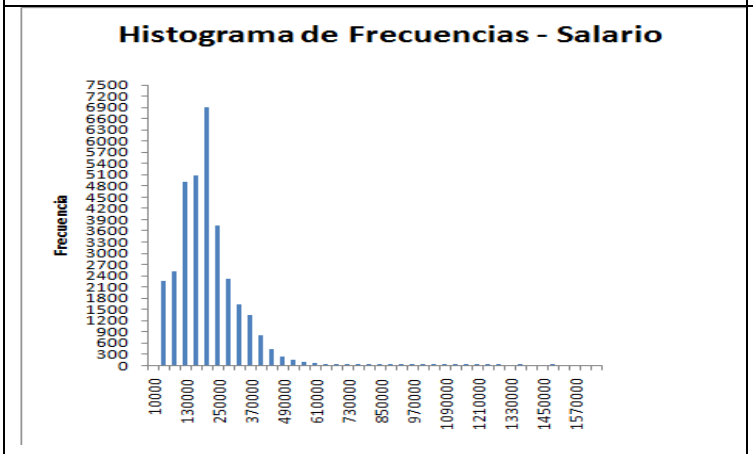


Debido a la gran cantidad de puntos existentes en los gráficos de dispersión (32561 puntos), no es fácil visualizarlos independientemente. La figura 21 presenta histogramas de frecuencias para las variables *Edad*, *Salario y Años_estudio*. Los histogramas permiten visualizar mejor la cantidad de veces que se repite un mismo valor, lo cual es indicativo de que tan atípico es un valor en el conjunto de datos. Las tablas a la derecha de los histogramas, corresponden a los valores que se presentan con menor frecuencia en los datos.

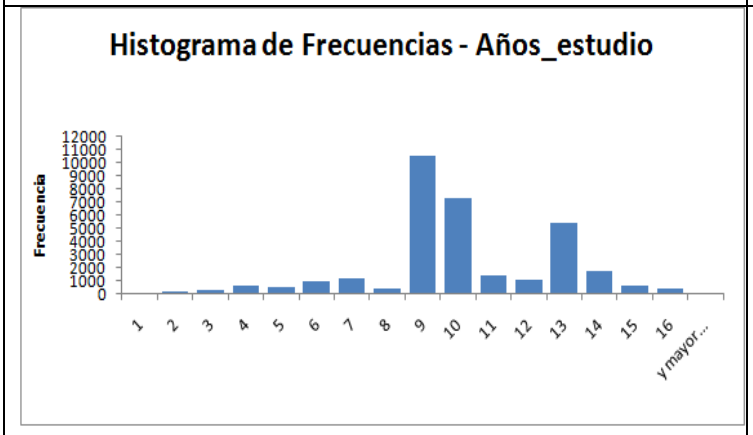
Figura 21. Histograma de Frecuencias variables Edad, Salario y Años_estudio



Edad	Frec (%)
76	0,141
77	0,089
78	0,071
79	0,068
80	0,068
81	0,061
82	0,037
83	0,018
84	0,031
85	0,009
86	0,003
87	0,003
88	0,009
89	0,000
90	0,132



Salario (x1000)	Frec (%)	Salario (x1000)	Frec (%)
450	1,907	950	0,012
500	0,832	1.000	0,009
550	0,455	1.050	0,009
600	0,252	1.100	0,006
650	0,129	1.150	0,003
700	0,111	1.200	0,006
750	0,080	1.250	0,003
800	0,028	1.300	0,003
850	0,015	1.400	0,003
900	0,012	1.500	0,006



Años estudio	Frec (%)
1	0,157
2	0,516
3	1,023
4	1,984
5	1,579
15	1,769
16	1,268

De los resultados anteriores, puede notarse que las diferentes técnicas arrojan resultados variables en cuanto a los valores atípicos identificados. Por ejemplo, para la variable *Años_estudio*, la técnica de Tukey arroja como resultado los valores menores a cinco años, Grubbs indica los menores de cinco años y aquellas personas con 16 años de estudio, Mahalanobis sólo indica aquellos datos con 1 año de estudio y Regresión Lineal Simple los menores de tres años (ver tabla 26). El análisis del histograma de frecuencias para esta variable (ver figura 21), muestra que la gran mayoría de los valores se encuentran entre 6 y 14 años de estudio, pudiéndose considerar que los atípicos reales en este caso son los menores de 6 años y los mayores de 14. La técnica que arrojó los resultados más cercanos, fue la técnica de Grubbs. Aunque la variable *Años_estudio* no se ajusta exactamente a una distribución normal, no está muy alejada como puede juzgarse por sus valores de asimétria y curtosis cercanos a ± 0.5 (Ver tabla 18). Tukey debido a que la mediana de los datos es 10 y la distribución es ligeramente sesgada hacia la derecha, no logra detectar los atípicos localizados en ese extremo de la distribución. La regresión lineal simple, al no haber una relación lineal entre ninguna de las variables, tampoco logra detectar los atípicos en ambos lados de la distribución. Mahalanobis, hace el trabajo de detección más deficiente debido a que no existe correlación entre las variables (ver tabla 22).

5.5. Guía Metodológica para Selección de Técnicas para Detección de Valores Atípicos.

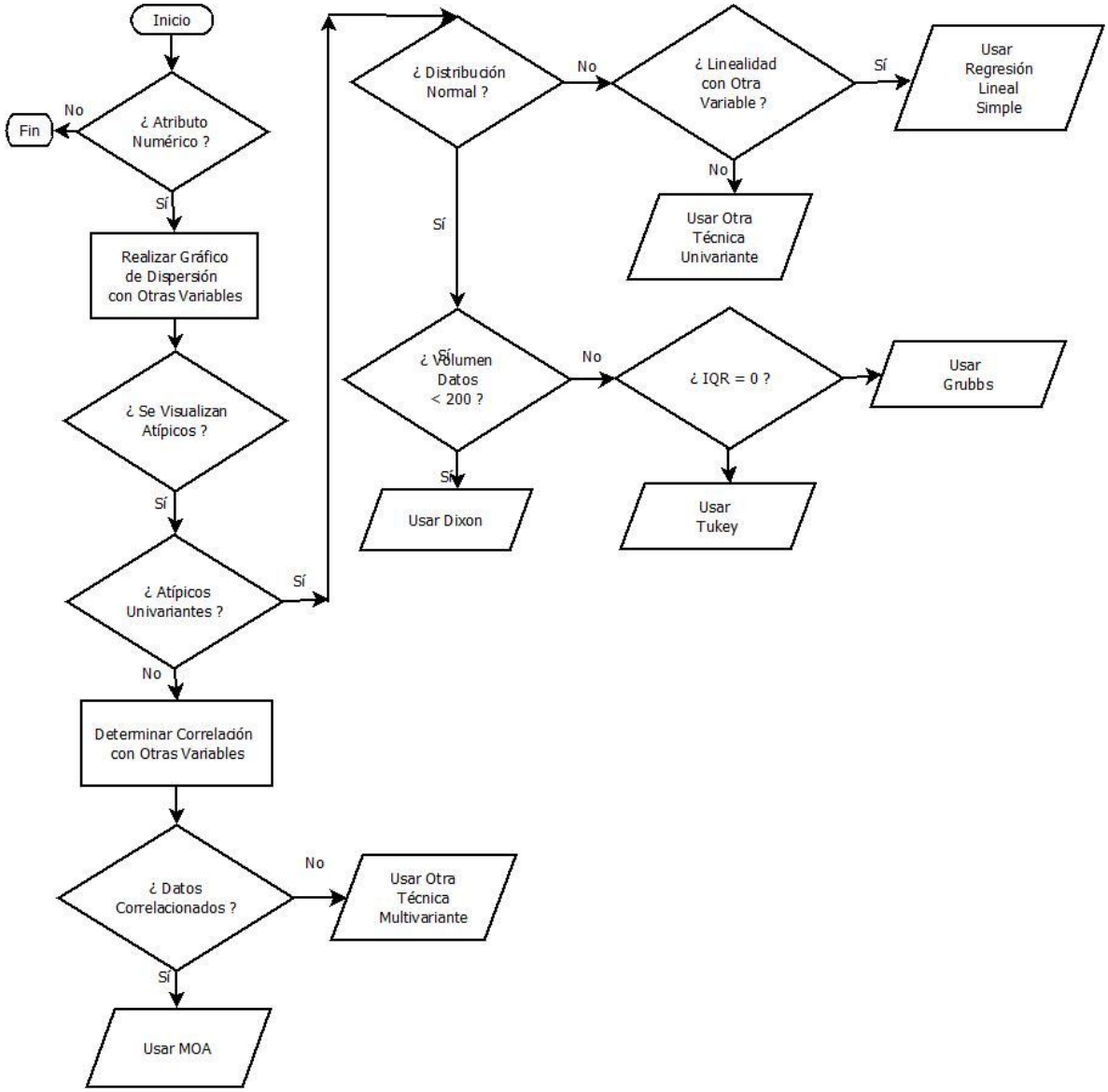
La figura 22 presenta la guía para el problema de la detección de valores atípicos, mediante un diagrama de flujo de datos.

El diagrama comienza indagando si el tipo de datos de un atributo al cual se desee hacer limpieza, es de tipo numérico, ya que las técnicas para detección de atípicos operan sobre este tipo de datos. Para atributos numéricos, la guía solicita realizar gráficos de dispersión con otras variables. Esta tarea, fácilmente realizable en Excel o en algún paquete estadístico, permite detectar visualmente valores alejados de los demás, en cuyo caso la respuesta a la pregunta ¿Se visualizan atípicos? sería verdadera. En caso contrario, el proceso termina.

A continuación, se indaga si los datos atípicos son univariantes. Para dar respuesta a esta pregunta el usuario debe entender el concepto de atípico univariante y multivariante. Recuérdese que un valor atípico univariante es un punto de datos que es muy diferente al resto para una sola variable. Un valor atípico multivariante, es un caso que es un valor extremo para una combinación de variables. Por tanto, el usuario debe evaluar si los datos son atípicos vistos individualmente o si lo son porque no son valores normales en relación con los valores que toman otras variables. Por ejemplo un salario de

dos millones de pesos colombianos puede no ser considerado atípico, pero seguramente lo es si corresponde a una persona de 15 años de edad, en cuyo caso se estaría en presencia de un atípico multivariante.

Figura 22. Diagrama para la selección de técnicas para Valores Atípicos



La única técnica disponible, entre las evaluadas, que es capaz de tratar atípicos multivariantes, es MOA (*Mahalanobis Outlier Analysis*). Esta técnica sólo debe

aplicarse si la variable bajo análisis está correlacionada con otras variables ya que la distancia de Mahalanobis toma en cuenta dicha correlación, por tanto esta es la técnica recomendada para esa situación. Para atípicos multivariantes, pero sin correlación con otras variables, debe aplicarse alguna otra técnica no estudiada en este trabajo.

Para la situación de atípicos univariantes, si se distribuyen normalmente y son pocos datos (menos de 200), se recomienda la prueba de Dixon. Para una cantidad de datos superior, debe verificarse si el IQR o rango intercuantil es cero, ya que en ese caso todos los valores diferentes de cero serán catalogados como atípicos. Si el IQR es diferente de cero, puede usarse la prueba de Tukey; si es igual a cero puede usarse la prueba de Grubbs. Para atípicos univariantes que no se distribuyen normalmente, si existe linealidad con otra variable, se recomienda regresión lineal simple. Si no existe linealidad, debe aplicarse alguna otra técnica no estudiada en este trabajo.

5.6. Conclusiones y Trabajo Futuro sobre Técnicas para Detección de Valores Atípicos.

Utilizar técnicas para detección de valores atípicos inadecuadas puede llevar a que no se detecten valores que son sospechosos o por el contrario a que se trate como sospechoso a un valor que no lo es. Si dichos valores se eliminan o imputan, puede tener graves consecuencias sobre la calidad de los datos y sobre los procesos realizados posteriormente. Existen diversas técnicas para tratar con los valores atípicos, pero deben aplicarse con buen juicio. Decidir cuál técnica utilizar en un caso particular no es un asunto trivial, el cual puede ser facilitado a los usuarios mediante una guía metodológica que recomiende la adecuada en una situación dada.

Este objetivo se logra en este trabajo, mediante el cumplimiento de los objetivos espécificos planteados, así:

- ✓ Se logró identificar cinco técnicas para detección de valores atípicos (prueba de Dixon, prueba de Grubbs, prueba de Tukey, Mahalanobis y regresión lineal simple).

- ✓ Se identificaron las características de cada técnica determinando sus propiedades, funcionalidad, fortalezas y debilidades.

- ✓ Se compararon las diferentes técnicas. Para esto fue necesario:

    - o Identificar una métrica para realizar la comparación (atípicos detectados mediante inspección visual realizada en gráficas de dispersión entre las variables e histogramas de frecuencias de los datos).

- Obtener un conjunto de datos de prueba (censo USA).
- Construir gráficos de dispersión e histogramas de frecuencias para varias variables.

✓ Se diseñó un diagrama que sirve como guía al usuario para la selección de la técnica más adecuada según la naturaleza de los datos a depurar.

Como trabajo futuro, se plantea incorporar a la guía otrás técnicas no analizadas.

6. REFERENCIAS BIBLIOGRÁFICAS

[Acock, 2005] Acock, C., Alan. Working With Missing values, *Journal of Marriage and Family* 67, November. 2005.

[Acuña y Rodríguez, 2009] Acuña, E., Rodríguez, C. The treatment of missing values and its effect in the classifier accuracy: Four different methods to deal with missing values S.p.i. 2-3p. 2009.

[Aluja, 2000] Aluja, T. Classification and Regresion Trees. Curso Árboles de Decisión. Universidad Politécnica de Cataluña. 2000.

[Ampanthong y Suwattee, 2009] Ampanthong, p. y Suwattee, P. A Comparative Study of Outlier Detection Procedures in Multiple Linear Regression. Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, Marza 18 - 20, 2009, Hong Kong.

[Anderson *et. al.*, 1983] A.B. Anderson, A. Basilevsky, and D.P.J. Hum, Missing Data: Review of the Literature,º Handbook of Survey Research. P.H. Rossi, J.D. Wright and A.B. Anderson eds., New York: Academic Press, pp. 415-492, 1983.

[Angiulli et. Al., 2006] Angiulli, F., Basta, S., and Pizzuti, C. 2006. Distance-Based Detection and Prediction of Outliers. *IEEE Trans. on Knowl. and Data Eng.* Vol. 18, No. 2 (Feb. 2006), pp. 145-160. DOI= http://dx.doi.org/10.1109/TKDE.2006.29

[Álvarez, 2007] Álvarez R. Estadística aplicada a las ciencias de la salud. Ed. Diaz de Santos. 2007. 1030p.

[Arslan y Eğecioğlu, 2001] Arslan, A.N. y Eğecioğlu, O. "A New Approach to Sequence Comparison: Normalized Sequence Alignment", Bioinformatics, vol. 17, no. 4, pp. 327-337, 2001.

[Ávila, 2002] Ávila, C. Una aplicación del procedimiento Hot Deck como método de imputación. Lima, 2002. Trabajo de grado. (Licenciado en estadística). Universidad Nacional Mayor de San Marcos. Facultad de Ciencias Matemáticas.

[Badler et. al., 2005] Badler, C., Alsina, S., Puigsubirá, C. y Vitelleschi, M. Utilización de metodología para el tratamiento de información Faltante y/o confusa en el diagnóstico de la desocupación. Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística (IITAE). 2005.

[Baeza-Yates y Gonnet, 1992] Baeza-Yates, R., y Gonnet, G.H. "A new approach to Text Searching", Communications of the ACM, vol. 35, no. 10, pp. 74-82, 1992.

[Bakar et. al., 2006] Bakar, Z., Ahmad, M., y Deris, M. A. 2006. Comparative Study for Outlier Detection. En: IEEE Conference on Cybernetics and Intelligent Systems CIS 2006 (Bangkok, Tailandia, Junio 7-9, 2006).

[Barnett y Lewis, 1984] Barnett, V. y Lewis, T. Outliers in statistical data, 2nd Edición. New York, John Wiley & Sons. 1984.

[Barnett y Lewis, 1994] Barnett, V. y Lewis, T. Outliers in statistical data, 3rd edition. Chichester, John Wiley & Sons, 1994, 584 pp.

[Becker y Gather, 1999] Becker, C. y Gather, U. (1999) The masking breakdown point of multivariate outlier identification rules. Journal of the American Statistical Association, 94, pp. 947-955.

[Beckman y Cook, 1983] Beckamn, R. J. y Cook, R. D. Outlier…….s, Technometrics Vol 25, No. 2. pp 119-149.

[Ben-Gal, 2005] Ben-Gal, I. Outlier Detection, En: Maimon, O. y Rockach, L. "Data Mining and Knowledge Discovery Handbook: A complete Guide for Practitioners and Researchers." Kluwer Academic Publishers, 2005.

[Berk, 1987] Berk, K. (1987). "Computing incomplete repeated measures", Biometrics, 43, 269-291.

[Bilenko y Mooney, 2003] Bilenko, M. y Mooney, R.J. "Learning to Combine Trained Distance Metrics for Duplicate Detection in Databases", Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 39-48, 2003.

[Breimer y Goldberg, 2002] Breimer, E. y Goldberg, M. "Learning Significant Alignments: An Alternative to Normalized Local Alignment", Proceedings of the 13th International Symposium on Foundations of Intelligent Systems, pp. 37-45, 2002.

[Breuning et. Al., 2000] Breuning, M., Kriegel, H., Ng, R. y Sander, J. 2000. Lof: Identifying density-based local outlier. En: Proceedings SIGMOD 2000 (Dallas, Texas, Mayo 14-19, 2000), 93-104.

[Brick et. al., 2005] Brick, J., Jones, M., Kalton, G. y Valliant, R. Variance Estimation with Hot Deck Imputation: A Simulation Study of Three Methods, Survey M ethodology, Vol 31, pp 151-159.

[Brown, 1994] R.L. Brown, Efficacy of the Indirect Approach for Estimating Structural Equation Models With Missing Data: A Comparison of Five Methods, Structural Equation Modeling, vol. 1, no. 4, pp. 287-316, 1994.

[Browne, 1983] C.H. Browne, Asymptotic Comparison of Missing Data Procedures for Estimating Factor Loadings, Psychometrics, vol. 48, no. 2, pp. 269-291, 1983.

[Cañizares et. al., 2004] Cañizares, M. Barroso, I., y Alfonso, K. Datos incompletos: una mirada crítica para su manejo en estudios sanitarios. *Gac Sanit* [online]. 2004, vol.18, n.1, pp. 58-63. ISSN 0213-9111.

[Cartwright et. al., 2003] Cartwright, M., Shepperd, M.J., and Song, Q. (2003). "Dealing with Missing Software Project Data", In *Proc. of the 9th Int.Symp .on Software Metrics 2003*, 154-165.

[Chambers et. al., 1983] Chambers, J., Cleveland, W., Kleiner, B. y Tukey, P. Graphical Methods for Data Analysis, Wadsworth. 1983.

[Chandola et. al., 2007] Chandola, V., Arindam, B., y Vipin K,. 2007. Outlier detection: A survey. Technical Report Department of Computer Science and Engineering. University of Minnesota. Agosto 15, 2007.

[Chandola et. al., 2009] Chandola, V., Arindam, B., y Vipin K,. 2009. Anomaly detection: A survey. ACM Computing Surveys. Septiembre 2009.

[Christen, 2006] Christen, P. "A Comparison of Personal Name Matching: Techniques and Practical Issues", Sixth IEEE International Conference on Data Mining, pp. 290-294, 2006.

[Cohen, 1997] Cohen, D. "Recursive Hashing Functions for n-Grams", ACM Transactions on Information Systems, vol. 15, no. 3, pp. 291-320, 1997.

[Cohen, 1998] Cohen, W.W. 1998. Integration of Heterogeneous Databases without Common Domains Using Queries Based on Textual Similarity. En: Proceedings of the SIGMOD International Conference Management of Data SIGMOD'98 (Seattle, Washington, Junio 2-4, 1998), 201-212.

[Cohen et. al., 2003] Cohen, W.W., Ravikumarand, P. Fienberg, S.E. "A Comparison of String Distance Metrics for Name-Matching Tasks", International Joint Conference on Artificial Intelligence, pp. 73-78, 2003

[CQU, 1997] CQUniversity. The exploring data. How to Draw a Boxplot [En linea]. Education Queensland, 1997.
<http://exploringdata.cqu.edu.au/box_draw.htm> [Consulta: 18 Nov. 2009].

[CRISP-DM, 2000] CRISP-DM Consortium. CRISP-DM 1.0 Step-by-step data mining guide. Chicago, IL.: SPSS Inc., 2000, 13

[da Silva et. al., 2007] da Silva, R., Stasiu, R., Orengo, V.M. y Heuser, C.A. "Measuring Quality of Similarity Functions in Approximate Data Matching", Journal of Informetrics, vol. 1, no. 1, pp. 35-46, 2007.

[Dasgupta y Forrest, 1996] Dasgupta, D. y Forrest, S. Novelty Detection in Time Series Data Using Ideas from Immunology. En: *Proceedings of the Fifth International Conference on Intelligent Systems*. 1996.

[Dasu et. al., 2003] Dasu, T., Vesonder, G. T., y Wright, J. R. 2003. Data quality through knowledge engineering. En: Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining ACM SIGKDD 2003 (Washington, D.C., Agosto 24 - 27, 2003). KDD '03. ACM, Nueva York, NY, 705-710. DOI= http://doi.acm.org/10.1145/956750.956844

[Davis y McCuen, 2005] Davis, A. y McCuen, R. StormWater Management for Smart Growth. Springer. 2005. p.58.

[DeCoste y Levine, 2000] DeCoste, D. y Levine, M. B. Automated Event Detection in Space Instruments: A Case Study Using IPEX-2 Data and Support Vector Machines. En: *Proceedings of the SPIE Conference on Astronomical Telescopes and Space Instrumentation*. 2000.

[Dixon y Massey, 1983] Dixon, W., Massey, F. Introduction to Statistical Analysis (Fourt edition), Edited by Wilfrid J. Dixon. McGraw-Hill Book Company, New York, 1983. pp.377.

[Dunn, 1946] Dunn, H.L. "Record Linkage", Americal Journal of Public Health, vol. 36, no. 12, pp. 1412-1416, 1946.

[EAM, 2008] Departamento Administrativo Nacional de Estadística DANE. Encuesta anual manufacturera 2008.

[Edwards, 1976] Edwards, A. An Introduction to Linear Regression and Correlation. San Francisco, CA. 1976.

[Eğecioğlu e Ibel, 1996] Eğecioğlu, O. y Ibel, M. "Parallel Algorithms for Fast Computation of Normalized Edit Distances", Proceedings of the 8th IEEE Symposium on Parallel and Distributed Processing, pp. 496-503, 1996.

[Elmagarmid et. al. 2007] A.K. Elmagarmid, P.G. Ipeirotis and V.S. Verykios, "Duplicate Record Detection: A Survey", IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, 2007.

[El-Emam y Birk, 2000] K.E. Emam and A. Birk, Validating the ISO/IEC 15504 Measure of Software Requirements Analysis Process Capability, IEEE Trans. Software Eng., vol. 26, no. 6, pp. 541-566, June 2000.

[Ertoz et. al., 2003] Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P., Dokas, P., Kumar, V. y Srivastava, J. Detection of novel network attacks using data mining. En: *Proceedings of the 2003 ICDM Workshop on Data Mining for Computer Security*, Melbourne, Florida, USA, November 2003.

[Farhangfar et. al., 2007] Farhangfar, A., Kurgan, L.A. y Pedricks, W. A Novel Framework for Imputation of Missing Values in Databases. 2007.1p. (692)

[Fellegi y Sunter, 1969] I.P Fellegi y A.B. Sunter, "A Theory for Record Linkage", Journal of the American Statistical Association, vol. 64, no. 328, pp. 1183-1210, 1969.

[Filzmoser, 2004] P. Filzmoser. A multivariate outlier detection method. En: Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling, Vol. 1, pp. 18-22, Belarusian State University, Minsk, 2004.

[Gartner, 2007] Gartner. Dirty Data is a Business Problem, Not an IT Problem. [En línea]. 2007. http://www.gartner.com/it/page.jsp?id=501733 [Consulta: Octubre 10 de 2008]

[Gelbukh et. al., 2009] Gelbukh, A., Jiménez, S., Becerra, C. y González, F. "Generalized Mongue-Elkan Method for Approximate Text String Comparison", Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing, pp. 559-570, 2009.

[Gill, 1997] Gill, L.E. 1997. OX-LINK: The Oxford Medical Record Linkage System. En: Proceedings of International Record Linkage Workshop and Exposition, (Arlington, Estados Unidos, Marzo 20-21, 1997), 15-33.

[Goicoechea, 2002] Goicoechea, P. Imputación basada en árboles de clasificación. EUSTAT. 2002.

[Gold y Bentler, 2000] M.S. Gold and P.M. Bentler, Treatment of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximation, Structural Equation Modeling, vol. 7, no. 3, pp. 319-355, 2000.

[Gotoh, 1982] Gotoh, O. "An Improved Algorithm for Matching Biological Sequences", Journal of Molecular Biology, vol. 162, no. 3, pp. 705-708, 1982.

[Grahramani y Jordan, 1997] Ghahramani, Z. y Jordan, M.I. "Mixture models for learning from incomplete data," in *Computational Learning Theory and Natural Learning Systems, vol. 4, Making Learning Systems Practical*, R. Greiner, T. Petsche, and S. J. Hanson, Eds. Cambridge, MA: MIT Press, 1997, pp. 67–85.

[Grubbs, 1969] Grubbs, F. Procedures for Detecting Outlying Observations in Samples, Technometrics, Vol 11, No. 1, pp 1-21.

[Guilford y Fruchter, 1984] Guilford JP, Fruchter B. 1984. Métodos y problemas especiales de correlación. En: Estadística aplicada a la sicología y la educación. Editorial MacGraw–Hill. p. 265–333.

[Han y Kamber, 2006] Han, J. y Kamber, M. Data Mining: Concepts and Techniques. 2nd Ed. Morgan Kaufmann Publishers, 2006.

[Hanson, 1978] Hanson, R. The Current Population Survey. Bureau of the Census, Technical paper. 1978.

[Hawkins, 1980] Hawkins, D. M. Identification of Outliers. London, Chapman & Hall. 1980.

[Heuser et. al., 2007] Heuser, C.A., Krieser, F.N. y Orengo, V.M. "SimEval - A Tool for Evaluating the Quality of Similarity Functions", Tutorials, posters, panels and industrial contributions at the 26th International Conference on Conceptual Modeling, pp. 71-76, 2007.

[Hintze y Nelson, 1998] Hintze JL, Nelson RD. Violin Plots: A Box Plot-Density Trace Synergism. The American Statistician, Vol. 52, No. 2, pp.181-184.

[Hodge y Austin, 2004] Hodge, V. y Austin, J. A Survey of Outlier Detection Methodologies. *Artif. Intell. Rev.* Vol. 22, No. 2 (Oct. 2004), pp. 85-126. DOI= http://dx.doi.org/10.1023/B:AIRE.0000045502.10941.a9

[Hofmann *et. al.*, 2006] Hofmann, H., Kafadar, K. y Wickham, H. Letter Value Boxplot. Iowa state university. 2006.

[Huber y Vandervieren, 2008] Hubert, M. and Vandervieren, E. 2008. An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal.* 52, 12 (Aug. 2008), 5186-5201. DOI= http://dx.doi.org/10.1016/j.csda.2007.11.008

[Hyyrö, 2002] Hyyrö, H. "A Bit-Vector Algorithm for Computing Levenshtein and Damerau Edit Distances". The Prague Stringology Conference '02.

[Iglewicz y Hoaglin, 1993] Iglewicz, B. y Hoaglin, D. How to detect and handle outliers. American Society for Quality. Statistics Division. 1993.

[ITL, 2006] Information Technology Laboratory Itl. Statistical Reference Data Sets Archives. 2006.
Disponible en: http://www.itl.nist.gov/div898/strd/general/dataarchive.html

[Jaro, 1976] Jaro, M.A. 1976. Unimatch: A Record Linkage System: User's Manual, technical report, US Bureau of the Census, Washington, D.C.

[Johnson, 1989] Johnson, E.G. (1989). "Considerations and techniques for the analysis of NAEP data", *Journal of Educational Statistics.*, 14, 303-334.

[Juárez, 2003] Juárez, C. Fusión de Datos: Imputación y Validación. S.p.i.. Trabajo de grado. Universitad politécnica de Cataluña. Departamento de estadística e investigación operativa.

[Juster y Smith, 1998] Juster, F. T. y Smith, J. P. Improving the quality of economic data, Lesson from the HRS and AHEAD, *Journal of the American Statistical Association*. 1998.

[Kalton y Kasprzyk, 1982] Kalton, G. y D. Kasprzyk (1986), The treatment of missing survey data, Survey Methodology, Vol. 12. (1982), Imputing for Missing Surveys Responses, Proceedings of the Section on Survey Research Methods, American Statistical Association.

[Kaufman, 1998] Kaufman, C.J. (1988). "The application of logical imputation to household measurement", *Journal of the Market Research Society*, 30, 453-466.

[Keskustalo et. al., 2003] Keskustalo, H., Pirkola, A., Visala, K., Leppänen, E. y Järvelin, K. "Non-adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants", International Symposium on String Processing and Information Retrieval, pp. 252-256, 2003.

[Kim et. al.,2003] Kim, W., Choi, B.J., Hong, E.K., Kim, S.K., y Lee, D. 2003. A Taxonomy of Dirty Data. Data Mining and Knowledge Discovery, 7, 2003. 81-99.

[Kurgan et. al., 2005] Kurgan, L. A., Cios, K. J., Sontag, M. y Accurso, F. J. "Mining the cystic fibrosis data," in *Next Generation of Data-Mining Applications*, J. Zurada and M. Kantardzic, Eds. Piscataway, NJ: IEEE Press, 2005, pp. 415–444.

[Laaksonen, 2000] Laaksonen, S. *How to Find de Best Imputation Technique? Tests with Various Methods."* Statistics Finland.

[Lakshminarayan et. al., 1999] Lakshminarayan, K., Harp, S. A. y Samad, T. "Imputation of missing data in industrial databases," *Appl. Intell.*, vol. 11, no. 3, pp. 259–275, Nov./Dec. 1999.

[Lavrakas, 2008] Lavrakas, P. Encyclopedia of Survey Research Methods. SAGE publications. 2008.

[Levenshtein, 1966] Levenshtein, V.I. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals", Soviet Physics Doklady, vol. 10, no. 8, pp. 707-710, 1966.

[Li y Edwards, 2001] D. Li y E. Edwards. Automatic Estimation of Dixon's Test for Extreme Values Using a SAS Macro Driven Program. PharmaSug 2001. Toronto.

[Little y Rubin, 1987] Little, R. y Rubin, D. Statistical Analysis with missing data. New York: John Wiley & Sons, 1987.

[Little y Rubin, 1990] Little, R. y Rubin, D. The Analysis of Social Science Data with Missing Values, Sociological Methods and Research, vol. 18, nos. 2/3, pp. 292-326, Feb. 1990.

[Little y Rubin, 2002] Little, R. y Rubin, D. Statistical Analysis with missing data. Second Edition. New York: John Wiley & Sons, 2002.

[Lowrence y Wagner, 1975] Lowrence, R. y Wagner, R.A. "An Extension of the String-to-String Correction Problem", Journal of the ACM, vol. 22, no. 2, pp. 177-183, 1975.

[Maesschlck et. al., 2000] Maesschalck R., Jouan-Rimbaud, D. y Massart, D. The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems, Vol. 50, No. 1, Enero 2000, pp. 1-18.

[Magnani, 2004] Magnani, M. Techniques for Dealing with Missing Data in Knowledge Discovery Tasks. http://magnanim.web.cs.unibo.it/data/pdf/missingdata.pdf. 2004.

[Markou y Singh, 2003] Markou, M. y Singh, S. Novelty detection: a review-part 2: neural network based approaches. Signal Processing, Vol 83, No. 12. Pp.2499-2521. 2003.

[Martínez, 2005] Martínez, C. Estadística y Muestreo. 12ª Edición. Ecoe editores, 2005, 1100p.

[Marzal y Vidal, 1993] Marzal, A. y Vidal, E. "Computation of Normalized Edit Distance and Applications", IEEE Transactions on Patern Analysis and Machine Intelligence, vol. 5, no. 9, 1993.

[Masek, 1980] Masek, W.J. "A Faster Algorithm for Computing String Edit Distances", Journal of Computer and System Sciences, vol. 20, pp. 18-31, 1980.

[Matsumoto et. al., 2007] Matsumoto, S., Kamei, Y., Monden, A., y Matsumoto, K. 2007. Comparison of Outlier Detection Methods in Fault-proneness Models. En: Proceedings of the First international Symposium on Empirical Software Engineering and Measurement ESEM 2007 (Madrid, España, Septiembre 20 - 21, 2007). IEEE Computer Society, Washington, DC, 461-463. DOI= http://dx.doi.org/10.1109/ESEM.2007.34

[McGill et. al., 1978] McGill R, Tukey JW, Larsen WA . Variations of Box Plots. The American Statistician, Vol. 32, No. 1, 1978. pp.12-16.

[Mcknight et. al., 2007] MCKNIGHT E, Patrick; MCKNIGHT M, Katherine, SIDANI, Souraya, FIGUEREDO, Aurelio Jose. Missing Data: A gentle Introduction. S.p.i. 179 p. 2007

[Medina y Galván, 2007] Medina, F., y Galván, M. Imputación de datos: Teoría y práctica: Patrones de comportamiento de los datos omitidos. Serie Estudios estadísticos y prospectivos CEPAL. División de estadística y proyecciones económicas. Chile. Julio 2007. p26-27.

[Metha et. al., 2007] Mehta, K., Rustagi, M., Kohli, S. y Tiwari, s. Implementing Multiple Imputation in an Automatic Variable Selection. Statistics and Data Analysis. NorthEast SAS Users Group Conference NESUG 2007.

[Monge y Elkan, 1996] Monge, A.E. y Elkan, C.P. "The Field Matching Problem: Algorithms and Applications", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 267-270, 1996.

[Montgomery y Runger, 2003] Montgomery, D. y Runger, G. Applied Statistics and Probability for Engineers. Third Edition. John Wiley & Sons, Inc. 2003.

[Moreau et. al., 2008] Moreau, E., Yvon, F. y Cappé, O. "Robust Similarity Measures for Named Entities Matching", Proceedings of the 22nd International Conference on Computational Linguistics, pp. 593-600, 2008.

[Morillas y Díaz, 2007] Morillas, A. y Díaz, B. El problema de los outliers multivariantes en el análisis de sectores clave y cluster industrial. II Jornadas Españolas de Análisis Input – Output. Zaragosa. 2007.

[Müller y Freytag, 2003] Müller, H., y Freytag, J.C. 2003. Problems, Methods, and Challenges in Comprehensive Data Cleansing. Technical Report HUB-IB-164, Humboldt University, Berlin.

[Musil et. al., 2002] Musil, C.M., Warner, C.B., Yobas, P.K., and Jones, S.L. (2002). "A Comparison of Imputation Techniques for handling Missing Data", *Western Journal of Nursing Research*, 24 (5), 815-829.

[Myrtveit *et. al.*, 2001] Myrtveit, I., Stensrud, E., y Olsson, U. Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods. Discussion of MDTs. 2001. p10-11 (1009)

[Narita y Kitagawa, 2008] Narita, K., y Kitagawa, H. 2008. Outlier Detection for Transaction Databases using Association Rules. En: Proceedings of the Ninth International Conference on Web-Age Information Management iiWAS2007, (Jakarta, Indonesia, Diciembre 3-5, 2007).

[Needleman y Wunsh, 1970] Needleman, S.B. y Wunsh, C.D. "A General

Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins", vol. 48, no. 3, 1970.

[Newcombe, et. al., 1959] Newcombe, H., Kennedy, J., Axford, S. y James, A. "Automatic Linkage of Vital Records", Science, vol. 130, no. 3381, pp. 954-959, 1959.

[Newcombe y Kennedy, 1962] Newcombe, H y Kennedy, J. "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information", Communications of the ACM, vol. 5, no. 11, 1962.

[Nisselson et. al, 1983] Nisselson, H., Madow, W., Olkin, I. Incomplete Data in sample Surveys: Treatise. Wonder Book. New York. 1983.

[NIST, 2003] NIST/SEMATECH e-Handbook of Statistical Methods. Grubbs' Test for Outliers: [En linea]: 2003 <http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm> [Consulta: 20 Nov. 2009].

[Olinsky et. al., 2002] Olinsky, A., Chen, S. y Harlow, L. The comparative efficacy of imputation methods for missing data in structural equation modeling. European Journal of Operational Research. 2002.

[Oliveira et.al., 2005] Oliveira, P., Rodrigues, F., Henriques, P., y Galhardas, H. 2005. A Taxonomy of Data Quality Problems. En: Second International Workshop on Data and Information Quality IQIS 2005 (Porto, Portugal, Junio 13-17, 2005).

[Osborne et. al., 2001] Osborne, J. W., Christiansen, W. R. I., & Gunter, J. S. (2001). *Educational psychology from a statistician's perspective: A review of the quantitative quality of our field.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

[Palomino, 2004] Palomino, R. Notas breves sobre estadística descriptiva., Escuela de Estadística, Universidad Nacional de Colombia. 2004.

[Peat y Barton, 2005] Peat, J. y Barton, B. Medical Statistics: A guide to data analysis and critical appraisal. Blackwell Publishing. 2005.

[Petrovskiy, 2003] Petrovskiy, M. Outlier Detection Algorithms in Data Mining Systems. Programming and Computer Software. Vol 29, No. 4. pp.228-237. 2003.

[Philips, 1990] Philips, L. 1990. Hanging on the Metaphone, Computer Language Magazine, 7(12), 39-44, Diciembre, 1990.

[Philips, 2000] Philips, L. 2000. The Double Metaphone Search Algorithm," C/C++ Users J., 18(5), Junio, 2000.

[Poirier y Rudd, 1983] Poirier, D.J. and Rudd, P.A. (1983). "Diagnostic testing in missing data models", *International. Economic Review*, 24, 537-546.

[Pollock y Zamora, 1984] Pollock, J.J. y Zamora, A. "Automatic Spelling Correction in Scientific and Scholarly Text", Communications of the ACM, vol. 27, no. 4, pp. 358-368, 1984.

[Rahm y Do, 2000] Rahm, E., y Do, H. H. 2000. Data Cleaning: Problems and Current Approaches. IEEE Bulletin of the Technical Committee on Data Engineering, 24 (4).

[Ramírez y López, 2006] Ramírez, F. y López, E. (2006) "Spelling Error Patterns in Spanish for Word Processing Applications", *Proceedings of Fifth international conference on Language Resources and Evaluation, LREC 2006*.

[Rasmussen, 1988] Rasmussen, J. L. Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D. *Multivariate Behavioral Research, Vol. 23 No.* 2, pp. 189-202. 1988.

[Ristad y Yianilos, 1998] Ristad, E., y Yianilos, P. 1998. Learning string edit distance. 1998. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 5, pp. 522-532, Mayo, 1998.

[Rittman, 2006] Rittman, M. Data Profiling and Automated Cleansing Using Oracle Warehouse Builder 10*g* Release 2, Septiembre, 2006.

[Roberts y Tarassenko, 1995] Roberts, S. and Tarassenko, L.: 1995, 'A probabilistic resource allocating network for novelty detection'. *Neural Computation* **6**, 270–284.

[Robiah et. al, 2003] Robiah, A., Setan, H. y Mohd, M. Multiple Outliers detection Procedures in linear Regression. Matematika, Vol 19, No. 1. 2003, pp. 29-45.

[Rousseeuw, 1984] Rousseeuw, P. J. Least median of squares regression. 1984. Journal of the American Statistical Association. USA.

[Rousseeuw y Leroy, 1987] Rousseeuw, P. y Leroy, A. Robust Regression and Outlier Detection. 2a Ed. New York, John Wiley & Sons, 1987.

[Rousseeuw y Leroy, 1996] Rousseeuw, P. y Leroy, A. Robust Regression and Outlier Detection. 3a Ed. New York, John Wiley & Sons, 1996.

[Rousseeuw y Van Driessen, 1999] Rousseeuw, P.J. y Van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics, Vol. 41, pp. 212-223.

[Rousseuw y Van Zomeren, 1990] Rousseeuw, P.J. y Van Zomeren, B.C. (1990) Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, Vol. 85, pp. 633-639.

[Rosenthal, 2001] Rosenthal, A., Wood, D., y Hughes, E. 2001. Methodology for Intelligence Database Data Quality. Julio, 2001.

[Rovine y Delaney, 1990] M.J. Rovine and M. Delaney, Missing Data Estimation in Developmental Research, Statistical Methods in Longitudinal Research: Principles and Structuring Change, A. Von Eye ed., vol. 1, pp. 35-79, New York: Academic, 1990.

[Rubin, 1977] Rubin, D. "Formalizing subjective notions about the effect of nonrespondents in sample surveys," *J. Amer. Stat. Assoc.*, vol. 72, no. 359, pp. 538–543, Sep. 1977.

[Rubin, 1996] Rubin, D. "Multiple imputation after 18+ years," *J. Amer. Stat. Assoc.*, vol. 91, no. 434, pp. 473–489, Jun. 1996.

[Russell, 1918] Russell, R.C. 1918 Index, U.S. Patent 1,261,167, http://patft.uspto.gov/netahtml/srchnum.htm, Apr. 1918.

[Russell, 1922] Russell, R.C. 1922. Index, U.S. Patent 1,435,663, http://patft.uspto. gov/netahtml/srchnum.htm, Noviembre. 1922.

[SAS, 2003] SAS. Enterprise Miner SEMMA [En línea]. Cary, NC: SAS Institute Inc., 2003.
<http://www.sas.com/technologies/analytics/datamining/miner/semma.html>
[Consulta: Mayo 10 de 2009]

[Schwager y Margolin, 1982] Schwager, S. J., & Margolin, B. H. Detection of multivariate outliers. *The annals of statistics, Vol. 10*, pp. 943-954. 1982.

[Shafer, 1997] Shafer, J. Analysis of Incomplete Multivariate Data. London, U.K.: Chapman & Hall, 1997.

[Shafer y Graham, 2002] Schafer, J.L., y Graham, J. W. Missing Data: Our view of the state of the art. Psychological Methods. Vol 7 No. 2, pp. 147-177.

[Shekhar et. al., 2001] Shekhar, S., Lu, C., and Zhang, P.: 2001, 'Detecting Graph-Based Spatial Outliers: Algorithms and Applications'. In: *Proceedings of*

the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[Smith y Waterman, 1981] Smith, T.F. y Waterman, M.S. "Identification of Common Molecular Subsequences", Journal of Molecular Biology, vol. 147, no. 1, pp. 195-197, 1981.

[Song et. Al., 2005] Song, Q., Shepperd, M y Cartwright, M. A short note on safest default missingness mechanism assumptions, *Empirical Software Engineering: An International Journal* Vol. **10** No. 2, 2005, pp. 235–243.

[Strike *et. al.*, 2001] K. Strike, K.E. Emam, and N. Madhavji. Strike, K., El-Emam, K.E., Madhavji, N. (2001). "Software Cost Estimation with Incomplete Data", *IEEE Transaction on Software Engineering*, 27 (10), 890-908.

[Sutinen y Tarhio, 1995] Sutinen, E. y Tarhio, J. "On Using Q-Gram Locations in Approximate String Matching", Proceedings of the Third Annual European Symposium on Algorithms, pp. 327-340, 1995.

[Taft, 1970] Taft, R.L. 1970. Name Search Techniques. Technical Report Special Report No. 1, Nueva York State Identification and Intelligence System, Albany, N.Y., Febrero, 1970.

[Tan et. al., 2006] Tan, P., Steinbach, M. y Kumar, V. Introduction to data mining: Missing values. 2006, p40-41.

[Taylor y Cihon, 2004] Taylor, J. y Cihon, C. Statistical techniques for data analysis. 2.ed. New York: Chapman & Hall/CRC, 2004. p.103.

[Tierstein, 2005] Tierstein, Leslie. A Methodology for Data Cleansing and Conversion, White paper W R Systems, Ltd. 2005.

[Tiwary et. al, 2007] Tiwary, K., Metha, K., Jain, N., Tiwari, R. y Kanda, G. Selecting the Appropriate Outlier Treatment for Common Industry. SAS Conference Proceedings: NESUG 2007. Noviembre 11-14, 2007, Baltimore, Maryland.

[Triola et. al., 2004] Triola, M. y otros. Estadística. 9.ed. S.l: Pearson Addison Wesley, 2004. p.376

[Tukey, 1977] Exploratory Data Analysis. Addison-Wesley Publishing Co, London. 1977.

[UCLA, 2009] How can I see the number of missing values and patterns of missing values in my data file. UCLA: Academic Technology Services, Statistical consulting Group. [En línea] <http://www.ats.ucla.edu/stat/Stata/faq/nummiss_stata.htm [Fecha consulta: Noviembre 10 de 2009].

[Ukkonen, 1997] Ukkonen, E. "Aproximate String-Matching With Q-grams and Maximal Matches", Theoretical Computer Science, vol. 92, no. 1, pp. 191-211, 1992.

[Ullmann,1977] Ullmann, J.R. 1977. A Binary n-Gram Technique for Automatic Correction of Substitution, Deletion, Insertion, and Reversal Errors in Words. The Computer J., 20(2), 141-147.

[Useche y Mesa, 2006] Useche, L. y Mesa, D. Una Introducción a la imputación de valores perdidos. Terra Nueva Etapa. Vol XXII, No. 031. Universidad Central de Venezuela. Caracas, Venezuela. pp. 127-151.

[Van den Broeck, 2005] Van den Broeck Jan, Argeseanu Cunningham Solveig, Eeckels Roger, Herbst Kobus. Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLoS Medicine. 2005

[Verbeke y Molenberghs, 2000] Verbeke, G. y Molenberghs, G. Linear mixed models for longitudinal data. New York: Springer-Verlag, 2000.

[Verma y Quiroz, 2006] Vernal, S y Quiroz, A. Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering. En: Revista Mexicana de Ciencias Geológicas. Vol. 23, No. 2, 2006, p. 133.

[Vidal et.al., 1995] Vidal, E., Marzal, A. y Aibar, P. "Fast Computation of Normalized Edit Distances", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 9, pp. 899-902, 1995.

[Wagner y Fischer, 1974] Wagner, R.A. y Fischer, M.J. "The String-to-String Correction Problem", Journal of the ACM, vol. 21, no. 1, pp. 168-173, 1974.

[Wagstaff, 2004] Wagstaff, K. Clustering with missing values: No imputation required. Classification, Clustering, and Data Mining Applications. En Proceedings of the Meeting of the International Federation of Classification Societies. pp. 649–658. Springer. 2004

[Wang, 2003] Wang, J. Data mining opportunities and challenges. IGI Publishing, Hershey. PA, USA. 468p.

[Waterman et.al., 1976] Waterman, M., Smith, y T., Beyer, W.A. 1976. Some biological sequence metrics. Advances in Math., 20(4), 367-387, 1976.

[Weigel y Fein, 1994] Weigel, A. y Fein, F. "Normalizing the Weighted Edit Distance", Proceedings of the 12th IAPR International Conference on Pattern Recognition, pp. 399-402, 1994.

[Webster, 2000] Webster, A. Estadística Aplicada a los negocios y la economía. 3ª Edición. McGrawHill, 2000.

[Winkler, 1989] Winkler, W.E. "Frequency-Based Matching in the Fellegi-Sunter Model of Record Linkage", Proceedings of the Section on Survey Research Methods, pp. 778-783, 1989.

[Winkler, 1990] Winkler, W.E. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage", Proceedings of the Section on Survey Research Methods, pp. 354-359, 1990.

[Winkler, 1993] Winkler, W.E. "Improved Decision Rules in the Felligi-Sunter Model of Record Linkage", Proceedings of the Section on Survey Research Methods, pp. 274-279, 1993.

[Winkler, 2000] Winkler, W.E. "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage", Proceedings of the Section on Survey Research Methods, pp. 667-671, 2000.

[YALE, 1998] Universidad de Yale. Departamento de estadística. Linear Regression. 1998. [En línea].
Disponible en: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[Yancey, 2006] Yancey, W.E. "Evaluating String Comparator Performance for Record Linkage", Proceedings of the Fifth Australasian Conference on Data mining and Analytics, pp. 23-21, 2006.

[Yujian y Bo, 2007] Yujian, L. y Bo, L. "A Normalized Levenshtein Distance Metric", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1091-1095, junio 2007.

[Zhang et. al., 2008] Zhang, C., Zhou, X., Gao, C. y Wang, C. On Improving the Precision of Localization with Gross Error Removal. En: Proceedings 28th International Conference on Distributed Computing Systems Workshops. (2008: Shanghai). p.147

[Zimmerman, 1994] Zimmerman, D. W. A note on the influence of outliers on parametric and nonparametric tests. *Journal of General Psychology, Vol. 121 No.* 4, pp. 391-401. 1994.

[Zimmerman, 1995] Zimmerman, D. W. Increasing the power of nonparametric tests by detecting and downweighting outliers. *Journal of Experimental Education, Vol. 64 No.* 1, pp. 71-78. 1995.

[Zimmerman, 1998] Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education,* Vol. 67 No. 1, pp. 55-68.

ANEXOS

ANEXO 1

Artículo Hacia una metodología para la selección de técnicas de depuración de datos.

El siguiente artículo fue publicado en la revista Avances en Sistemas e Informática (categoría C), Volumen 6, Número 1, Junio 2009. ISSN 1657-7663.

Adicionalmente, se presentó como ponencia en el Cuarto Congreso Colombiano de Computación (4CCC), realizado en Bucaramanga (Abril 23-25, 2009) y quedó publicado en las memorias del evento ISBN 978-958-8166-43-8.

ANEXO 2

Artículo Funciones de Similitud sobre Cadenas de Texto: Una Comparación Basada en la Naturaleza de los Datos

El siguiente artículo se presentó como ponencia en el Congreso CONF – IRM2010 realizado en Jamaica en Mayo 16 a 18 del 2010.

ANEXO 3

Artículo Detección de Duplicados: Una Guía Metodológica

El siguiente artículo se presentó como ponencia en el Quinto Congreso Colombiano de Computación realizado en Cartagena en Abril 14 a 16 del 2010.

Adicionalmente fue seleccionado entre los mejores artículos del congreso y será publicado en la Revista Colombiana de Computación (categoría C) en la edición de Junio del 2010.