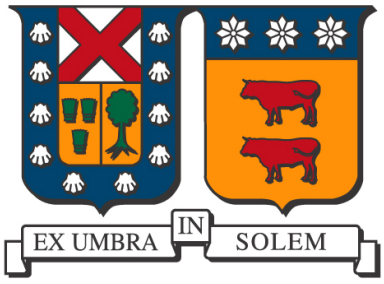


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE MATEMÁTICA

VALPARAÍSO - CHILE

Detección de fraudes en el consumo de agua potable

Memoria de Título presentada por

Francisco Alfaro Medina

como requisito parcial para optar al título de

Ingeniero Civil Matemático

Profesor Guía

Dr. Ronny Vallejos

Martes 22 de Agosto, 2017.

Índice general

Capítulo 1

Introducción

Desde la llegada de la modernidad, la dinámica de acceso a los servicios básicos se ha presentado de distintas formas, por ejemplo la provisión de agua para beber comenzó siendo responsabilidad individual de los usuarios, quienes acudían a fuentes diversas, posteriormente concurrían a pozos centralizados y regularizados, sin embargo, la llegada de la energía eléctrica domiciliaria cambió la forma de acceso, pues no era posible para un usuario individual generar su propia energía, teniendo que acudir a terceros que proveían el servicio, generándose un monopolio natural. El incremento de los requisitos básicos para el consumo de agua, junto con el aumento de la población en los núcleos urbanos, hizo necesario que se normalizara y centralizara el acceso, dando paso a las empresas de servicios sanitarios.

En Chile, la cobertura se ha incrementado de manera sostenida en el tiempo (ver Tabla C.1), llegando a un nivel de acceso del 99,9 % de la población (Superintendencia de Servicios Sanitarios, 2014), colocando a Chile por sobre el promedio latinoamericano, el cual es de un 93 % (Soulier Faure, M., Ducci, J., & Altamira, M., 2013), ubicándose a niveles comparables con países como Noruega o Dinamarca (OECD,2007).

En términos de los consumos y pérdidas, la Asociación Internacional de Agua (IWA) en conjunto con la Asociación Americana de Trabajos en Agua (AWWA), definen los tipos de consumo que puede seguir un cliente y los tipos de pérdidas que existen en los sistema de agua (American Water Works Association, 2012). Para comprender la relación entre estas terminologías, es necesario definir algunos conceptos previos, cuya relación se muestra en la Tabla C.2 (Ver Apéndice C).

- Suministro de Agua Potable: Corresponde al volumen anual de agua potable inyectada al sistema.

- Consumos autorizados: Corresponde al volumen anual de agua medida y/o no medida entregada a clientes autorizados.
- Pérdidas de Agua: Diferencia entre el agua inyectada al sistema y los consumos autorizados.
- Pérdidas aparentes: Consumos no autorizados, todo tipo de imprecisiones de medición, y errores sistemáticos de manejo de datos.
- Pérdidas Físicas: El volumen anual de pérdidas mediante todo tipo de filtraciones, fugas, roturas o rebalse en redes, almacenamiento o puntos de servicio, hasta el punto de medición del cliente.

De acuerdo a los datos contenidos en el Informe de Gestión del Sector Sanitario 2014, presentado por la Superintendencia de Servicios Sanitarios (SISS), en Chile las empresas sanitarias produjeron en su conjunto un total de 1.670 millones de metros cúbicos de agua potable de la cual un 33,65 % no fue facturada, de ese total estimaciones de la SISS establecen que el 74 % se originan por pérdidas físicas, mientras que el 26 % restante corresponde a pérdidas aparentes, esto equivale a cerca de 146 millones de metros cúbicos de agua potable. Si se estima que el metro cúbico de agua cuesta alrededor de $700 pesos chilenos (Superintendencia de Servicios Sanitarios, 2016), en total se estaría perdiendo más 100.000 millones de pesos por año, cifras alarmantes para las empresas sanitarias.

El objetivo principal de este trabajo se enfoca en desarrollar un modelo matemático capaz de identificar y predecir aquellos clientes que tienen un consumo anómalo. El tipo de problema abordado corresponde a un problema clásico de "Modelos de Detección de Fraude". Este tipo de problemas ha sido ampliamente estudiado en las matemáticas financieras, dado que es aquí donde las empresas tienen las mayores pérdidas económicas.

El fenómeno a estudiar tiene una dificultad adicional, la clasificación de los clientes se encuentra sesgada, es decir, existen clientes que presentan consumos anómalos que de momento no han sido detectados como tal, encontrándose registrados como clientes de consumo regular. Por tanto, se debe proponer una nueva metodología de investigación para dar solución a la problemática, tomando un enfoque diferente al usual abordado en este tipo de problemas.

La estructura de trabajo es la siguiente: En el Capítulo 2 se presenta una serie de definiciones y resultados que permiten definir de manera formal los modelos en este trabajo (Murphy, 2012). En el mismo capítulo se habla de enfoques alternativos encontrados en la literatura, desde cómo ha sido abordado el mismo problema en otras partes del mundo: Caso Kampala, Uganda

(Humaid & Barhoom, 2012) y Medidores inteligentes, hasta problemas similares abordados en este ámbito: Detección de crímenes y mapas de calor (PredPol), y aplicaciones de la ley de Benford para la detección de fraudes en contabilidad.

En el Capítulo 3 se presenta la metodología de investigación, divido en tres etapas: La primera etapa corresponde al análisis del conjunto de datos, seleccionando la información relevante para comprender el consumo de los clientes de la empresa ESVAL mediante gráficos y tablas. La segunda etapa corresponde a la modelación, donde se detallan los pasos para dar solución al problema. Para comprender el comportamiento del consumo de los clientes se ajustaran modelos de clasificación sobre el conjunto de datos, tales como: regresión logística, bosque aleatorio y máquinas de soporte vectorial. Los resultados de los modelos serán comparados por diferentes medidas de rendimiento, tales como: accuaracy, precision, recall y f-score. Por otro lado, se realiza pruebas de aleatoriedad espacial (basados en cuadrantes y en distancias) sobre las coordenadas espaciales de los clientes que siguen un comportamiento fraudulento, para concluir si el comportamiento de estos clientes sigue algún patrón o su comportamiento es completamente aleatorio.

Finalmente, en el Capítulo 4 se concluye el presente trabajo con algunos comentarios de los resultados obtenidos y se presenta una propuesta para continuar la investigación a modo de trabajos futuros.

Capítulo 2

Preeliminares

En este capítulo se introduce algunas definiciones y resultados básicos sobre los modelos ocupados en este trabajo. También, se realiza una investigación de cómo ha sido abordado este problema en diferentes partes del mundo y de cómo problemas similares a este, podrían dar una visión más amplia en la resolución del mismo.

2.1. Modelos de regresión

Para definir correctamente los modelos de regresión, es conveniente establecer la siguiente notación:

- $\boldsymbol{x}^{(i)} = (x_1^{(i)}, ..., x_k^{(i)})^\top$: Vector de entrada.
- $y^{(i)}$: Variable de salida.
- $\{(\boldsymbol{x}^{(i)}, y^{(i)});\ i = 1, ..., m\}$: Conjunto de entrenamiento.
- $\mathcal{X}$,$\mathcal{Y}$: Espacio de entrada y de salida, respectivamente.

Dado un conjunto de entrenamiento $\{(\boldsymbol{x}^{(i)}, y^{(i)});\ i = 1, ..., m\}$, un modelo de regresión consiste en estimar el valor de $\boldsymbol{y} = (y^{(1)}, ..., y^{(m)})^\top \in \mathcal{Y}$ mediante los valores de $\boldsymbol{X} = (\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(m)})^\top \in \mathcal{X}$.

Existen dos enfoque para resolver estos problemas: paramétrico y no-paramétrico. El **enfoque paramétrico** asume que los datos del conjunto de entrenamiento vienen de una población que sigue una distribución probabilística, basados en un sistema finitos de parámetros. Mientras que para el **enfoque**

no paramétrico, la estructura del modelo no se especifica a priori, sino que se determina a partir del conjunto de entrenamiento.

Dentro de los modelos de regresión, se encuentran los modelos de clasificación, en el cual la variable de salida $y^{(i)}$ solo puede tomar un número pequeño de valores discretos, es decir, $y^{(i)} \in \{0, 1, 2, ..., K\}$ para todo $i = 1, 2, ..., m$. Si $K = 2$, se tiene un problema de **clasificación binaria**, en donde la variable de salida puede tomar los valores 0 ó 1. Si $K > 2$, se tiene un problema de **clasificación multiclase**.

Este trabajo esta enfocado principalmente a los modelos de clasificación binaria. Si bien existe una variedad de este tipo de modelos (para más detalles (Murphy, 2012)), se da énfasis a tres modelos en particular: regresión logística, máquinas de soporte vectorial y bosque aleatorio.

2.1.1. Regresión logística

Corresponde al más simple de los modelos de clasificación binaria. El objetivo de este problema es realizar un símil con un problema de regresión lineal. Para definir adecuadamente el modelo, se considera los siguientes previos:

i) Se define la **función logística** o **función sigmoide** por:

$$g(z) = 1/(1 + e^{-z}).$$

Esta función cumple con las siguientes propiedades:

a) $g(z)$ tiende a 1 cuando $z \to \infty$ y $g(z)$ tiende a 0 cuando $z \to -\infty$.

b) $g'(z) = g(z)(1 - g(z))$.

ii) Sea $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_k) \in \mathbb{R}^{k+1}$ un vector de parámetros. En base a la función sigmoide, se define la función:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) = g(\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)}) = \frac{1}{1 + e^{\boldsymbol{\theta}^\top \boldsymbol{x}^{(i)}}} \quad , \quad \text{con} \quad \boldsymbol{\theta}^\top \boldsymbol{x}^{(i)} = \theta_0 + \sum_{j=1}^{k} \theta_j x_j^{(i)}$$

Luego, el modelo de regresión logística se define considerando algunos supuestos distribucionales sobre el conjunto de entrenamiento $\{(\boldsymbol{x}^{(i)}, y^{(i)});\ i = 1, ..., m\}$, estos son

$$p(y^{(i)} = 1 | \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})$$
$$p(y^{(i)} = 0 | \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) = 1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})$$

Es decir, $y^{(i)} \sim \text{Bernoulli}(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))$. Lo anterior, se escribe de forma compacta mediante

$$p(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\theta}) = (h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))^{y^{(i)}}(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))^{1-y^{(i)}} \tag{2.1}$$

Suponiendo que las m variables son generadas independientemente, la función de verosimilitud queda expresada por

$$\begin{aligned} L(\boldsymbol{\theta}) &= p(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta}) \\ &= \prod_{i=1}^{m} p(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\theta}) \\ &= \prod_{i=1}^{m} (h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})^{y^{(i)}}(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))^{1-y^{(i)}} \end{aligned} \tag{2.2}$$

Luego, la función de log-verosimilitud es

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \log\, L(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} y^{(i)} \log h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})) \end{aligned} \tag{2.3}$$

Derivando la ecuación (2.3) respecto a la j-ésima componente de $\boldsymbol{\theta}$, esto queda

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}_j}\ell(\boldsymbol{\theta}) &= (\boldsymbol{Y}\frac{1}{g(\boldsymbol{\theta}^\top \boldsymbol{X})} - (\mathbb{1} - \boldsymbol{Y})\frac{1}{g(\boldsymbol{\theta}^\top \boldsymbol{X})})\frac{\partial}{\partial \theta_j} g(\boldsymbol{\theta}^\top \boldsymbol{X}) \\ &= (\boldsymbol{Y}\frac{1}{g(\boldsymbol{\theta}^\top \boldsymbol{X})} - (\mathbb{1} - \boldsymbol{Y})\frac{1}{g(\boldsymbol{\theta}^\top \boldsymbol{X})}) g(\boldsymbol{\theta}^\top \boldsymbol{X})(\mathbb{1} - g(\boldsymbol{\theta}^\top \boldsymbol{X}))\frac{\partial}{\partial \theta_j}\boldsymbol{\theta}^\top \boldsymbol{X} \\ &= (\boldsymbol{Y}(\mathbb{1} - g(\boldsymbol{\theta}^\top \boldsymbol{X})) - (\mathbb{1} - \boldsymbol{Y}) g(\boldsymbol{\theta}^\top \boldsymbol{X})\boldsymbol{X}_{(j)} \\ &= (\boldsymbol{Y} - h_{\boldsymbol{\theta}}(\boldsymbol{X}))\boldsymbol{X}_{(j)} \end{aligned} \tag{2.4}$$

Existen varios métodos para estimar el valor de $\boldsymbol{\theta}$, los más usados se presentan a continuación: método de puntuación de Fisher, gradiente descendente, gradiente descendente estocástico, Newton-Rampson, entre otros métodos (Solomon, 2015).

2.1.2. Máquinas de soporte vectorial

El modelo de máquinas de soporte vectorial (abreviado con las siglas SVM), es un modelo de clasificación multiclase, sin embargo, en esta sección se

desarrolla la teoría para el problema de clasificación binaria. Para definir adecuadamente el modelo, se consideran los siguientes previos:

i) En el conjunto de entrenamiento $\{(\boldsymbol{x}^{(i)}, y^{(i)});\ i = 1, ..., m\}$, se considera que $y^{(i)} \in \{-1, 1\}$, para todo $i = 1, ..., m$.

ii) Un conjunto de entrenamiento $\{(\boldsymbol{x}^{(i)}, y^{(i)});\ i = 1, ..., m\}$ se dice que es **separable** si los conjuntos $\{\boldsymbol{x}^{(i)}, y^{(i)} = 1\}$ y $\{\boldsymbol{x}^{(i)}, y^{(i)} = -1\}$ pueden ser separados por un hiperplano:

$$H = \{\boldsymbol{x}\ ;\ R(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b = \sum_{j=1}^{k} w_j x_j + b\},$$

Es decir, $y^{(i)} R(\boldsymbol{x}^{(i)}) > 0$, para todo $i = 1, ..., m$.

Existen dos escenarios respecto al conjunto de entrenamiento: el caso separable y el caso no separable.

Caso Separable

Cuando el conjunto de entrenamiento es separable, el objetivo es encontrar un hiperplano $H(\boldsymbol{x})$ que separe correctamente el conjunto de entrenamiento, que al mismo tiempo maximice la distancia perpendicular entre dicho hiperplano con el punto más cercano al mismo.

Para comprender mejor el planteamiento del problema, se analiza un conjunto de entrenamiento en el espacio euclidiano (ver Figura 2.1). En este caso, el hiperplano $H(x)$ correspondería a una recta que separa el conjunto de entrenamiento en dos regiones: $\mathcal{R}_1$ y $\mathcal{R}_2$. Se dice que el punto $x \in \mathcal{R}_1$ si $R(x) > 0$, de lo contrario el punto pertenece a la región $\mathcal{R}_2$. Por otro lado, sea $\boldsymbol{x}$ un punto en el espacio, r la distancia de $\boldsymbol{x}$ al hiperplano $H(\boldsymbol{x})$ (el valor de r se denomina **margen**), $\boldsymbol{w}$ el vector normal de r y $\boldsymbol{x}_\perp$ la proyección ortogonal de $\boldsymbol{x}$. Entonces, $\boldsymbol{x}$ se puede expresar como

$$\boldsymbol{x} = \boldsymbol{x}_\perp + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

Por tanto, el hiperplano luce como en la Figura 2.1, quedando definido matemáticamente por la expresión

$$H(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b = (\boldsymbol{w}^\top \boldsymbol{x}_\perp + b) + r\frac{\boldsymbol{w}^\top \boldsymbol{w}}{\|\boldsymbol{w}\|}$$

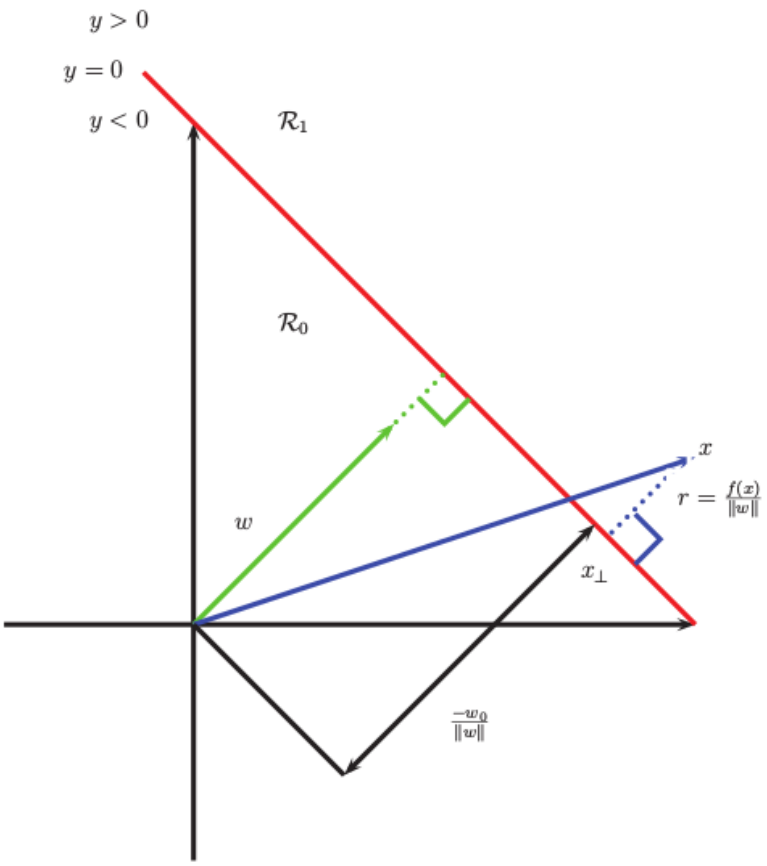


Figura 2.1: Esquema del modelo SVM en el plano Euclidiano.

Notar que: $H(\boldsymbol{x}_\perp) = 0$, así $\boldsymbol{w}^\top \boldsymbol{x}_\perp + b = 0$. Entonces, el hiperplano se puede escribir como: $H(\boldsymbol{x}) = r\frac{\boldsymbol{w}^\top \boldsymbol{w}}{\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}$ y $r = \frac{H(\boldsymbol{x})}{\|\boldsymbol{w}\|}$.
En este contexto, se estaría buscando que la distancia $r = H(\boldsymbol{x})/\|\boldsymbol{w}\|$ sea lo más grande posible, como se muestra en la Figura 2.2. En particular, podrían haber muchas líneas que separen perfectamente los datos de entrenamiento (especialmente si se trabaja en un espacio de alta dimensión), pero intuitivamente, el mejor a escoger es aquel que maximiza el margen, es decir, la distancia perpendicular al punto más cercano. Además, se busca que asegurar que cada punto se encuentre en la región correcta del hiperplano, es decir, se quiere que $H(x_i)y_i > 0$.

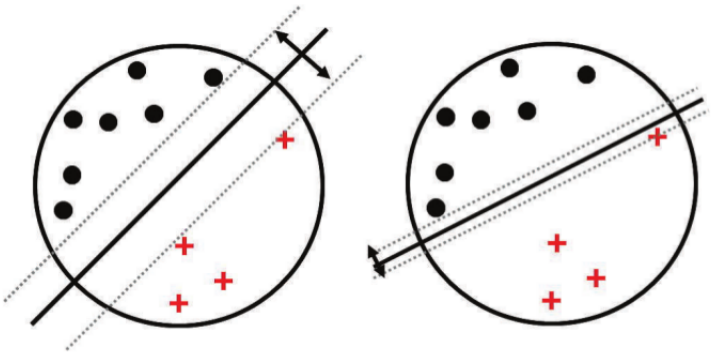


Figura 2.2: Ilustración del largo de los margenes principales. Izquierda: un hiperplano separador con un margen grande. Derecha: un hiperplano separador con un margen pequeño.

Por lo tanto, la función objetivo de este problema queda formulada matemáticamente por

$$(\mathrm{P}) : \max_{\boldsymbol{w},b}\Big(\min_{i=1,\dots,m} \frac{y^{(i)}(\boldsymbol{w}^\top \boldsymbol{x}^i + b)}{\|\boldsymbol{w}\|}\Big) \tag{2.5}$$

Re-escalando los parámetros por: $\boldsymbol{w} \to k\boldsymbol{w}$ y $b \to kb$, lo anterior no cambia la distancia de todos los puntos al hiperplano en cuestión, esto se debe a que el factor k se cancela cuando se divide por $\|\boldsymbol{w}\|$. Luego, se define el valor de escala $y^{(i)}H(\boldsymbol{x}^{(i)}) = 1$ (con $i = 1, ..., m$) para los puntos más cercanos al hiperplano.
Por lo tanto, la ecuación (2.5) se puede reescribir por

$$\text{(P):} \begin{cases} \min\limits_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{s.a.} : y^{(i)}(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 \ , \ i = 1, ..., n \end{cases} \tag{2.6}$$

La restricción indica que todos los puntos que se encuentren en el lado correcto del hiperplano, tengan al menos un margen de 1. Por esta razón, se dice que un SVM es un ejemplo de **clasificador de grandes margenes**.

Caso No Separable

En primera instancia, la idea es llevar el problema del caso no separable a un problema del caso separable. Para ello, se busca una función $\phi : \mathbb{R}^k \to \mathbb{R}^K$ (usualmente $K \gg k$), tal que los conjuntos $\{\boldsymbol{z}^{(i)} = \phi(\boldsymbol{x}^{(i)}), y^{(i)} = 1\}$ y $\{\boldsymbol{z}^{(i)} = \phi(\boldsymbol{x}^{(i)}), y^{(i)} = -1\}$ sean linealmente separables.

Por otro lado, si no existen restricciones respecto a la asignación de ϕ, esto puede conducir a que el problema de hallar un clasificador lineal en $\mathbb{R}^K$, sea computacionalmente intratable.

La solución, es escoger funciones particulares de ϕ, denominadas funciones de **kernelización**. Para esto se escoge $\phi : \mathbb{R}^k \to \mathbb{R}^K$ tal que:

$$\langle \phi(x), \phi(y) \rangle = \kappa(x, y)$$

para algún kernel $\kappa : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$. Aquí, $\langle \cdot, \cdot \rangle$ denota el producto escalar en $\mathbb{R}^K$ (Para más detalles, consultar el Apéndice C) .

Si después de aplicar las funciones de kernelización sobre el conjunto de entrenamiento, este sigo siendo un conjunto no separable, se re-formula la ecuación (2.6), añadiendo un vector de holgura $\boldsymbol{\xi} = (\xi_1, ..., \xi_m)$, con $\xi_i \geq 0$, para todo $i = 1, ..., m$.

Se tiene que: $\xi_i = 0$, si el punto se encuentra en o dentro del borde marginal correcto, en caso contrario, $\xi_i = |y^{(i)} - H(\boldsymbol{x}^{(i)})|$. Si $0 < \xi_i \leq 1$, el punto se encuentra dentro del margen pero no en el lado correcto del hiperplano. Si $\xi_i > 1$, el punto se encuentra en el lado equivocado del hiperplano (ver Figura 2.3).

Ahora, se reemplaza la condición $y^{(i)}H(\boldsymbol{x}^{(i)} \geq 1$ por $y^{(i)}H(\boldsymbol{x}^{(i)}) \geq 1 - \xi_i$ (esta condición se conoce con el nombre de **condición de margen suave**). Por lo tanto, la nueva función objetivo es

$$\text{(P):} \begin{cases} \min\limits_{\boldsymbol{\xi},\boldsymbol{w},b} \ \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{n}\xi_i \\ \text{s.a. :} \quad \xi_{(i)} \geq 0 \ , \ i = 1, ..., n \\ \quad y^{(i)}(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_{(i)} \ , \ i = 1, ..., n \end{cases} \tag{2.7}$$

donde el parámetro C corresponde a un parámetro de regularización, cuya función es controlar el número de errores que se está dispuesto a tolerar en el conjunto de entrenamiento. Comúnmente se define $C = 1/\nu n$, donde $0 \leq \nu \leq 1$ controla la fracción de los puntos mal clasificados permitidos.

La ecuación (2.7) corresponde a un problema con restricciones que puede ser resuelto usando los multiplicadores de Lagrange. La función objetivo adopta la forma de un problema tipo **QP**("Quadratic Program"), cuyo tiempo de ejecución es del orden de $\mathcal{O}(n^3)$. Existen algoritmos especializados que evitan soluciones genéricas QP, por ejemplo, el algoritmo SMO ("Sequential minimal optimization") (Platt, 1998), que en la práctica tiene un orden de $\mathcal{O}(n^2)$. No obstante, esto puede ser lento si n es demasiado grande, por lo que se frecuenta ocupar los SVM lineales, cuyo tiempo ejecuación es de $\mathcal{O}(n)$ (Joachims (2006) ; Bottouet al. (2007)).

Se puede probar que la solución de la ecuación (2.7) es:

$$\hat{\boldsymbol{w}} = \sum_i \alpha_i \boldsymbol{x}_i, \tag{2.8}$$

donde $\alpha_i = \lambda_i y_i$ y $\alpha = (\alpha_1, ..., \alpha_m)$ es vector con una gran cantidad de ceros. Los x_i para los cuales $\alpha_i > 0$ son llamados vectores de soporte. Estos puntos son clasificados incorrectamente o son clasificados correctamente pero se encuentran en o dentro del margen.

Por tanto, la predicción viene dada por:

$$\hat{y}(\boldsymbol{x}) = \text{sgn}(\hat{\boldsymbol{w}}^\top \boldsymbol{x} + \hat{b}). \tag{2.9}$$

2.1.3. Bosque aleatorio

Un árbol de decisión, se define como una partición recursiva del espacio de entrada R, definiendo con ello un modelo local en cada partición realizada.

Sea R_m es la m-ésima partición de la región, w_m es la respuesta media en aquella región y $\boldsymbol{v}_m$ codifica la elección de la variable a partir de la cual esta se dividirá, y el valor del umbral, es la trayectoria de la raíz a la m-ésima hoja. El modelo se escribe como:

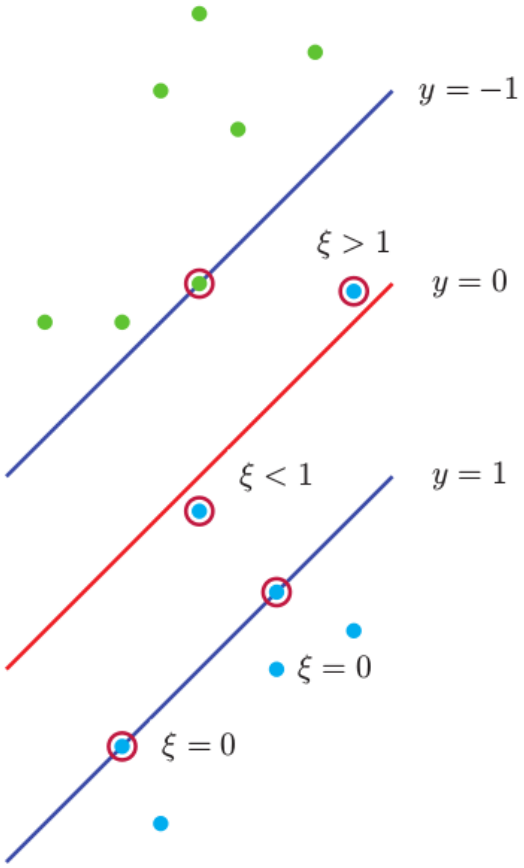


Figura 2.3: Ilustración del margen principal suave. Los puntos con círculos a su alrededor son vectores de soporte. También se indica el valor de las correspondientes variables de holgura. Basado en la Figura 7.3 de (Obispo 2006a).

$$f(x) = \mathbb{E}(y \mid \boldsymbol{x}) = \sum_{m=1}^{M} w_m \mathbb{1}_{\{\boldsymbol{x} \in R_m\}} = \sum_{m=1}^{M} w_m \phi(\boldsymbol{x} \ ; \ \boldsymbol{v}). \tag{2.10}$$

Se observa que un árbol de decisión queda definido por funciones de bases adaptativas. Por otro lado, estas estimaciones tienen un número considerables de variables, buscando alternativas para reducirlas. Un camino para solucionar este problema es considerar el promedio conjunto de muchas estimaciones. Por ejemplo, se puede entrenar M árboles diferentes (f_m) definidos por la ecuación (2.10) sobre subconjuntos diferentes de datos elegidos al azar con reemplazo, y luego calcular la función

$$f(x) = \sum_{m=1}^{M} \frac{1}{M} f_m(x) \tag{2.11}$$

Esta técnica es llamada "Bagging" (Breiman, 1996), lo que significa "bootstrap aggregating".

Las desventajas que posee este método es que al volver a ejecutar el mismo algoritmo de aprendizaje en diferentes subconjuntos de datos, puede resultar en predictores altamente correlacionados, lo que limita la reducción de la varianza.

Esta técnica es conocida como Bosque Aleatorio (Breiman, 2001). Este tipo de modelos suelen tener una precisión predictiva muy buena (Caruana & Niculescu-Mizil , 2001) y se han usado ampliamente en muchas aplicaciones, por ejemplo, para el reconocimiento de la pose del cuerpo usando el sensor del Kinect de Microsoft (Shotton et al., 2011).

También es posible desarrollar un enfoque bayesiano del aprendizaje de árbol. En particular (Wu et al., 2007), realizaron inferencia aproximada sobre el espacio de los árboles (tanto en la estructura como los parámetros) utilizando técnicas del tipo Markov Chain Monte Carlo (MCMC). Esta metodología redujo la varianza.

2.2. Medidas de rendimiento de los modelos

La **matriz de confusión** es una herramienta que permite la visualización del desempeño de los modelos de clasificación. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real.

Para el caso de los problemas de clasificación binaria, la matriz de confusión corresponde a una matriz cuadrada de 2×2. Las entradas de esta matriz se explican considerando el siguiente experimento: se tienen P instancias positivas y N instancias negativas. Por lo tanto, la matriz de confusión se define según la Figura 2.4, en donde:

- **Verdadero Positivo**(TP): Número de veces donde hubo una instancia positiva P, dado que hubo una instancia positiva P.
- **Falso Positivo**(FP): Número de veces donde hubo una instancia positiva P, dado que hubo una instancia negativa N.
- **Verdadero Negativo**(TN): Número de veces donde hubo una instancia una instancia negativa N, dado que hubo una instancia negativa N.
- **Falso Negativo**(FN): Número de veces donde hubo una instancia una negativa N, dado que hubo una instancia positiva P.

		Clase real	
		Clase referencia	***Clase no referencia***
Clase estimada	Clase referencia	TP	FP
	Clase no referencia	FN	TN

Figura 2.4: Matriz de confusión para la clasificación binaria.

A partir de los elementos de la matriz de confusión, se definen las siguientes medidas de rendimientos:

a) **accuracy**: En el campo de la recuperación de información, la medida "accuracy" corresponde al total de documentos recuperados en la consulta. Esta se define por:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.12}$$

b) **precision**: En el campo de la recuperación de información, la medida "presicion" corresponde a la fracción de documentos recuperados que son relevantes para la consulta. Esta se define por:

$$\text{precision} = \frac{TP}{TP + FP} \tag{2.13}$$

c) **recall**: En el campo de la recuperación de información, la medida "recall" corresponde a la fracción de los documentos relevantes que se recuperan con éxito. Esta se define por:

$$\text{recall} = \frac{TP}{TP + TN} \tag{2.14}$$

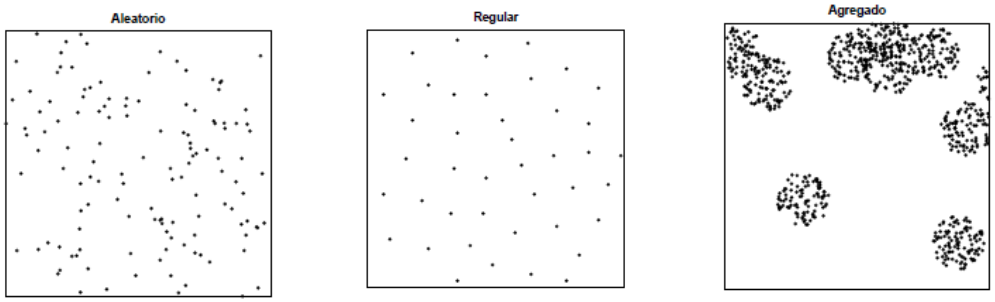
d) **f-score**: Corresponde a la media armónica entre las medidas "presicion" y "recall". Esta se define por:

$$\text{f-score} = 2\ \frac{\text{accuracy} \cdot \text{precision}}{\text{accuracy} + \text{precision}} \tag{2.15}$$

2.3. Análisis de patrones puntuales

Sea $\{Z(s) : s \in D \subset \mathbb{R}^d\}$ un proceso estocástico, en el que s es la ubicación en el espacio Euclidiano d-dimensional. Un **proceso puntual** se define como un arreglo o patrón de puntos en un conjunto aleatorio D. Estos puntos se denominan los eventos del proceso. Si los eventos son observados sólo parcialmente, el patrón se llama un patrón muestreado ("Sampled point mapped"). Cuando todos los eventos de la realización se registran, se dice que este es puntual.

Sea D un conjunto aleatorio, el experimento que genera una realización particular puede ser visto como un sorteo de lugares en D de los eventos que son observados. Los patrones puntuales son realizaciones de experimentos aleatorios y se distingue entre patrones (completamente) aleatorios, agrupados (especialmente agregados), y los regulares, lo cual no debe conducir a la falsa impresión de que los dos últimos tipos de patrones no contienen ningún tipo de aleatoriedad. Ejemplo de estos tres tipos de patrones aleatorios se observan en la Figura 2.5.



(a) Patrones puntuales con distribución aleatoria.

(b) Patrones puntuales regulares.

(c) Patrones puntuales agregados.

Figura 2.5: Tipo de patrones puntuales.

Un patrón puntual se llama completamente aleatorio si se cumplen los siguientes requisitos:

a) El promedio de eventos por unidad de área, la intensidad $\lambda(s)$, es homogénea a lo largo D.

b) El número de eventos en dos sub-regiones que no se solapan, A_1 y A_2 son independientes.

c) El número de eventos en cualquier sub-región sigue una distribución de Poisson.

Por lo tanto, los eventos se distribuyen uniforme e independiente a lo largo del dominio, proceso que se reconoce matemáticamente como un proceso Poisson homogéneo, y que sirve como hipótesis nula para muchas investigaciones estadísticas en los patrones puntuales.

Test de aleatoriedad

Método basado en cuadrantes

Otro tipo de métodos para probar la aleatoriedad espacial de un patrón son basados en la división del dominio D en sub-regiones no traslapadas (cuadrantes) $A_1, ..., A_k$ de igual tamaño. Sean $A_1, ..., A_k$ tal que $\cup_{i=1}^{k} A_i = D$, donde A_i y A_j no se traslapan.

Test de Bondad de ajuste Chi-Cuadrado

Esta prueba de ajuste estadístico, prueba la hipótesis nula, en que los n puntos estarían distribuidos uniformemente independientes a lo largo de D, o en otras palabras, los conteos de sitios por cuadrante son variables Poisson independientes con media común.
El estadístico de prueba esta dado por:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(rc-1)} \tag{2.16}$$

Ahora, si $O_{ij} = n_{ij}$, $E_{ij} = \overline{n} = n/rs$, con r y s definidos, $\hat{\lambda} = n/|A|$, $N(A) \sim$ Poisson(λA). La ecuación (2.16) se puede escribir de la siguiente forma

$$X^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(n_{ij} - \overline{n})^2}{n_{ij}} \sim \chi^2_{(rc-1)} \tag{2.17}$$

Índice de Dispersión

Si $X \sim$ Poisson(λ),$\mathbb{E}(X) = \mathbb{V}(X) = \lambda$. Entonces, el índice de dispersión se define por:

$$I = \frac{\mathbb{V}(N(A))}{\mathbb{E}(N(A))} = \frac{S^2}{\overline{n}}, \tag{2.18}$$

donde $S^2 = \sum\sum(n_{ij} - \overline{n})^2/rc - 1$. Considerando que $(I-1)$ indica el tamaño del cluster, se definen las siguientes reglas de decisión:

I) Si $(I-1) \approx 0$, el patrón es completamente aleatorio.

II) Si $(I-1) > 0$, el patrón es agregado.

III) Si $(I-1) < 0$, el patrón es regular.

Métodos basados en distancias

La elección de la forma y el número de cuadrantes para probar la hipótesis de aleatoriedad basados en conteos por áreas es un elemento que puede influenciar los resultados. Las pruebas estadísticas basadas en distancias entre eventos o entre puntos muestreados y eventos elimina estas características. En esta sección se nombran las pruebas que tienen en cuenta tales distancias entre eventos.

Función G(h)

Esta función tiene en cuenta la mínima distancia entre eventos(distancia al vecino mas cercano). Se define:

- d_i = Distancia mínima entre un evento y sus vecinos.
- d = Variable aleatoria de mínima distancia de un punto a un evento.
- $G(d)$ = Función de distribución de d.

La función de distribución empírica de $G(d)$ viene dada por:

$$\hat{G}(d) = \frac{\sum \mathbb{1}_{\{d_i \leq d\}}}{n} \tag{2.19}$$

Por otro lado, la distribución teórica de $G(d)$ viene dada por

$$\begin{aligned} G(d) &= \mathbb{P}((D \leq d)) \\ &= 1 - \mathbb{P}((D > d)) \\ &= 1 - \mathbb{P}(\text{No hay eventos en } A = \pi d^2). \end{aligned} \tag{2.20}$$

Si $N(A) \sim \text{Poisson}(\lambda\pi d^2)$, entonces $P(N(A) = 0) = e^{-\lambda\pi d^2}$. Por lo tanto:

$$G(d) = 1 - e^{-\lambda\pi d^2} \tag{2.21}$$

de donde

- Si $\hat{G}(d)$ crece rápidamente en distancias cortas, los eventos son agregados.
- Si los eventos son regularmente espaciados, $\hat{G}(d)$ crece lentamente hasta cierta distancia (espacio eventos) y después crece rápidamente.

Adicionalmente, se hace un cálculo de las bandas de confianza para la función en donde se calcula:

1) $\hat{G}_0(h)$ = Función $G(H)$ calculada con los datos observados.
2) $\hat{G}_1(h), ..., \hat{G}_g(h)$ = Función $G(h)$ calculada con g realizaciones de un proceso completamente aleatorio.
3) $G_L(h) = \min\limits_{i=1,...,g} \hat{G}_i(h)$ y $G_U(h) = \max\limits_{i=1,...,g} \hat{G}_i(h)$

Función F(h)

Tiene en cuenta la mínima distancia punto-evento, donde un punto es un evento escogido aleatoriamente dentro de la región de estudio. Para su construcción se siguen los siguientes pasos:

1) Seleccionar aleatoriamente m puntos $\{p_1, ..., p_m\}$.
2) Calcular $d_i = d(p_i, s_i)$, la distancia de cada punto escogido al sitio del evento más cercano.

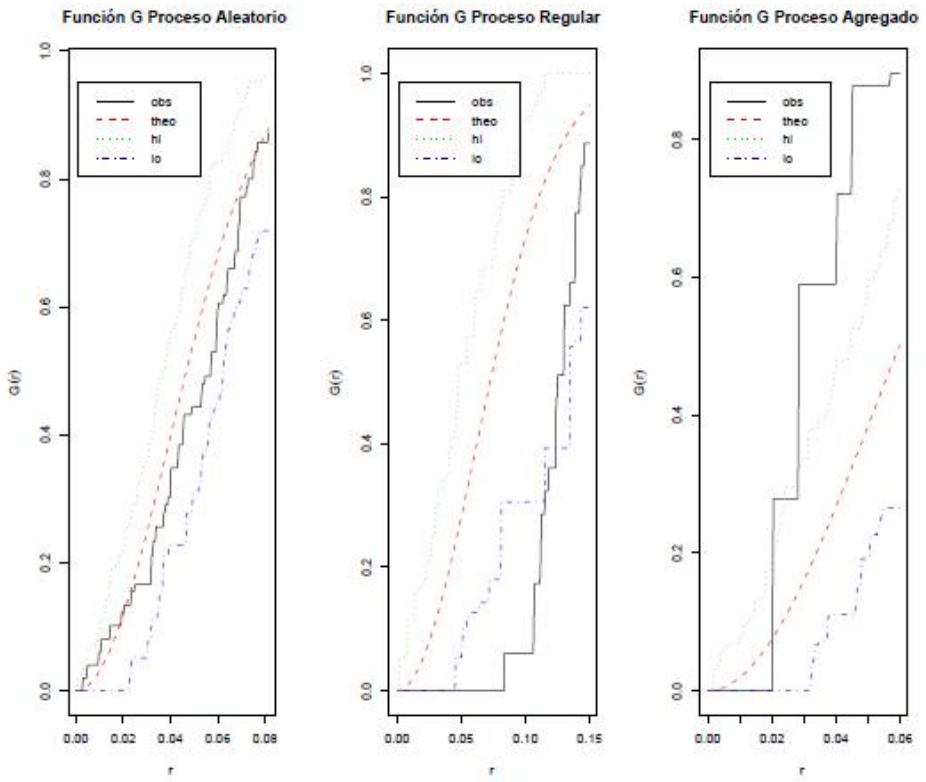
Sea d la variable aleatoria de mínima distancia de un punto a un evento, entonces:

$$F(d) = \mathbb{P}(D \leq d) \quad , \quad \hat{F}(d) = \frac{\sum \mathbb{1}_{\{d_i \leq d\}}}{m} \tag{2.22}$$

Luego, la función teórica es

$$F(d) = 1 - e^{-\lambda\pi d^2}. \tag{2.23}$$

de donde



(a) Patrones puntuales con distribución aleatoria.

(b) Patrones puntuales regulares.

(c) Patrones puntuales agregados.

Figura 2.6: Gráficas del la función G para cada tipo de proceso.

- Si $\hat{F}(d)$ crece lentamente al comienzo y rápidamente para distancia largas, el patrón es agregado.
- Si $\hat{F}(d)$ crece rápido al comienzo, el patrón es regular.

Por otro lado, sea x la distancia mínima de un punto a un evento. Para un patrón completamente aleatorio se tiene que:

$$F(x) = 1 - e^{-\lambda \pi x^2}, \tag{2.24}$$

donde

$$\mathbb{E}(x) = \frac{1}{2\sqrt{\lambda}} \quad , \quad \mathbb{V}(x) = \frac{4-\pi}{4\pi\lambda}. \tag{2.25}$$

Ahora, si se tiene $x_1, ..., x_m$:

$$\mathbb{E}(\overline{x}) = \frac{1}{2\sqrt{\lambda}} \quad , \quad \mathbb{V}(\overline{x}) = \frac{4-\pi}{4\pi\lambda m}. \tag{2.26}$$

Luego, si se define $G(y) = 1 - e^{-\lambda\pi y^2}$, con y = distancia mínima al evento más cercano. Dado $y_1, ..., y_m$, se observa que

$$\mathbb{E}(\overline{y}) = \frac{1}{2\sqrt{\lambda}} \quad , \quad \mathbb{V}(\overline{y}) = \frac{4-\pi}{4\pi\lambda m}. \tag{2.27}$$

De la ecuación (2.26) y la ecuación (2.27), se concluye que bajo aleatoriedad $G(y)$ y $F(y)$ son iguales.

Función K(h) de Ripley

La función K de Ripley (1976) es una función de λ_2 para procesos estacionarios e isotrópicos. También es conocida como la medida reducida del segundo momento (Cressie, 1998), y es la función reducida del segundo momento.

Sea $\mathcal{A}$ = Número de eventos adicionales a una distancia d de un evento elegido aleatoriamente. La función K de Ripley se define por

$$K(d) = \frac{\mathbb{E}(\mathcal{A})}{\lambda}. \tag{2.28}$$

Sea $\hat{\lambda} = n/A$, con A = Área de la región D, d_{ij} = distancia entre el i-ésimo y j-ésimo evento y $I_d(d_{ij}) = \begin{cases} 1 \text{ si } d_{ij} < 0 \\ 0 \text{ en otro caso.} \end{cases}$.

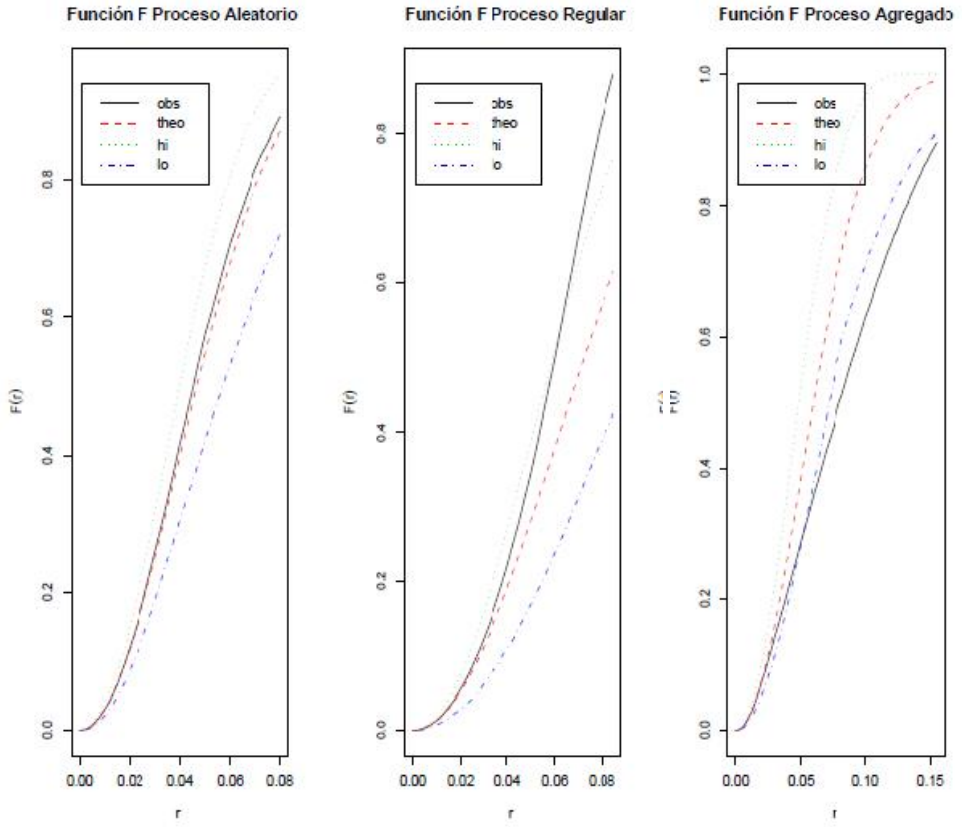
Luego, la función K de Ripley estimada, esta dada por

$$\hat{K}(d) = \frac{\sum\sum_{i \neq j} I_d(d_{ij})}{\hat{\lambda}}. \tag{2.29}$$

Bajo aleatoriedad:

$$K(d) = \lambda\pi d^2, \tag{2.30}$$

en donde: 1) si $\hat{K}(d) < \lambda\pi d^2$, hay regularidad, 2) si $\hat{K}(d) > \lambda\pi d^2$, hay agregación.



(a) Patrones puntuales con distribución aleatoria.

(b) Patrones puntuales regulares.

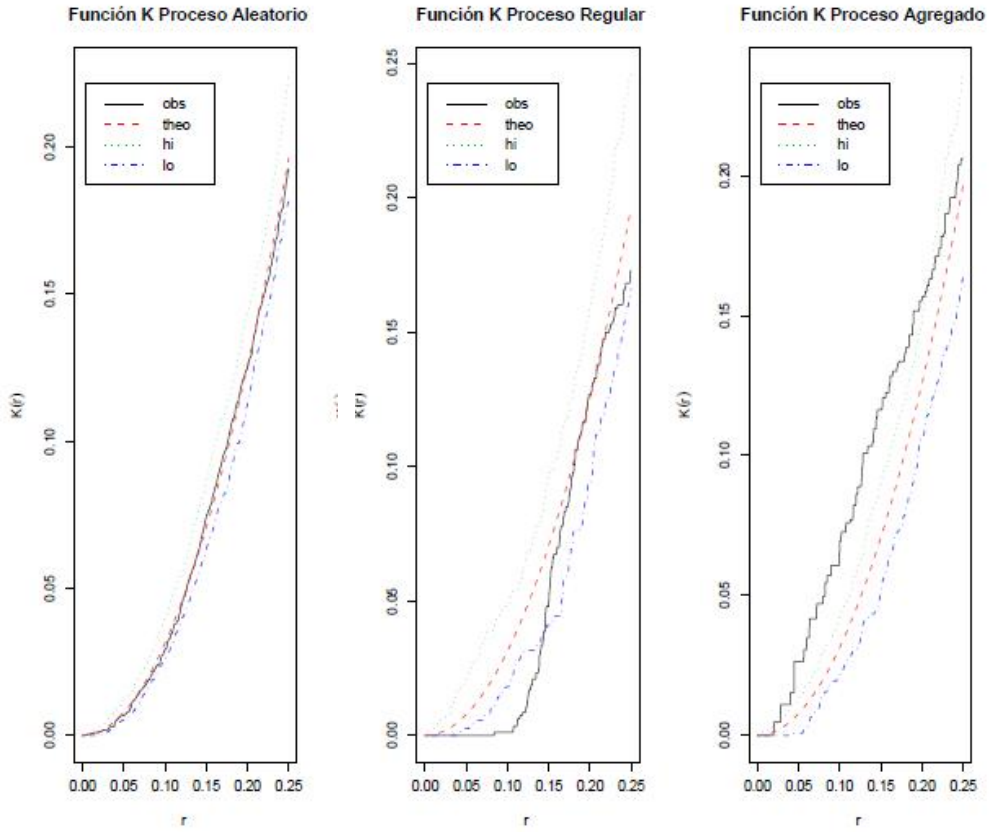
(c) Patrones puntuales agregados.

Figura 2.7: Gráficas del la función F para cada tipo de proceso.

2.4. Estudios alternativos

En términos generales el problema, tanto la cuantificación como la detección de pérdidas en sistemas de agua potable ha sido estudiado de manera consistente, sin embargo, en términos mayoritarios el enfoque investigativo se ha inclinado por la detección de pérdidas físicas y en particular en lo referido a las filtraciones en líneas de distribución, de esta forma es posible encontrar estudios enfocados en la detección mediante el análisis de vibraciones en tuberías soterradas, utilización de algoritmos genéticos, etc. Sin embargo el estudio de las pérdidas aparentes resulta ser mucho más fragmentario.

En esta sección se estudia cómo ha sido abordado el problema, tanto a nivel local (enfocados solo en el ámbito de agua potable) como a nivel global



(a) Patrones puntuales con distribución aleatoria.

(b) Patrones puntuales regulares.

(c) Patrones puntuales agregados.

Figura 2.8: Gráficas del la función K de Ripley para cada tipo de proceso.

(modelos de detecciones de fraudes en otros ámbitos).

2.4.1. Caso Kampala, Uganda

Manteniendo siempre en consideración que los valores particulares en torno a las pérdidas no son necesariamente extrapolables entre poblaciones de características disímiles, resulta valioso considerar el enfoque de cuantificación de pérdidas aparentes o comerciales utilizado en la ciudad de Kampala, Uganda (Humaid & Barhoom, 2012).

Se proponen separar el análisis en base a grupos causales de pérdidas:

i) **Medidores imprecisos**: bajo la premisa de que, al igual que cualquier dispositivo mecánico, los medidores se ven sometidos al desgaste, su-

friendo, como consecuencia, una merma en su precisión, de esta manera, y siguiendo métodos de muestreo estadístico previamente establecidos, se consideraron categorías de medidores, de acuerdo a su antigüedad, y se realizaron pruebas con distintos niveles de flujo.

ii) **Errores en la lectura**: Dado que en Kampala, al igual que en Chile, la recolección de datos de consumo, se realiza de manera manual, por inspectores relativamente poco capacitados, existiendo errores humanos inherentes en la lectura.

iii) **Manejo de datos y errores de facturación**: al comparar los datos entregados por los inspectores al realizar la lectura se detectan nuevos errores.

iv) **Consumos no autorizados**: se considera la alteración de medidores, la instalación de bypass, reposición no autorizada de servicio y conexiones no autorizadas a la red. Una vez detectados, se calcula el volumen de consumo no autorizado, como un promedio de los consumos mensuales previos a la falta, lo anterior en base a la detección mediante auditorías en terreno.

La estimación de la desagregación de las pérdidas aparentes se expresan por:

i) **Medidores imprecisos**: $22 \pm 2\,\%$ de la facturación.

ii) **Errores en la lectura**: $1,4 \pm 1\,\%$ de la facturación.

iii) **Manejo de datos y errores de facturación**: $3,5 \pm 0,5\,\%$ de la facturación.

iv) **Consumos no autorizados**: $10 \pm 2\,\%$ de la facturación.

Para la solución del problema, se propone una estrategia de modelación, desde la consistencia de los datos hasta la propuesta de los modelos con sus respectivos resultados. Los modelos abordados son: máquinas de soporte vectorial y redes neuronales.

2.4.2. Medidores inteligentes de agua potable

Un enfoque alternativo al desarrollo clásico de este tipo de problemas, es reemplazar los antiguos medidores de agua potable por medidores inteligentes. Las ventaja que se tiene sobre los antiguos medidores de agua potables son múltiples, en los que se destacan:

i) **Información**: Existe un sensor remoto que se utiliza para recepcionar la información del cliente a 100 metros de distancia del medidor inteligente. Con esto se evita errores de medición humano cometidos por los inspectores que van a tomar el estado del agua a terreno.

ii) **Manipulación**: Se tiene conocimiento si el medidor ha sido dañado o intervenido por terceros.

iii) **Cortes**: EL medidor inteligente cuenta con una válvula de corte de agua. Esto evita el rompimiento del suelo para intervenir la matriz, evitando costos a las empresas y malos ratos a los clientes.

Este tipo de medidores está muy lejos de ser implementados por las empresas distribuidoras de agua en Chile, debido a los enormes costos asociados.

2.4.3. Fraude en el sistema financiero

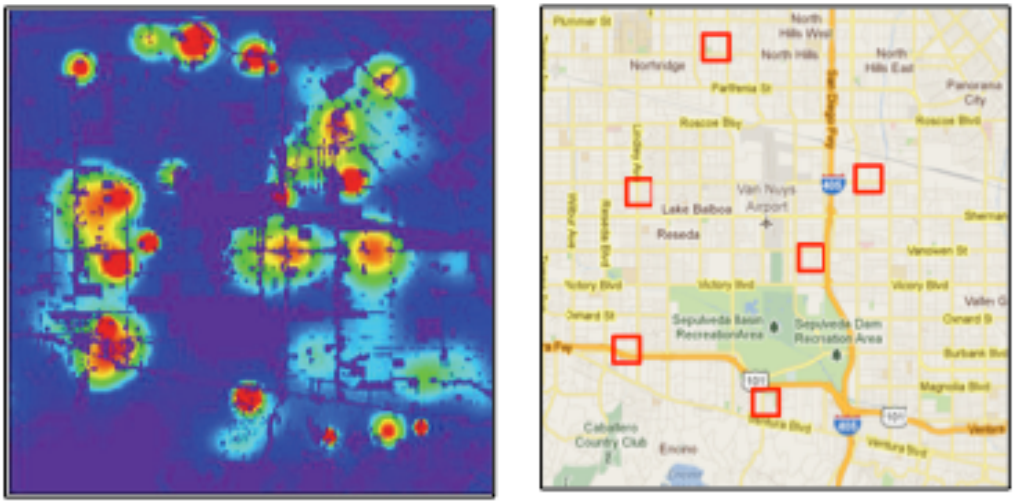
El fraude es un problema serio que enfrentan los emisores de tarjetas de crédito. Las transacciones con tarjeta de crédito tuvieron una pérdida total de 800 millones dólares de fraude en los Estados Unidos el 2004. En Reino Unido, en el mismo año, la pérdida causada por el fraude con tarjeta de crédito fue de 425 millones de libras (aproximadamente 750 millones de dólares EE.UU.). EL retraso de la gestión del riesgo se convierte en uno de los mayores obstáculos para el crecimiento de la rentabilidad. Por tanto, la gestión de riesgos de tarjetas de crédito se convierte en uno de los temas más importantes para los investigadores en el sector financiero privado.

Debido a la importancia del asunto, se ha desarrollado una basta teoría de modelos de predicción en el fraude financiero. La mayoría de estas investigaciones va enfocada en modelos estudiados en la rama de estadística y la computación. Es en esta última rama donde se han enfocado con mayor énfasis, debido al gran impacto que han causado la sub-ramas de la programación como lo son el aprendizaje de máquinas y la minería de datos.

Para más información, ver algunas de las siguientes referencias: Quah & M. Sriganesh (2008), Duman & Ozcelik (2011), Ngai et al. (2011).

2.4.4. PredPol

Predpol (en inglés “Predictive Policing”), es un software que tiene como objetivo mostrar las zonas geográficas en donde es más probable que se cometa un crimen dado que se dio un crimen en cierto espacio-tiempo. El algoritmo que utiliza PredPol necesita tres entradas: tipo de crimen, lugar donde



(a) Mapa de Calor. **(b)** Lugares más probables de crimenes.

Figura 2.9: Esquema del software PredPol.

se cometió y fecha exacta cuando tuvo lugar. Una vez ingresada esta información, el software muestra dos imágenes. La primera imagen (Figura 2.9a) corresponde un mapa de calor, en el cual se señala donde es más probable que suceda un crimen en las proximas horas. La segunda imagen (Figura 2.9b) corresponde a un mapa satelital, en el cual se señala con sectores rectangulares los lugares donde deben visitar los agentes de turnos.

Este software ha sido implementado en varias oficinas de los Estados Unidos: Florida, Miami, Atlanta, Los Angeles, entre otros estados. Por ejemplo, en el año 2013 , en el estado de Atlanta, el número de crímenes disminuyó en un 13 %.

Por otro lado, no existen publicaciones científicas que detallen los modelos ocupados en el algoritmo, limitándose los alcance del método.

Capítulo 3

Modelamiento y resultados

En esta sección se detallan los pasos a seguir en el modelamiento y resolución del problema. Al final de esta sección se presentan los resultados obtenidos durante la investigación.

3.1. Análisis de los datos

El conjuntos de datos recabados corresponden a los clientes de la Empresa de Servicios Sanitarios de Valparaíso (ESVAL). La información proporcionada se encuentra segregada en los siguientes conjuntos de datos.

a) **Datos de consumo**: Información del consumo mensual de los clientes entre los años 2010 al 2014.

b) **Datos de localía**: Información de la localía (o comuna) a la cual pertenece el cliente.

c) **Datos de localización espacial**: Información de la ubicación del cliente en coordenadas espaciales.

d) **Datos de clientes fraudulentos**: Información de la fecha y tipo de fraude cometidos por el cliente.

3.1.1. Preparación de los datos

Del conjunto de datos disponibles, se crea una nueva base de datos con toda la información proporcionada. En la tabla Tabla 3.1 se definen las variables del modelo.

Tabla 3.1: Definición de las Variables

Variable	Tipo	Definición
id cliente	Caracter	Identificador único del cliente
localidad	Caracter	Indica la localidad a la cual pertenece el cliente
coordenadas	Numérico	Coordenadas espaciales del cliente
fecha fraude	Caracter	Indica la fecha del fraude
tipo fraude	Caracter	Indica el tipo de fraude
indicador fraude	Numérico	1: Se comete fraude durante el 2010-2014, 0: e.o.c.
consumo	Numérico	Consumo del cliente en el mes yy durante el año xx.

3.1.2. Análisis exploratorio de los datos

En esta sección se presenta un resumen de las principales características del conjunto de datos, mediante tablas y/o gráficos. El fin del análisis exploratorio es comprender la naturaleza de la problemática, y con ello establecer los criterios necesarios para dar paso a la etapa de la modelación.

a) Información general del conjunto de datos

El conjunto de datos contiene la información de aproximadamente 850 mil clientes de la empresa ESVAL, estos se encuentran repartidos en 54 comunas de la V región. Respecto a los datos de consumo, se tiene el consumo mensual de los clientes entre los años 2010 y 2014, correspondientes a un total de 60 meses de consumo.

Por otro lado, existen clientes que presentan información incompleta en su consumo, localía y/o ubicación. , es decir, se está en presencia de un problema de **datos perdidos**. Dado que no se asume de momento ningún supuesto distribucional, se opta por separar el conjunto de datos en dos conjuntos:

- **Información Real**: En ella se tiene la información de todos los clientes.
- **Información Efectiva**: En ella se tiene la información de todos los clientes que no presentan problemas de datos perdidos.

En la Tabla 3.2 se compara la información disponibles en ambos conjuntos.

De aquí en adelante, se considera el conjunto de datos correspondiente a la información efectiva disponible.

Tabla 3.2: Comparación entre la información real y la información efectiva.

Descripcion	Informacion total	Informacion efectiva	% de uso
Total clientes	843.833	259.796	30.79
Total clientes fraudulentos	59.692	25.793	43.21
Total clientes fraudulentos esp.	11.932	3.915	32.81

En la Tabla C.3 (ver apéndice C) se muestra la información de los tipos de clientes y los porcentajes relativos de clientes fraudulentos correspondiente a cada comuna. Al considerar solo los clientes fraudulentos especiales, el porcentaje relativo de estos no supera el 5 % del total de la población. Cabe recordar que según la Superintendencia de Servicios Sanitarios alrededor de un 5 % a un 10 % corresponde a clientes fraudulentos, por lo tanto existen clientes fraudulentos que no han sido detectados.

b) Tipo de cliente

Los clientes son clasificados en dos categorías basado en su consumo histórico: clientes con un consumo normal y clientes con un consumo fraudulento.

El perfil de un cliente que sigue un tipo de consumo normal, es caracterizado por tener una curva de consumo variable mes a mes, debido a los diversos factores que influyen en el consumo de un persona, por ejemplo: efectos estacionales, abandono del hogar, fugas, aumento del número de integrantes entre otros factores. Existen casos donde se mantiene una tendencia en su consumo anual, por ejemplo, los clientes de la comuna de Cartagena mantienen un consumo elevado durante el periodo de verano (Diciembre, Enero y Febrero), mientras que el resto del año su consumo baja considerablemente. En la Figura 3.1, se muestra el consumo mensual entre los años 2010 al 2014 de un cliente de la comuna de Cartagena. Por otro lado, existen casos donde los clientes tienden a mantener un consumo anual más caótico, donde no es posible definir claramente algún tipo de tendencia, por ejemplo, en la comuna de Valparaíso, el consumo anual de los clientes no mantiene ninguna tendencia. En la Figura 3.2, se muestra el consumo mensual entre los años 2010 al 2014 de un cliente de la comuna de Valparaíso.

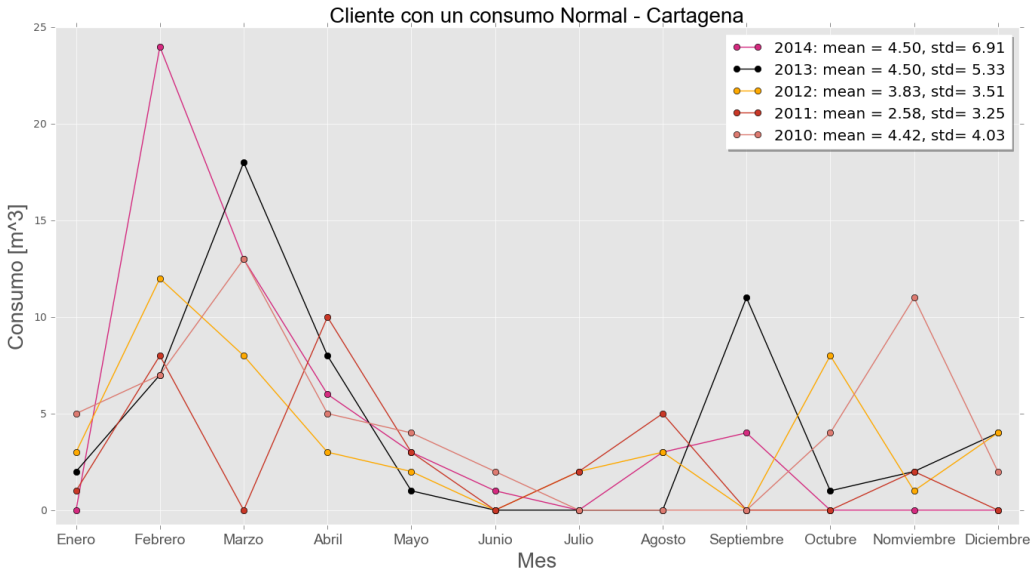


Figura 3.1: Perfil de un cliente Normal con tendencia en su consumo anual.

La característica esencial de que un cliente sigue un consumo normal es su curva de consumo promedio anual, ya que dicho consumo se mantiene relativamente estacionario año a año (ver Figura 3.3).

Por otro lado, el perfil de un cliente que sigue un tipo de consumo fraudulento, es caracterizado por la existencia de periodos donde el consumo se mantiene relativamente constante y significativamente bajo respecto al promedio anual (ver Figura 3.4). Cuando se analiza el consumo promedio anual, se observa que la curva ya no es estacionaria como el caso de un cliente que sigue un consumo normal (ver Figura 3.5).

3.2. Modelación

En esta sección se establece los pasos a seguir para dar solución al problema. En primer lugar, se toma en cuenta el conjunto de datos que no presenta problemas de datos perdidos. En segundo lugar, debido a la magnitud del conjunto de datos, se opta por trabajar el problema por comunas. Esto disminuye el gasto computacional y se logra realizar un estudio más especializado por sector.

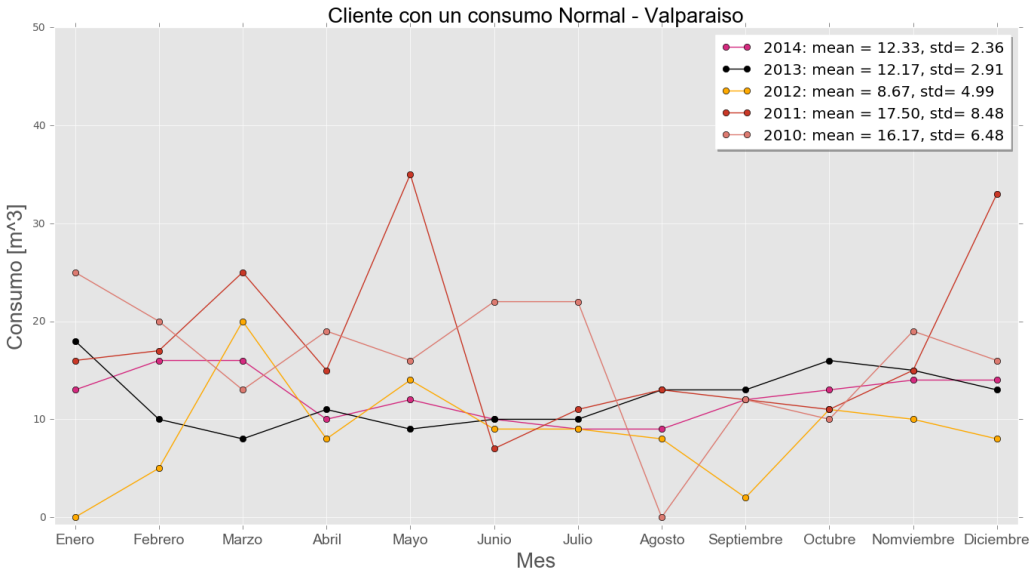


Figura 3.2: Perfil de un cliente Normal sin tendencia en su consumo anual.

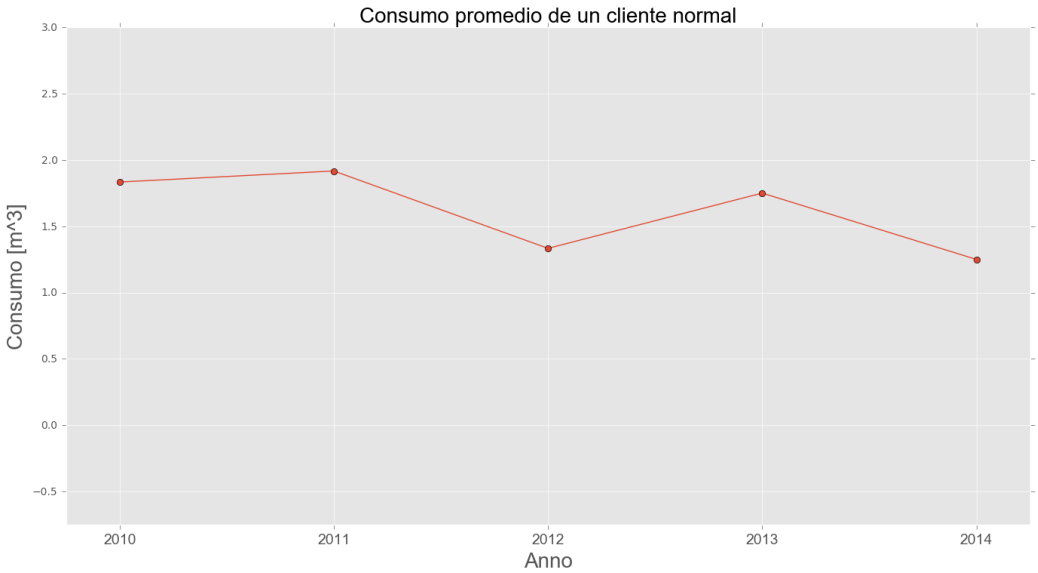


Figura 3.3: Perfil del consumo promedio de un cliente normal.

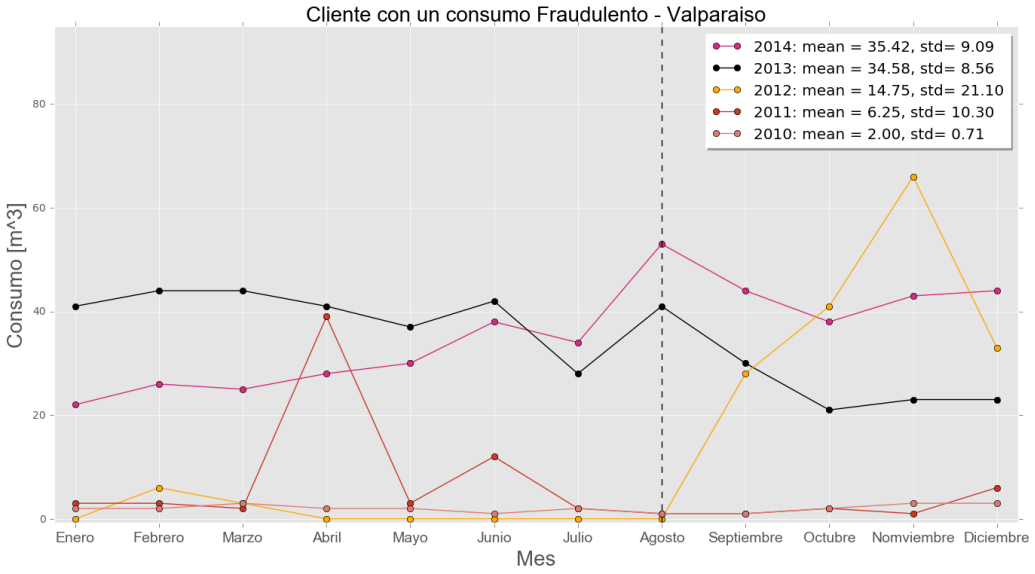


Figura 3.4: Perfil de un cliente fraudulento.

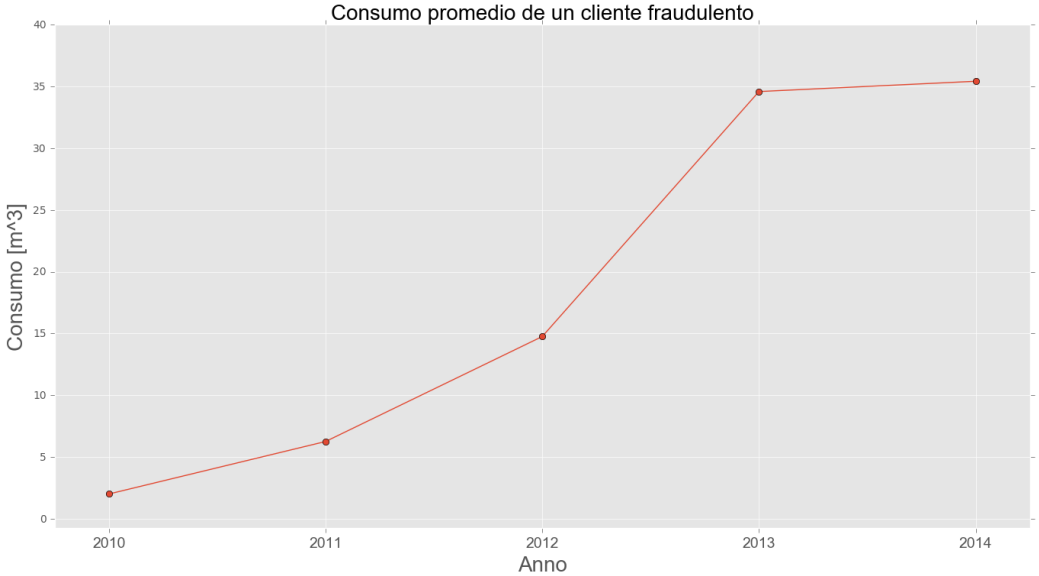


Figura 3.5: Perfil del consumo promedio de un cliente fraudulento.

3.2.1. Aplicación e interpretación de los modelos de regresión

El conjunto de entrenamiento se define por:

- $\boldsymbol{x}^{(i)} = (x_1^{(i)}, ..., x_{60}^{(i)})$: Consumo mensual del cliente i durante los años 2010 y 2014.
- $y^{(i)}$: 1, si el cliente i cometió algún tipo de fraude durante los años 2010 y 2014. 0, en otro caso.

Definido el conjunto de entrenamiento, se procede a aplicar los modelos descritos en la Sección 2.1: regresión logística (lr), bosque aleatorio (rf) y máquinas de soportes vectorial (svm). Para comparar los resultados de los distintos modelos, se ocupan las medidas de rendimiento descritos en la Sección 2.2: accuracy, precision, recall y f-score. Para obtener una banda de confiabilidad de las medidas de rendimientos se calculan intervalos de confianza asintóticos, para ello el conjunto de entrenamiento se divide aleatoriamente en dos conjuntos: "training dataset" y "testing dataset", cada uno tiene el 70 % y 30 % de la información del conjunto de datos, respectivamente. El procedimiento es el siguiente: se ajustan los modelos sobre el conjunto "training dataset", luego se procede a predecir los resultados ocupando el conjunto "testing dataset", con estas predicciones se calculan las medidas de rendimiento. El procedimiento anterior se repite 100 veces. Con esta información se calcula el promedio y la desviación estándar de las medidas de rendimientos y se da paso para encontrar los intervalos de confianza asintóticos del promedio mediante la ley de los grandes números. Se debe tener cuidado con la interpretación de las medidas de rendimiento, debido a que el conjunto de datos se encuentra desproporcionado (aproximadamente se tiene proporciones de 95 % de clientes normales y 5 % de clientes fraudulentos sobre el total de clientes). Estos resultados serán buenos a medida que el total verdaderos positivos sea alto , es decir, el total de verdaderos positivos sea lo más cercano al total de clientes fraudulentos, mientras que el número de falsos positivos sea lo más bajo posible (idealmente igual a cero).

Aplicados los diferentes modelos, se aborda el problema de los clientes fraudulentos que no fueron detectados como tal en el conjunto de datos. En este caso, un falso positivo corresponde a un cliente que sigue un consumo normal pero que el modelo detecto como cliente fraudulento y considerando el hecho que aproximadamente entre un 5 % a 10 % de la población del país corresponde a clientes fraudulentos, es probable que al menos uno de estos falsos positivo corresponda efectivamente a un cliente fruadulento (si es que

el número total de clientes fraudulentos es menor al 5 % de la población total de la comuna). Por tanto, estos falsos positivos serán clasificados como **clientes sospechosos**. Se definen tres tipos de clientes sospechosos:

- **Cliente sospechoso A**: Cliente detectado como un falso positivo por un modelo solamente.
- **Cliente sospechoso B**: Cliente detectado como un falso positivo por dos modelos simultáneamente.
- **Cliente sospechoso C**: Cliente detectado como un falso positivo por los tres modelos simultáneamente

Otro aspecto importante es que la suma de los clientes sospechosos más los clientes fraudulentos no debe superar el 10 % de la población.

El objetivo de este trabajo es encontrar aquellos clientes fraudulentos que no han sido detectados por las empresas de servicios de agua potable. En este caso, los candidatos a ser este tipo de clientes serán los clientes detectados como sospechosos. Para verificar si un cliente sospechoso corresponde efectivamente a un cliente fraudulento, las empresas deben mandar inspectores a las direcciones de los clientes detectados como sospechosos.

3.2.2. Análisis espacial de los clientes fraudulentos

Finalizada la parte de clasificación, se procede a realizar una análisis espacial de los datos. Aquí, se estudia la aleatoriedad de las coordenadas espaciales de lo clientes fraudulentos mediantes los test basados en cuadrantes (Test Xi-cuadrado) y en distancia (Función G,F,K-de Ripley), descritos en la sección 2.3.

Si los test indican que los clientes fraudulentos no siguen un patrón completamente aleatorio, se podría definir sectores específicos donde los inspectores puedan realizar la búsqueda de clientes fraudulentos, ahorrando tiempo y dinero a las empresas de servicios básicos de agua. La investigación realizada da los primeros pasos para el desarrollo de un software capaz de indicar las zonas geográficas que deben visitar en la búsqueda de los nuevos clientes fraudulentos, sin embargo, esto queda como trabajo a futuro.

3.3. Evaluación y pruebas

La comuna para desarrollar la metodología descrita es Cartagena. Esta comuna registra 1573 clientes con consumo continuo durante los años 2010 al

2014, con un total de 66 clientes fraudulentos (correspondiente a un 4.20 % de la población). El porcentaje de clientes fraudulentos se encuentra por debajo del intervalo límite establecido de clientes fraudulentos estimados (de alrededor de un 5 % a un 10 % por comuna).

3.3.1. Modelos de clasificación

En la Ecuación (3.1) se muestran los resultados para los distintos modelos vía la matriz de confusión. El número de verdaderos positivos de cada modelo (49 para el modelo de lr, 50 para el modelo de rf y 62 para el modelo de svm) esta cercano al número de clientes fraudulentos (66). Por otro lado, si se considera que a los más el número de clientes fraudulentos estaría rondando los 157 clientes, este número se encuentra por debajo del total de clientes fraudulentos detectado por cada modelo. Esto se debe al gran número de falsos positivos detectados en cada modelo.

$$CM_{lr} = \begin{pmatrix} 49 & 17 \\ 321 & 1186 \end{pmatrix} \ , \ CM_{rf} = \begin{pmatrix} 50 & 16 \\ 272 & 1235 \end{pmatrix} \ , \ CM_{svm} = \begin{pmatrix} 62 & 4 \\ 170 & 1337 \end{pmatrix} \tag{3.1}$$

En la Tabla 3.3 se muestran los intervalos de confianzas asintóticos de las medidas de rendimiento para los distintos modelos. Los resultados de medida de rendimiento "Accuracy" de los modelos son bastante similares entre si (alrededor de un 80 %), esto se debe a la cantidad de clientes normales que tiene el conjunto de entrenamiento. Por otro lado, en las otras tres medidas de rendimiento, el modelo SVM es levemente mejor que el modelo de LR y bastante mejor que el modelo RF.

Tabla 3.3: Intervalo de Confianza para el promedio de las medidas de rendimiento a nivel 0,95.

Medidas de Rendimiento	Regresión Logística	Bosque Aleatorio	Máquina de Vectores de Soporte
Accuracy	[77.94, 78.81]	[80.75, 82.79]	[80.75, 81.73]
Precision	[37.42, 41.78]	[18.48, 22.02]	[34.91, 39.50]
Recall	[7.68, 8.55]	[5.35, 6.55]	[8.42, 9.45]
F-score	[12.68, 14.06]	[8.07, 9.68]	[13.48, 15.04]

En lo que respecta al tipo de cliente, en la Tabla 3.4 se muestran los resultados obtenidos de los modelos ajustados. Cabe destacar que la suma de los clientes fraudulentos más los clientes sospechosos tipo C suman un total de 88 clientes (un 5.6 % de la población de Cartagena), es decir, si todos los clientes detectados como sospechoso tipo C fuesen fraudulentos, este número de clientes estaría dentro de los márgenes estipulados de clientes fraudulentos por comuna.

Tabla 3.4: Clasificación del tipo de cliente en la comuna de Cartagena.

Tipo de Cliente	Total
Normal	1065
Sospechoso A	284
Sospechoso B	136
Sospechoso C	22
Fraudulento	66

3.3.2. Análisis espacial

En la Figura 3.6 se muestra un mapa satelital de la comuna de Cartagena con la distribución de los clientes. Luego, en la Figura 3.7 se muestra un mapa satelital solo de los clientes fraudulentos. Ahora se procede a aplicar los test de aleatoriedad sobre las coordenadas espaciales de los clientes fraudulentos.

a) **Test Basado en los Cuadrantes**: Para este caso solo se considera el test Chi-cuadrado. Los resultados de la Tabla 3.5 indican que al aplicar el test se obtiene un p-valor < 0.05. Por lo tanto se concluye que las coordenadas espaciales de los clientes fraudulentos siguen un patrón puntual agrupado.

Tabla 3.5: Resultados del test de cuadratura en la comuna de Cartagena.

Test	χ^2	df	p-valor
Chi-Cuadrado	17.938	5	0.006052

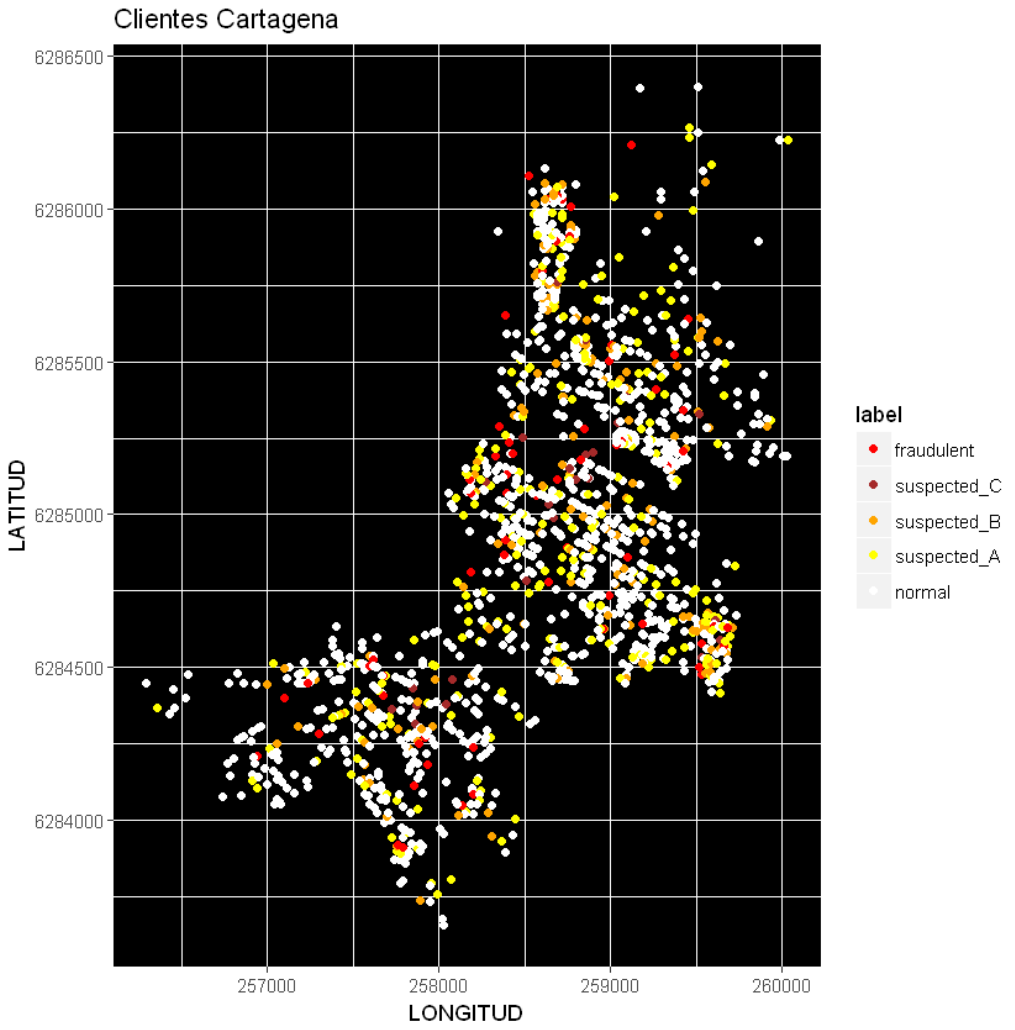


Figura 3.6: Mapa de los clientes de Cartagena clasificado según el tipo de cliente.

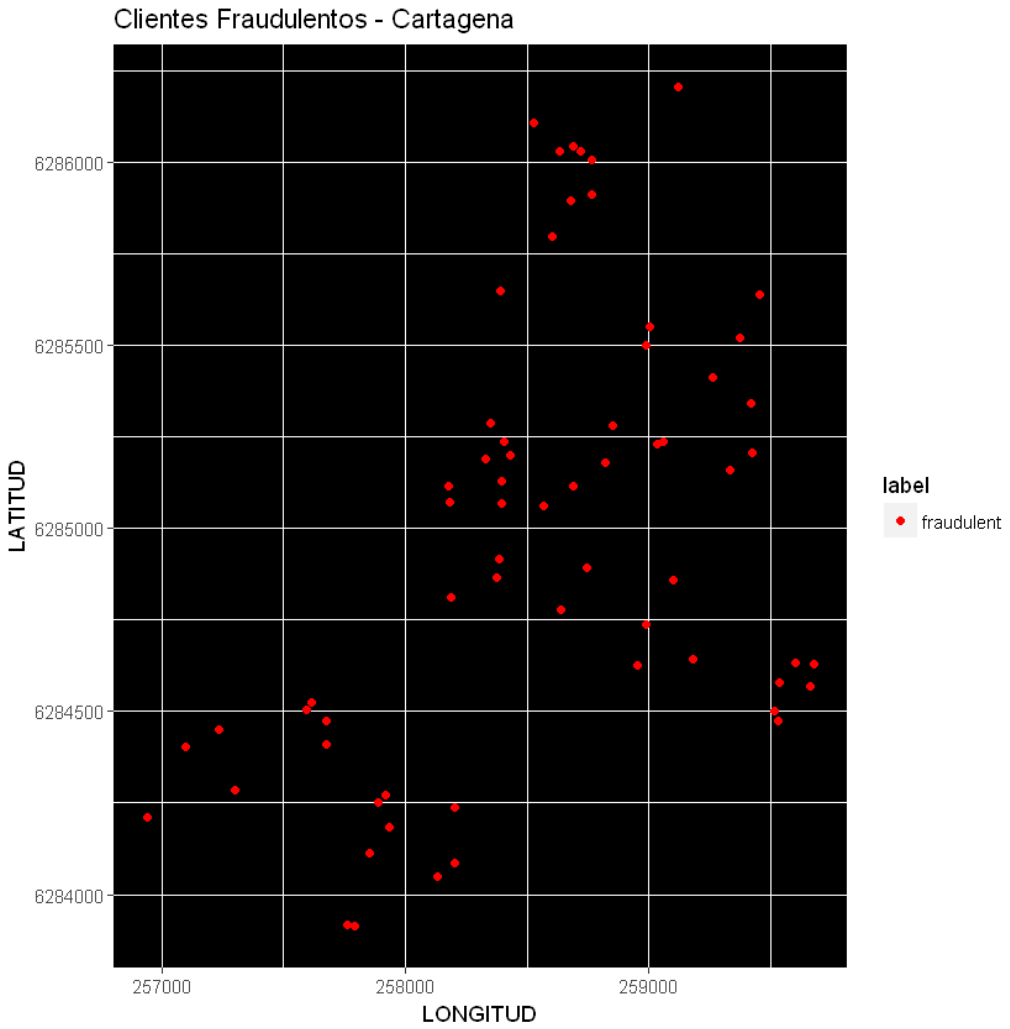
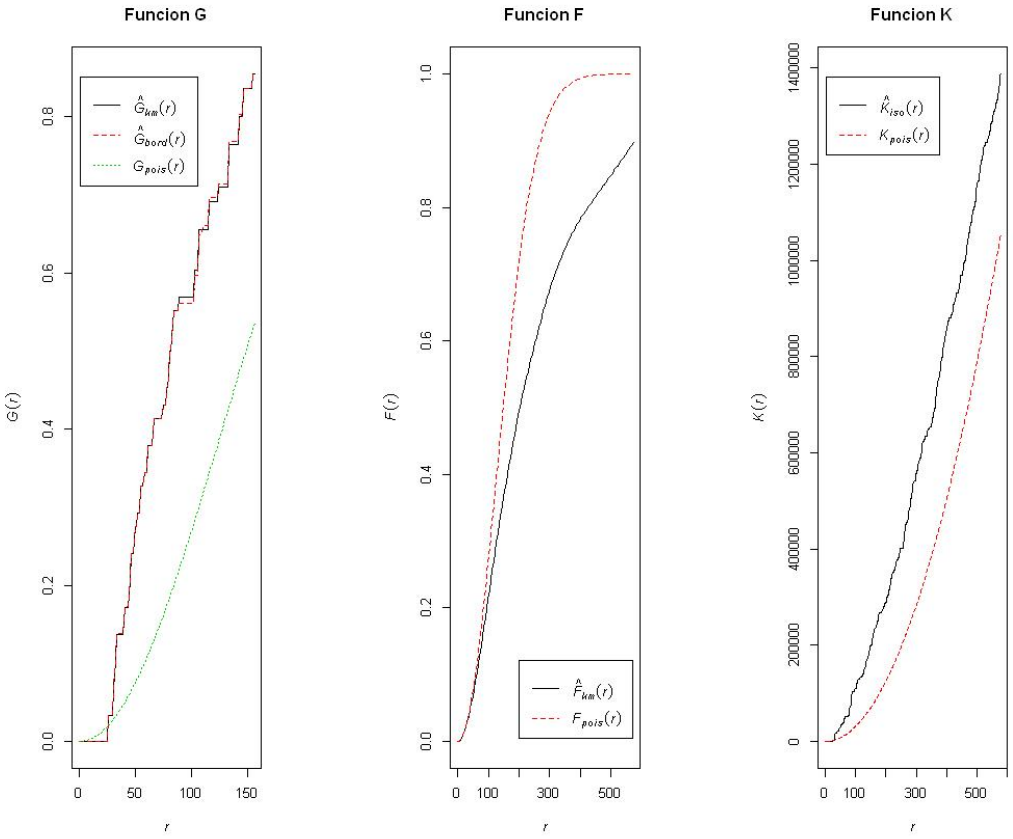


Figura 3.7: Mapa de los clientes de fraudulentos de Cartagena. Los test de aleatoriedad espacial indican que este patrón sigue el comportamiento de un patrón puntual agrupado.

b) **Test Basado en las distancias**: Para este caso se considera las funciones G, F y K de Ripley. En la Figura 3.8 se obtienen las gráficas de aplicar las distintas funciones de distancias. Según la forma que adoptan las gráficas, se concluye que las coordenadas espaciales de los clientes fraudulentos siguen un patrón puntual agrupado.



(a) Función G. **(b)** Función F. **(c)** Función K.

Figura 3.8: Resultados de los test basados en la distancia.

De los distintos test de aleatoriedad, se concluye que las coordenadas espaciales de los clientes fraudulentos siguen un patrón puntual agrupado. En la Figura 3.9 como estos clientes fueron agrupados mediante 4 grupos diferentes.

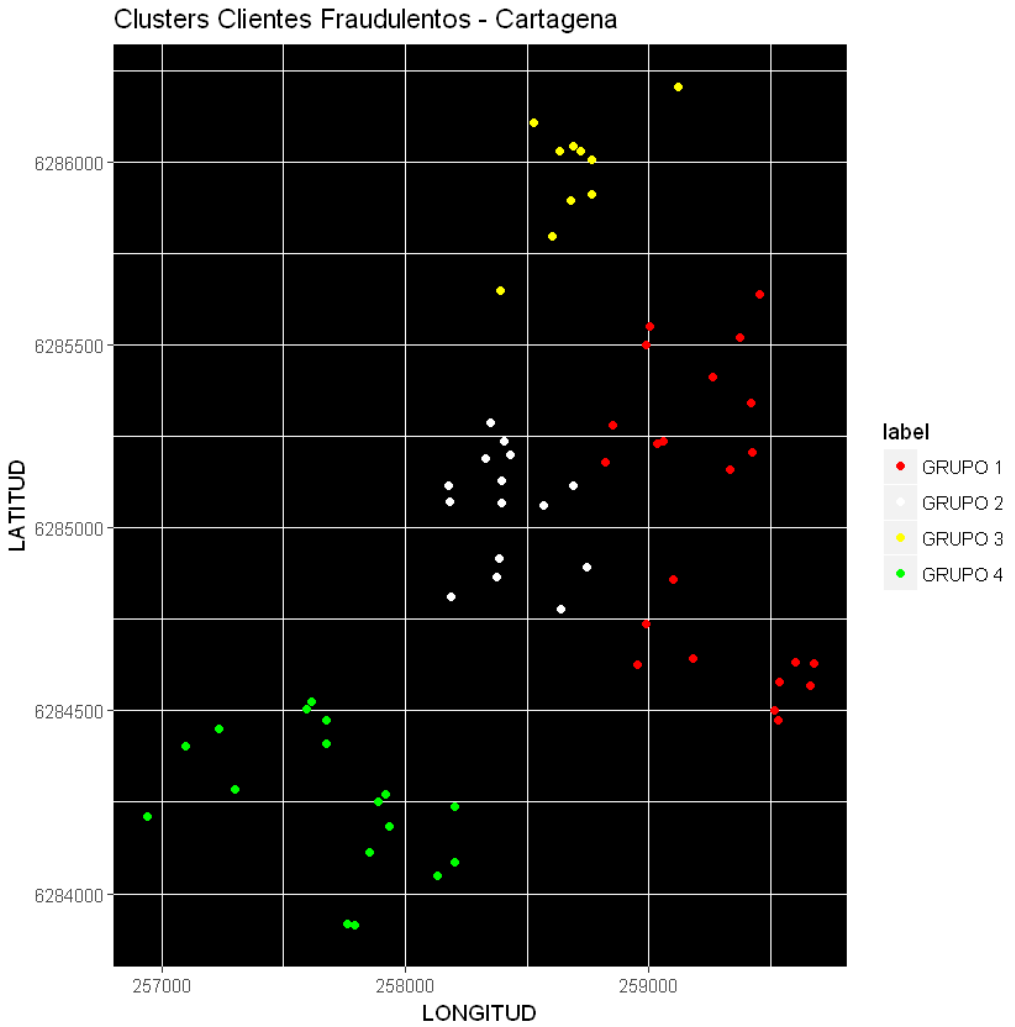


Figura 3.9: Mapa geográfico de los clientes fraudulentos de Cartagena separados por grupos.

Capítulo 4

Conclusiones

Los principales resultados obtenidos durante la investigación fueron:

a) En lo que respecta al análisis de datos, se establece las principales diferencias entre los consumos históricos de los clientes normales y los clientes fraudulentos. Además, se logra identificar patrones o tendencias en el consumo anual de ciertas comunas, por ejemplo, en la comuna de Cartagena los clientes durante el periodo de verano mantienen consumos más elevado en comparación al resto del año.

b) La metodología propuesta arroja resultados satisfactorios, los modelos propuestos se ajustan bastante bien al conjunto de entrenamiento, y se desarrolla un procedimiento para identificar los clientes fraudulentos que no han sido detectados, cumpliendo con el objetivo del problema. Sin embargo, estos resultados no fueron comprobados en terrenos por los inspectores, debido a la poca interacción que hubo con la empresa ESVAL.

c) Se realiza un primer avance respecto al análisis espacial del problema. El estudio enfocado a las coordenadas de los clientes fraudulento basado en las distintas pruebas de aleatoriedad indicaron que este proceso espacial sigue un patrón agrupado. La información que se puede rescatar del análisis espacial puede ser de utilidad para el desarrollo de un software capaz de identificar las zonas donde es más probable encontrar un cliente fraudulento. Además, se buscaría desarrollar una interfaz gráfica que muestre estos resultados, similar a la visualización del software Predpol (descrito en la Sección 2). Esto se deja como un eventual trabajo futuro.

Trabajos de investigación futuro corresponden a añadir nuevas variables de decisión al conjunto de entrenamiento, por ejemplo, la información socio-económica. Esta información serviría para contrastar la hipótesis si los clientes tienden a encontrarse en un estrato en particular o no. Adicionalmente, es de interés validar en terreno los resultados obtenidos durante la investigación, con el fin de contrastar la efectividad teórica respecto a la efectividad real.

Considerando que la metodología propuesta se enfoca principalmente en el consumo de sus clientes, esta se podría replicar sin problemas a otros servicios básicos, por ejemplo, servicios de luz, telefonía, cable, gas, entre otros.