UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICA

Minería de Datos para la predicción en series de tiempo: una aplicación a datos de tráfico de voz y tráfico de datos.

Trabajo Especial de Grado presentado ante la ilustre Universidad Central de Venezuela por el **Br. Adelis C. Nieves C.** para optar al título de Licenciado en Matemática.

Tutor: Dr. Ricardo Ríos

Caracas - Venezuela

28 de Abril de 2009

Nosotros, los abajo firmantes, designados por la Universidad Central de Venezuela como integrantes del Jurado Examinador del Trabajo Especial de Grado titulado "**Minería de Datos para la predicción en series de tiempo: una aplicación a datos de tráfico de voz y tráfico de datos**", presentado por el **Br. Adelis C. Nieves C.** titular de la Cédula de Identidad **V-15.315.164**, certificamos que este trabajo cumple con los requisitos exigidos por nuestra Magna Casa de Estudios para optar al título de **Licenciado en Matemática**.

Dr. Ricardo Ríos

Tutor

Dr. José Rafael León

Jurado

Dr. Mairene Colina

Jurado

Dedicado a:

Mi madre María Cordeiro.

Mis hermanos, Manuel Agustín Nieves Cordeiro,

Darvin y Luis Guillermo Rodríguez.

Agradecimiento

La elaboración de este trabajo ha sido producto de muchos días y noches de arduo esfuerzo. Esfuerzo que se traduce en alcanzar el éxito en una etapa muy importante de mi vida. Muchas personas merecen mi agradecimiento por su apoyo y por comprender lo importante que es para mi escribir estas palabras en este momento. De manera muy especial quiero agradecer a mi madre María Cordeiro por estar presente cada noche que pudo con su taza de café, a la abuela Josefa, a mis hermosos hermanos Darvin y Luis Guillermo por llegar a mi vida en los momentos más difíciles y brindarme sus bellas sonrisas y ocurrencias que me que me impulsaron a continuar, y a mi hermano Manuel Agustin que aunque no esté en cuerpo presente, siempre ha sido mi motivo de ser y contiuar.

Quiero agradecer especialmente a Elia Gonźalez (La Madre) por ser mi mejor amiga, apoyarme y estar siempre alli diciendo: " todo va a estar bien ", a Sofía, Lili, Freddy, Gledys, Dimitar, Zaira, Mileidys, Ubencio, Andrés, Elvis Lacruz, Daniela Palacios, Jesús Yerena, Luis Arleo y demás amigos por hacerme feliz, enseñarme y compartir conmigo cada día especial de mi vida.

Quiero agradecer a Samuel Carvajal por darme amor, apoyo y comprensión durante 6 años y a su madre Cecilia Mujica por brindarme su techo y hacer posible muchas de mis metas académicas.

Agradezco especialmente a Luis Araujo por formar parte de un proceso de renovación y cederme un espacio tranquilo para culminar este proyecto. A la doctora Juana Castillo por su sabiduría y permitirme conocerle, a ella especialmente agradezco la culminación exitosa de todas las fases de este trabajo.

Quiero agradecer de manera especial a mi tutor Ricardo Ríos por su sabiduría, comprensión, apoyo, confianza, dedicación, en fín, él ha sido el responsable de esto. Rico, gracias por tu amistad. A los profesores Mairene Colina y Rafael León por su colaboración, correciones y sugerencias.

Deseo agradecer a María Jiménez por haberme facilitado toda la información necesaria para la elaboración de este proyecto y además por sus conocimientos, apoyo y dedicación.

Finalmente quiero agradecer a Miguel Torres, por aparecer y permancer en mi vida, darme seguridad, confianza, apoyo, amor, sabiduría y mantenerme arriba, subiendo y bajando las más duras colinas, siempre juntos de la mano, GRACIAS!

Indice General

Indice de Figuras

Indice de Tablas

Introducción

El mercado de telefonía móvil en Venezuela cuenta con operadoras móviles de red propia que permiten la transferencia de grandes cantidades de datos. Con el origen de la telefonía móvil CDMA (de sus siglas en inglés Code Divison Multiple Access) y TDMA (Time Division Multiple Access) se proporciona la posibilidad de transferir tanto voz (una llamada telefónica) y datos no - voz (descarga de programas, intercambio de email y de mensajería de texto y multimedia). La primera es una forma de tecnología que permite numerosas señales para ocupar un solo canal de transmisión, optimizando el uso de ancho de banda posible. La segunda es utilizada en telefonía móvil digital y divide cada canal de la comunicación celular en tres franjas horarias con el fin de aumentar la cantidad de datos que se puedan llevar.

La evolución o transformación de este tipo de tecnología ayuda a cubrir las grandes demandas de comunicación inalámbrica que se generan a nivel mundial, [14]. Esa evolución y transformación depende, en gran parte, de las investigaciones y estudios que se hacen sobre los datos generados por los equipos que usan dicha tecnología, con la finalidad de asignar los recursos necesarios para ofrecer un mejor servicio y cubrir la alta demanda de la misma.

Este Trabajo Especial de Grado está enfocado hacia la aplicación de técnicas estadísticas creadas para el análisis, descripción y predicción en una gran base de datos continente de información temporal multivariada relacionada con aspectos varios de la telefonía celular en Venezuela.

La telefonía celular es una de las áreas de las telecomunicaciones que experimenta una mayor expansión y crecimiento en Venezuela, el interés de las empresas del ramo es mantenerse en un mercado de alta competencia; en ellas se generan grandes cantidades de datos diariamente, lo que dificulta mucho tomar decisiones en tiempo real. Por ejemplo: el tráfico de voz, mensajes de texto y multimedia se acumula en términos de series temporales vectoriales con una muy compleja interacción entre sus componentes. Esta información se presenta como flujo de datos (del inglés *Data Stream*). Un flujo de datos supone un conjunto de registros multidimensionales $X_1, X_2, \ldots, X_n$ en horas o fechas de llegada $t_1, t_2, \ldots, t_n$ lo que prácticamente señala una serie temporal; además, cada X_i es un registro multidimensional que contiene d dimensiones denotadas por $X_i = (x_i^1, \ldots, x_i^d)$ y cada registro X_i en

la formación del flujo de datos está asociado con una clase o etiqueta $\mathcal{C}_i$, [1], [4]. En el desarrollo de este trabajo se considerará nuestra base de datos como flujos de datos con estas características.

Tener una metodología de análisis, descripción y predicción para estas series, tan importantes en la definición de políticas varias en las empresas, se ha convertido en un tema de creciente interés, tanto desde el punto de vista teórico como práctico y será la directriz principal de este trabajo, [16]. Para ello estudiaremos un conjunto de datos, con múltiples factores pertenecientes a una serie temporal vectorial, correspondiente a los registros diarios por horas de cinco meses del año 2007, del tráfico de información de telefonía móvil en varias regiones de Venezuela, extraídos de radio bases (MTX) asignadas a una empresa de telefonía móvil venezolana. Nuestro trabajo se enmarca en los llamados procesos de conocimiento por descubrimientos en grandes bases de datos conocidos como KDD (por las siglas en inglés de Knowledge Discovery in Data Bases). Una herramienta poderosa a usar es la llamada Minería de Datos (Data Mining), que contempla el análisis de datos multivariados en conjunto con el estudio de series temporales vectoriales. Estas herramientas serán útiles para el establecimiento de patrones de llamadas telefónicas u otro tipo de información recolectada, modelos de cargas en redes, detección de fraude, modelos de predicción de ganancias y pérdidas, control sobre los planes ofrecidos a los usuarios, etc.

El trabajo está estructurado en 3 capítulos, el primero contentivo de la teoría básica del proceso KDD y Data Mining, Análisis de Componentes Principales e Histogramas Bivariados. El segundo capítulo contiene la aplicación de cada una de las fases del proceso de KDD y Data Mining a la base de datos y la evaluación e interpretación de los resultados extraídos a partir del uso de las técnicas explicadas en el capítulo uno. El tercero, está consagrado a presentar conclusiones y recomendaciones sobre la metodología lograda, su difusión y uso en la toma de decisiones. Finalmente, se presentan dos apéndices donde encontraremos información referida a fundamentos teóricos y resultados de aplicaciones hechas en el capítulo dos.

Capítulo 1

Fundamentos teóricos

1.1 Introducción

En este capítulo se estudian las bases teóricas sobre las cuales reposa el trabajo. Uno de los campos más estudiados en la actualidad está relacionado con la extracción de conocimiento a partir de fuentes masivas de datos, [5]. Para ello se emplean metodologías basadas en procesos que permiten el análisis exhaustivo de una base de datos para obtener información útil que permita tomar decisiones a posterior.

Estos procesos de extracción de conocimiento de las bases de datos se fundamentan en técnicas estadísticas muy poderosas, durante el desarrollo del capítulo se manejaraán conceptos fundamentales empleados en el texto: Descubrimiento de Conocimientos en Bases de Datos (KDD), Minería de Datos, Análisis de Componentes Principales e Histogramas Bivariados. La estructura de este capítulo esta inspirada en los autores señalados en [2], [5], [8] y [12].

1.2 Descubrimiento de conocimientos en bases de datos (KDD)

Actualmente la cantidad de datos que se almacena en las bases de datos excede nuestra habilidad para reducirlos y analizarlos sin el uso de técnicas de análisis automatizadas. La mayoría de las bases de datos, de cualquier origen, crecen a una proporción fenomenal, la idea subyacente es interpretar grandes cantidades de datos y encontrar relaciones o patrones. Para conseguirlo se requiere de técnicas de aprendizaje automatizado, estadística, reconocimiento de patrones y modelos predictivos.

KDD (de sus siglas en inglés Knowledge Discovery in Databases) tiene que ver con la extracción o descubrimiento de conocimiento en bases de datos. El KDD es *el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos.* En esta definición se resume cuáles deben ser las propiedades adecuadas del conocimiento extraído:

- **Válido:** hace hincapié en que los patrones deben seguir siendo precisos para datos nuevos (con un cierto margen de certidumbre), y no sólo para aquellos que han sido usados en su obtención.

- **Novedoso:** que aporte algo desconocido tanto para el sistema como para los usuarios.

- **Potencialmente útil:** la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.

- **Comprensible:** la extracción de patrones no comprensibles dificulta o impide su interpretación, revisión, validación y uso en la toma de decisiones. Más aún, una información incompresible no proporciona conocimiento al menos desde el punto de vista de su utilidad.

El proceso de KDD se inicia con la identificación de los datos. Para ello es necesario tener una idea de cuáles datos se necesitan, dónde se pueden encontrar y cómo extraerlos. Una vez que se dispone de los datos, se deben seleccionar aquellos que sean útiles para los objetivos propuestos y deben ser preparados para ponerlos en un formato adecuado.

Cuando se cuenta con los datos correctos se procede a la Minería de Datos, proceso en el que se seleccionarán las herramientas y técnicas adecuadas para lograr los objetivos pretendidos. Y tras este proceso se culmina con el análisis de los resultados que será lo que nos aportará el conocimiento esperado.

Como se deduce de la anterior definición, el KDD es un proceso complejo que incluye no sólo la obtención de los modelos o patrones (el objetivo de la Minería de Datos), sino también la evaluación y posible interpretación de los mismos, tal y como se refleja en la siguiente figura:

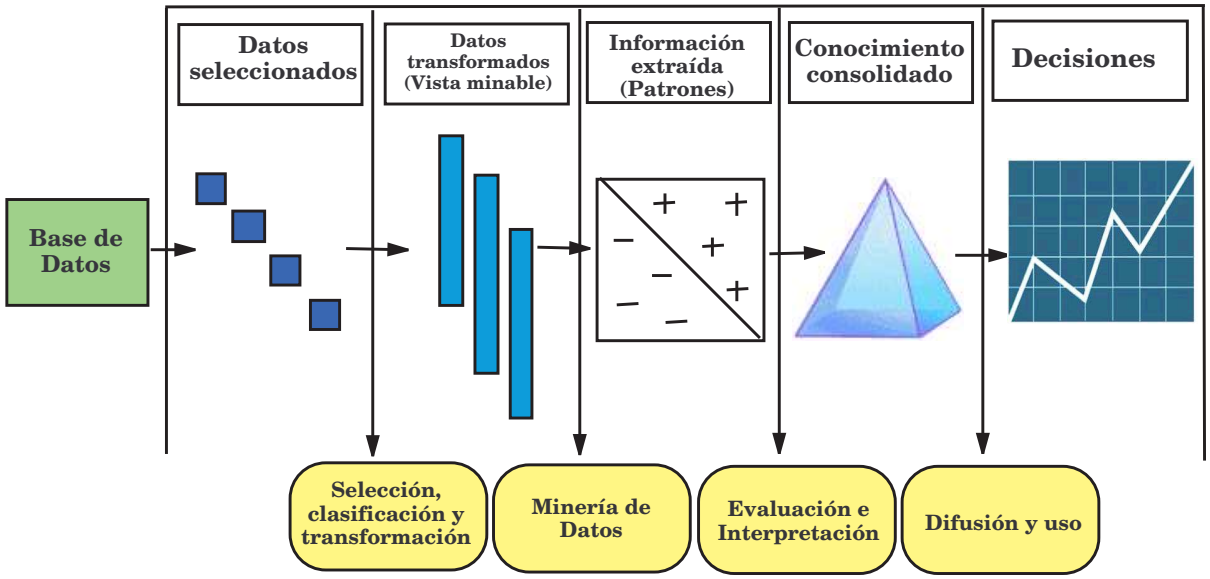

Figura 1.1: Metodología para el descubrimiento de conocimiento en bases de datos.

KDD es un proceso interactivo e iterativo, que involucra numerosos pasos e incluye muchas decisiones que deben ser tomadas por el usuario, y se estructura en las siguientes etapas:

- **Comprensión del dominio de la aplicación:** es decir, del conocimiento relevante de los objetivos del usuario final.

- **Creación del conjunto de datos:** consiste en la selección del conjunto de datos, o del subconjunto de variables o muestra de datos, sobre los cuales se va a realizar el descubrimiento.

- **Limpieza y preprocesamiento de los datos:** se compone de las operaciones, tales como: recolección de la información necesaria sobre la cual se va a realizar el proceso, decidir las estrategias sobre la forma en que se van a manejar los campos de los datos no disponibles, estimación del tiempo de la información y sus posibles cambios.

- **Reducción de los datos y proyección:** encontrar las características más significativas para representar los datos, dependiendo del objetivo del proceso. En este paso se pueden utilizar métodos de transformación para reducir el número efectivo de variables a ser consideradas o para encontrar otras representaciones de los datos.

- **Elegir la tarea de Minería de Datos:** decidir si el objetivo del proceso de KDD es: regresión lineal, clasificación, agrupamiento, etc.

- **Elección de algoritmos de Minería de Datos:** selección de métodos a ser utilizados para buscar los patrones en los datos. Incluye además la decisión sobre cuáles modelos y parámetros pueden ser los más apropiados.

- **Minería de Datos:** consiste en la búsqueda de los patrones de interés en una determinada forma de representación o sobre un conjunto de representaciones, utilizando para ello métodos de clasificación, reglas o árboles, regresión lineal, agrupación, etc.

- **Interpretación de los patrones encontrados:** dependiendo de los resultados, a veces se hace necesario regresar a uno de los pasos anteriores.

- **Consolidación del conocimiento descubierto:** consiste en la incorporación de este conocimiento al funcionamiento del sistema, o simplemente documentación e información a las partes interesadas.

Así, los sistemas de KDD permiten la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento; consolidar el conocimiento resolviendo posibles conflictos con conocimiento previamente extraído y hacer el conocimiento disponible para su uso. Además, este proceso puede involucrar varias iteraciones y puede contener ciclos entre dos cualesquiera de los pasos. La mayoría de los trabajos hechos con esta metodología se centran en la fase de Minería de Datos. Sin embargo, los otros pasos se consideran importantes para el éxito del mismo, en la figura

(1.2) se muestra el esfuerzo que requiere cada fase del proceso de KDD:

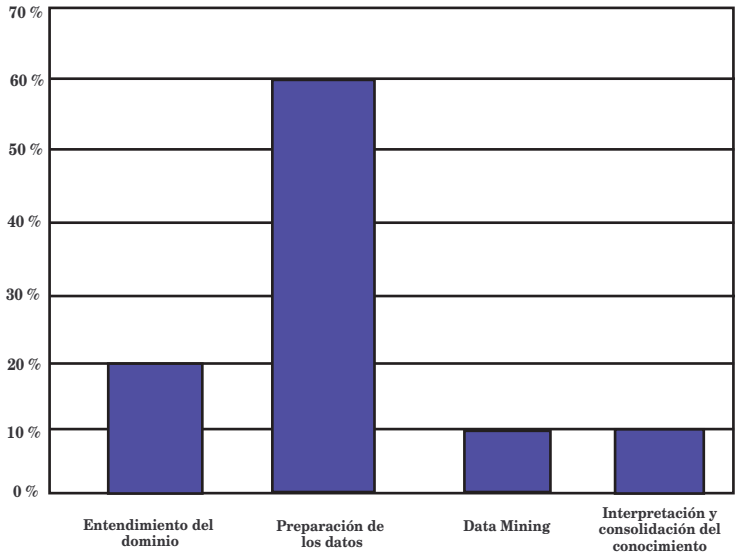

Figura 1.2: Esfuerzo requerido para cada fase del KDD.

Como se observa gran parte del esfuerzo del KDD recae sobre la fase de preparación de los datos, fase crucial para tener éxito en la interpretación y que se expondrá con más detalle en el Capítulo 2.

1.3 Minería de Datos (MD)

Las definiciones dadas anteriormente establecen una visión clara de la relación que existe entre el KDD y la Minería de Datos: el primero es el proceso global de descubrir conocimiento útil desde la base de datos mientras que la Minería de Datos se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtención de patrones y modelos.

La Minería de Datos se define como *el proceso de planteamiento de distintas consultas y extracción de información útil, patrones y tendencias previamente desconocida desde grandes cantidades de datos posiblemente almacenados en bases de datos.*

Considerando que el proceso de Minería de Datos convierte datos en conocimiento, si nos refe-

rimos al contexto anterior del KDD, en el que estuvimos preparando unos datos para aplicar una técnica de minería de datos, la perspectiva anterior se resume en la siguiente figura:

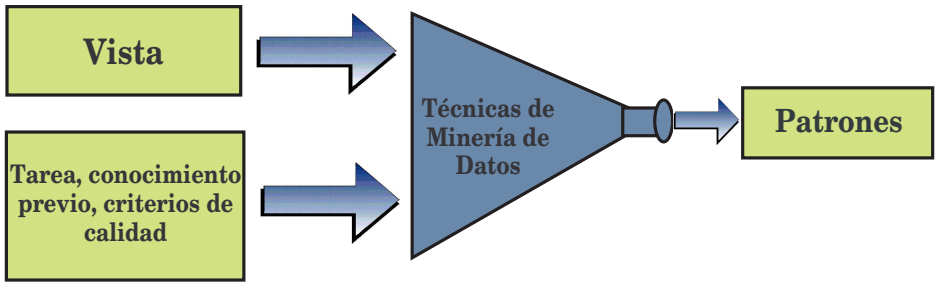

Figura 1.3: Proceso ideal de minería de datos.

En la figura (1.3) las técnicas de Minería de Datos aparecen como un colador donde al introducirse los datos produce, sin complicaciones, unos patrones. Aunque la figura ilustra muy bien y de manera simplista el caso, las cosas no resultan tan triviales. En general los procesos que extraen patrones son computacionalmente muy costosos, más aún cuando los patrones requeridos son expresivos, novedosos, comprensibles y útiles.

Para este objetivo, no trivial, se requieren tareas y técnicas. Una tarea de Minería de Datos es *un tipo de problema de minería de datos.* Por otra parte, las técnicas de Minería de Datos constituyen *el enfoque conceptual para extraer la información de los datos, y, en general son implementadas por varios algoritmos.* Es importante establecer la diferencia entre una **tarea** y una **técnica**. Pasemos a ver, en primer lugar, las tareas y, posteriormente, las técnicas más importantes en este contexto.

1.3.1 Tareas de la Minería de Datos

Los tipos de tareas más importantes son: clasificación, regresión, agrupamiento y reglas de asociación, acompañadas de ejemplos.

Antes de definir las tareas es preciso definir el conjunto de ejemplos con los que se va a trabajar. Definamos E como el conjunto de todos los posibles elementos de entrada. Las instancias posibles dentro de E generalmente se representan como un conjunto de valores para una serie de atributos

(nominales o numéricos). Es decir, $\mathrm{E} = \mathrm{A}_1 \times \mathrm{A}_2 \times, \cdots, \times \mathrm{A}_n$ donde A_i con $i = 1, \ldots, n$ son los atributos y un ejemplo e es una tupla $\langle a_1, a_2, \ldots, a_n \rangle$ tal que $a_i \in A_i$.

Veamos a continuación las tareas más importantes en la Minería de Datos.

- **Predictivas:** se tratan de problemas en los que hay que predecir uno o más valores para uno o más ejemplos. Los ejemplos van acompañados de una salida (clase, categoría o valor numérico) o un orden entre ellos. Dependiendo de como sea la correspondencia entre los ejemplos y los valores de salida y la presentación de los ejemplos se pueden definir varias tareas predictivas: Clasificación, Clasificación Suave, Estimación de Probabilidad de Clasificación, Categorización, Preferencias o Priorización y Regresión. En el desarrollo del capítulo 2 se trabajarán solamente las tareas de Clasificación y Clasificación Suave que a continuación detallamos.

  - **Clasificación:** los ejemplos se presentan como un conjunto de pares de elementos de dos conjuntos , $\delta = \{< e, s >: e \in \mathrm{E}, s \in \mathrm{S}\}$, donde S es el conjunto de valores de salida. Los ejemplos e, al ir acompañados de un valor de S, se denominan ejemplos etiquetados $< e, s >$ y, en consecuencia, δ se denomina conjunto de datos etiquetado. El objetivo es aprender una función $\lambda : \mathrm{E} \rightarrow \mathrm{S}$, denominada clasificador, que represente la correspondencia existente en los ejemplos, es decir, para cada valor de E tenemos un único valor para S. Además, S es nominal, es decir, puede tomar un conjunto de valores $c_1, c_2, ..., c_m$, denominados clases. La función aprendida será capaz de determinar la clase para cada nuevo ejemplo sin etiquetar, es decir, dará un valor de S para cada valor de e. Por ejemplo: clasificar un mensaje de correo electrónico como *spam* o no, clasificar entre varios medicamentos cuál es el mejor para una determinada patología, etc.

  - **Clasificación suave:** la presentación del problema es semejante a la de clasificación, pares de elementos en δ. Además de la función $\lambda : \mathrm{E} \rightarrow \mathrm{S}$ se aprende otra función $\theta : \mathrm{E} \rightarrow \mathbb{R}$ que significa el grado de certeza de la predicción hecha por la función λ. Es importante tener un clasificador suave que acompañe a las predicciones de una medida de certeza de dichas predicciones aunque sea una pequeña estimación. Por ejemplo: clasificar un mensaje de correo electrónico como *spam* o no, proporcionando, además, la certeza de la clasificación,

clasificar los medicamentos como en la tarea anterior pero dando el grado de certeza, etc.

- **Descriptivas:** aquí los ejemplos se presentan como un conjunto $\delta^{'} = \{e : e \in \mathrm{E}\}$, sin etiquetar ni ordenar de ninguna manera. El objetivo no es predecir nuevos datos sino describir los existentes. Esto puede hacerse de muchas maneras y la variedad de tareas se incrementa notablemente. Algunas de las técnicas usadas son: detección de valores anómalos, clustering o agrupamiento, correlaciones y factorizaciones, dependencias funcionales y reglas de asociación . En el desarrollo del capítulo 2 se contemplan las tareas de correlaciones y detección de valores anómalos, para profundizar sobre los demás tipos de tareas consultar [5].

  - **Correlaciones y factorizaciones:** los estudios de correlaciones y factorizaciones se centran exclusivamente en los atributos numéricos. El objetivo es ver, dados los ejemplos del conjunto $\mathrm{E} = \mathrm{A}_1 \times \mathrm{A}_2 \times, \cdots, \times \mathrm{A}_n$, si dos o más atributos numéricos A_i *y* A_j están correlacionados linelmente o relacionados de algún otro modo. Este tipo de relaciones son bidireccionales o no orientadas. Para ver si existe algún tipo de orientación (causa/efecto) se pueden utilizar modelos de regresión restringidos a sólo esos dos atributos, no para predecir un valor a partir de otro, sino para ver la dependencia de los dos valores, analizando, por ejemplo, el coeficiente de la regresión lineal obtenida.

  - **Detección de valores e instancias anómalas:** el objetivo de la dectección de atípicos es muchas veces la limpieza de los datos. No obstante, la detección de valores anómalos puede ser muy útil para detectar precisamente comportamientos anómalos, que pueden sugerir fraudes, fallos, intrusos o comportamientos diferenciados. La definición de instancia anómala es más general en el sentido que no sólo considera un único atributo, sino que los considera todos. La tarea se define con el objetivo de encontrar aquellas intancias que no son similares a ninguna (o muy pocas) de las otras instancias. La manera de abordar el problema es agrupar los ejemplos y ver aquellas instancias que se quedan " desplazadas" de los grupos mayoritarios. Para ello son especialmente útiles los agrupadores suaves o los estimadores de probabilidad de agrupamiento, ya que si un ejemplo tiene baja probabilidad de agrupamiento con todos los grupos se pueden considerar un caso " aislado" y, por tanto, anómalo.También se utilizan otros métodos no necesariamente basados en la

tarea de agrupamiento, como la medición de distancias (aquellas cuyo vecino más próximo esté muy lejos puede considerarse, en cierto modo, una instancia anómala). Ejemplo: encontrar, de las compras realizadas con tarjeta, aquellas que sean anómalas.

Como hemos visto, algunas tareas están ralacionadas, esto ha hecho que la terminología de algunas de ellas sea bastante diversa y, a veces, sea difícil aclararse con los nombres que utilizan algunas herramientas o textos. Por ejemplo, se suele utilizar el término " aprendizaje supervisado " para los métodos predictivos y el término " aprendizaje no supervisado " para los métodos no predictivos.

1.3.2 Técnicas de la Minería de Datos

Cada una de las tareas anteriores requiere de métodos, técnicas o algoritmos para resolverlas. Como veremos en la sección siguiente una misma técnica puede resolver un abanico de tareas, sin embargo, es importante señalar la relación que existe entre las tareas y técnicas. [15] clasifica las tareas en *supervisadas* o *predictivas* y *no supervisadas* o *descriptivas*, además establece una relación de correspondencia entre ellas que podemos observar en la figura (1.4).

A continuación se mencionan brevemente algunas de las técnicas existentes para llevar a cabo las tareas anteriores. La relación establecida será para dar una reseña de la variedad de las mismas. Más adelante, en este capítulo, se hará énfasis en las técnicas que se usarán para los capítulos siguientes. Los tipos de técnicas usadas en la minería de datos son las siguientes:

- **Técnicas algebraicas y estadísticas:** este tipo de técnica tiene su fundamento en la expresión de modelos y patrones mediante fórmulas algebraicas, funciones lineales y no-lineales, distribuciones y valores agregados estadísticos como media, varianza y correlaciones. Por lo general, cuando estas técnicas obtienen un patrón, lo hacen a partir de un modelo del cual se estiman unos coeficientes o parámetros, de ahí el nombre de técnicas *paramétricas*. Algunos de los algoritmos más conocidos dentro de este grupo de técnicas son la regresión lineal, la regresión logarítmica y la regresión logística. Los discriminantes lineales y no-lineales, basados en funciones predefinidas, es decir, determinantes paramétricos, entran dentro de esta categoría. Por otra parte, aunque el término " no paramétrico " se utiliza para englobar gran parte

de técnicas provenientes del aprendizaje automático, como son las redes de neuronas, también existen muchas técnicas de modelización de estadística *no paramétrica*, estimación por núcleos de la función de densidad, por ejemplo.

- **Técnicas basadas en conteos de frecuencias e histogramas bivariados:** el objetivo de esta es contar la frecuencia en la que dos o más sucesos se presentan conjuntamente, además de observar las variaciones de los estadísticos de orden en dos variables.

- **Técnicas basadas en análisis factorial y de componentes principales:** consiste en reducir la dimensionalidad por transformación cambiando los ejes de proyección de los datos obteniendo atributos independientes entre si.

La correspondencia entre las tareas y técnicas es muy variada. Esta variedad es una de las razones por las cuales es necesario conocer las capacidades de cada técnica, los ámbitos donde suelen funcionar mejor, la eficiencia, la robustez, etc. Es importante resaltar que la razón por la cual se pueden resolver ciertas tareas con las mismas técnicas tiene que ver con que las mismas se centran alrededor de la idea del " aprendizaje inductivo "; pueden considerarse presentaciones diferentes del mismo proceso.

El aprendizaje inductivo es *la eliminación de redundancia, vista como comprensión de información*, [10]. Visto de esta manera, el aprendizaje nos permite identificar irregularidades de un conjunto de observaciones. Estas irregularidades corresponden a redundancias representadas por patrones o modelos que los compriman o los definan. Estos patrones pueden ser utilizados para predecir observaciones futuras o explicar observaciones pasadas. Esta capacidad de predecir y explicar el entorno es fundamental para mejorar el comportamiento.

En los siguientes capítulos plantearemos problemas donde las tareas a resolver serán de clasificación y predicción, definiremos formalmente dos técnicas de la minería de datos que nos permitirán resolverlas: Análisis de Componentes Principales y los Histogramas Bivariados.

Figura 1.4: Técnicas y tareas de la Minería de Datos.

1.4 Análisis de Componentes Principales (ACP)

Un problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad: si es posible describir con precisión los valores de p variables por un pequeño subconjunto de $k < p$ de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información.

El Análisis de Componentes Principales tiene este objetivo: dadas n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales. La presente sección está contemplada en dos resumidas partes. En la primera se definen las CP, su determinación, estandarización y pesos en términos de los autovalores de la matriz de correlación. La segunda parte, explica brevemente la importancia del Teorema de Descomposición Espectral en el análisis de componentes principales. El contenido de esta sección del capítulo se seguirá como en [7] y [13].

1.4.1 Definición de las Componentes Principales (CP)

Sea $X \in \mathbb{R}^p$, un vector de p variables aleatorias con segundo momento finito, denotamos con $X^{'}$ el vector traspuesto de X. La **esperanza** del vector X, $E(X)$, es el vector cuyo k-ésimo elemento es la esperanza de la variable X_k, la k-ésima entrada de X. La varianza del vector X es definida como:

$$Var(X) = E[(X - E(X))^{'}(X - E(X))]$$

nótese que $Var(X) = tr(\Sigma)$, donde Σ es la matriz de covarianza de X, definida por:

$$\Sigma = E[(X - E(X))(X - E(X))^{'}].$$

Supongamos que $E(X) = 0$. Entonces la matriz de covarianza de X es $\Sigma = E(XX^{'})$, el (i,j)-ésimo elemento de ésta matriz es la covarianza de X_i y X_j cuando $i \neq j$ y la varianza de X_j cuando $i = j$. En este capítulo supondremos que la matriz Σ es positiva definida, para el caso invertible.

Supongamos que las variables aleatorias $X_1, X_2, \ldots, X_p$ poseen una distribución p - dimensional cualquiera, con vector de medias μ y matriz de covarianzas Σ como se definió anteriormente, que tiene como raíces características: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

De esta distribución se extrae una muestra de n observaciones independientes $x_1, x_2, \ldots, x_n$, representada por la matriz de datos $X = (x_{ij})$, para $i = 1, \ldots, n$; y $j = 1, \ldots, p$ donde:

$$x_i^{'} = (x_{i1}, \ldots, x_{ip})$$

es el vector fila i-ésimo de la matriz X, es decir, la observación multivariada i-ésima ($i = 1, \ldots, n$), expresado en forma matricial:

$$\mathbb{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & & & & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1^{'} \\ x_2^{'} \\ \cdots \\ x_j^{'} \\ \cdots \\ x_n^{'} \end{bmatrix}.$$

Con estos datos se calcula el vector de medias muestrales

$$\overline{x} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_n \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^{n} x_{i1}}{n} \\ \vdots \\ \frac{\sum_{i=1}^{n} x_{ip}}{n} \end{bmatrix}.$$

La matriz de covarianzas $\Sigma = s_{jk}$ donde

$$s_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)}{n-1}$$

y la matriz de correlaciones muestrales:

$$\mathbf{R} = (r_{jk}).$$

donde:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}.$$

Las CP muestrales son combinaciones lineales de las variables originales X_j, que no estan correlacionadas entre sí y que tienen una varianza muestral máxima.

Para precisar, consideremos las p combinaciones lineales siguientes, cada una evaluada sobre los n individuos de la muestra:

$$\begin{aligned}
y_1 &= a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p = a_1'x \\
y_2 &= a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p = a_2'x \\
&\vdots \\
y_i &= a_{1i}x_1 + a_{2i}x_2 + \cdots + a_{pi}x_p = a_i'x \\
&\vdots \\
y_p &= a_{1p}x_1 + a_{2p}x_2 + \cdots + a_{pp}x_p = a_{ap}'x,
\end{aligned}$$

donde:

$$a_i' = (a_{1i}, a_{2i}, \ldots, a_{pi})$$

es un vector fila de constantes $(i = 1, \ldots, n)$ y

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

es un vector columna que representa los valores de las variables X_j en un individuo cualquiera.

Se cumple que:

$$Var(y_i) = a_i'\Sigma a_i \quad i = 1, \ldots, p.$$

y

$$Cov(y_i, y_k) = a_i'\Sigma a_k \quad i, k = 1, \ldots, p.$$

Las CP muestrales son combinaciones lineales $y_1, y_2, \ldots, y_p$ como las anteriores, pero que tienen vectores de coeficientes a_i' muy particulares, tales que:

- $y_1, \ldots, y_p$ no están correlacionados entre sí.

- $Var(y_1)$, $Var(y_2), \ldots, Var(y_p)$ alcanzan los valores más grandes posibles.

- Los vectores coeficientes $a_i^{'}$ son tales que $a_i^{'}a_i = 1$, $i = 1, \ldots, p$.

La primera CP muestral y_1 maximiza su varianza muestral $Var(y_1) = a_1^{'}\Sigma a_1$ sujeta a la condición de que $a_1^{'}a_1 = 1$. La segunda CP y_2 maximiza $Var(y_2) = a_2^{'}\Sigma a_2$ sujeta a $a_2^{'}a_2 = 1$ y covarianza muestral $Cov(y_1, y_2) = 0$. La tercera CP y_3 maximiza $Var(y_3) = a_3^{'}\Sigma a_3$ sujeta a $a_3^{'}a_3 = 1$, $Cov(y_1, y_3) = 0$ y $Cov(y_2, y_3) = 0$. En general, la i-ésima CP maximiza $Var(y_i) = a_i^{'}\Sigma a_i$ sujeta a $a_i^{'}a_i = 1$ y $Cov(y_k, y_i) = 0$ para $k < i$.

1.4.2 Determinación de las Componentes Principales

Consideremos, por el momento, que el vector de variables aleatorias X tiene matriz de covarianza conocida, Σ. Si Σ tiene raíces características $\lambda_1, \lambda_2, \ldots, \lambda_p$ y vectores (columna) característicos normalizados correspondientes $V_1, V_2, \cdots, V_p$ con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, las CP se obtienen haciendo $a_i^{'} = V_i^{'}$, es decir, la CP muestral i-ésima y_i está dada por:

$$y_i = V_i X = v_{1i}x_1 + v_{2i}x_2 + \cdots + v_{ki}x_k + \cdots + v_{pi}x_p, \quad i = 1, \ldots, p,$$

donde

$$V_1^{'} = (v_{1i}, v_{2i}, \ldots, v_{pi}),$$

es el vector característico asociado a la raíz característica λ_i y además:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}.$$

Se tiene que:

- $Var(y_i) = \lambda_i, \quad i = 1, \ldots, p$

- $$\sum_{i=1}^{p} Var(y_i) = \sum_{i=1}^{p} Var(x_i),$$

es decir,

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = s_{11} + s_{22} + \cdots + s_{pp}.$$

Frecuentemente las primeras k CP con $k < p$, explican una alta proporción (80 % a 90%) de la variabilidad total de las X_i. Estas CP pueden, por lo tanto, representar a las p variables originales sin pérdida importante de información.

La contribución o importancia de la k-ésima variable a la i-ésima CP, se mide por el elemento $v_{ki}^{'}$ del vector normalizado $V_i^{'}$ asociado a la raíz característica λ_i de Σ, es decir que:

$$r_{y_i,x_k} = \frac{v_{ki}\sqrt{\lambda_i}}{\sqrt{s_{kk}}}, \quad i, k = 1, \ldots, p.$$

Nótese que v_{ki} es proporcional al coeficiente de correlación entre la i-ésima CP y_i y x_k. A este coeficiente de correlación se le llama también ***carga - componente*** de la variable k-ésima sobre la CP i-ésima. Estas cargas facilitan la interpretación de las CP, porque como coeficientes de correlación toman en cuenta las diferencias en las variazas de las variables originales y son más confiables que los coeficientes componentes v_{ki} que están influidos por diferencias en las escalas de medición o en el

rango de las variables.

1.4.3 Variables estandarizadas

En algunas situaciones, las variables de interés x_i se miden en escalas con amplitudes o unidades diferentes, lo que produce varianzas muy distintas en las x_i. Esto influye en la composición de las CP extraídas, de manera que las variables con varianzas muy grandes tendrán mayor peso en las primeras CP. Para evitar esta situación, los datos deben estandarizarse transformando los valores de las variables originales $x_1, x_2, \ldots, x_p$ en nuevas variables $z_1, z_2, \ldots, z_p$ donde:

$$z_i = \frac{x_i - \overline{x}_i}{\sqrt{s_{ii}}}; \quad i = 1, \ldots, p.$$

Obviamente, cada z_i tiene promedio igual a 0 y varianza igual a 1. La matriz de covarianzas de las z_i es entonces igual a la matriz de correlación $\mathbf{R}$ de las x_i, y las CP de las z_i se obtienen de los vectores característicos de $\mathbf{R}$.

Representamos con w_i a las CP obtenidas de las z_i, donde:

$$w_i = \hat{u}_{1i} z_1 + \hat{u}_{2i} z_2 + \cdots + \hat{u}_{pi} z_p = \hat{\mathbf{u}}_1^{'} \mathbf{z}.$$

Basados en resultados anteriores, los $\hat{u}_1^{'}$ son los vectores característicos normalizados asociados a las raíces características γ_i de la matriz de correlaciones $\mathbf{R}$. Los vectores y raíces característicos derivados de $\mathbf{R}$, en general, no son iguales a los derivados de la matriz Σ de covarianzas.

Se cumple que:

- $$\sum_{i=1}^{p} Var(w_i) = \sum_{i=1}^{p} Var(z_i) = p.$$

- Proporción de la varianza total de las z_i debida a la k-ésima CP

$$r_{w_i,z_k} = \hat{u}_{ki}\sqrt{\gamma_i} = \frac{\gamma_k}{p} \quad i,k = 1,\ldots,p$$

$$\sum_{k=1}^{p} r_{w_i^2,z_k} = \gamma_i \quad i = 1,\ldots,p.$$

1.4.4 Determinación del número de CP

Al utilizar el análisis de CP para fines de reducción o simplificación de los datos, surge la necesidad de decidir sobre el número de CP que se deben retener. Existen varios métodos para llegar a este número; la escogencia de uno u otro depende del juicio del investigador, y también si se usa Σ o $\mathbf{R}$ como punto de partida. Los criterios analíticos que examinaremos para determinar el número de componentes a retener son los siguientes: criterio de la ***media aritmética*** y gráfico de ***sedimentación*** (en caso de usar Σ), criterio del ***contraste de las raíces características no retenidas*** y ***porcentaje acumulado de varianza total*** (en caso de usar $\mathbf{R}$).

Criterio de la media aritmética

Este criterio selecciona aquellas componentes cuya raíz caracaterística λ_j excede de la media de las raíces características. Recordemos que la raíz característica de una CP es precisamente su varianza. Analíticamente, este criterio implica retener todas aquellas componentes en que se verifique que

$$\lambda_j > \overline{\lambda} = \frac{\sum_{i=1}^{p} \lambda_i}{p}. \tag{1.1}$$

Si se utilizan las variables tipificadas, entonces se verifica que

$$\sum_{i=1}^{p} \lambda_i = p.$$

Por lo tanto, el primer criterio (1.1) cuando se aplica a las variables tipificadas, se puede expresar así:

Criterio de la media aritmética para variables tipificadas. Se seleccionan aquellas componentes para las cuales

$$\lambda_k > 1.$$

Contraste sobre las raíces características no retenidas

Se puede considerar que las $p - r$ últimas raíces características poblacionales son iguales a 0. Si las raíces muestrales que observamos correspondientes a estas componentes no son exactamente igual a 0, se debe a los problemas de azar. Por ello, bajo el supuesto de que las variables originales siguen una distribución normal multivariante, se puede formular la siguiente hipótesis nula relativa a las raíces características poblacionales:

$$H_0 : \lambda_{m+1} = \lambda_{m+2} = \cdots = \lambda_p = 0.$$

El estadístico para contrastar esta hipótesis es

$$\mathcal{Q}^* = \left\{ n - \frac{2p+11}{6} \right\} \left\{ (p-r)\ln(\overline{\lambda}_{p-r}) - \sum_{j=r+1}^{p} \ln(\lambda_j) \right\}.$$

Bajo la hipótesis nula anterior, este estadístico se distribuye como una Chi-cuadrado con $(p - r + 2)(p - r + 1)/2$ grados de libertad. Para ver la mecánica de la aplicación de este contraste, supongamos que inicialmente se han retenido r raíces características (por ejemplo las que superan la unidad) al aplicar el criterio en (1.1). En el caso cuando se rechace la hipótesis nula, se tendrá que una o más de las raíces características no retenidas es significativa. La decisión a tomar en ese caso sería la de retener una nueva componente y aplicar el contraste a las restantes raíces características. Este proceso continuaría hasta que no se rechace la hipótesis nula.

El gráfico de sedimenteción

El gráfico de sedimentación se obtiene al representar en el eje de las ordenadas las raíces características y en el eje de las abcisas el número de la componente en orden decreciente. Uniendo todos los puntos se obtiene una figura que, en general, se parece al perfil de una montaña con una pendiente fuerte hasta llegar a la base, formada por una meseta con una ligera inclinación. Continuando con el símil de la montaña, en esa meseta es donde se acumulan los guijarros caídos desde la cumbre, es

decir, donde se sedimentan. De acuerdo con el criterio del gráfico se retienen todas aquellas componentes previas a la zona de sedimentación. Es decir, se ubica un punto o " codo ", donde la curva descendente se convierta en una recta descendente, entonces se retendrán el número de CP igual a la abcisa donde comienza el codo. Este método tiene la desventaja de que no siempre existe un codo y, a veces, puede haber más de uno. La aplicación de este criterio se verá en el Capítulo 2 y allí se podrá observar como luce este gráfico.

Porcentaje acumulado de varianza total

Emplear como criterio el porcentaje acumulado de varianza total explicado por las CP supone retener $m < p$ CP si:

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i}.$$

alcanza un valor " grande ", digamos entre $0,80$ y $0,90$.

1.4.5 Pesos de las CP

En un análisis de CP el investigador por lo general calcula el valor que toma cada CP en cada individuo. A estos valores se les denomina ***pesos de las CP***.

Si las CP se extraen de Σ, los pesos de las CP para el j-ésimo individuo ($j = 1, \ldots, n$) son:

$$\begin{aligned} y_{j1} = &\quad V_1' x_j \\ y_{j2} = &\quad V_2' x_j \\ \vdots &\quad \vdots \\ y_{jm} &\quad V_m' x_j, \end{aligned}$$

donde x_j es el vector columna de observaciones para el j-ésimo individuo, y m es el número de CP que se consideró apropiado para representar parsimoniosamente el conjunto de datos.

Si los individuos se extraen de $\mathbf{R}$, los pesos de las CP en el j-ésimo individuo serían

$$\begin{aligned} w_{j1} = &\quad \hat{u}_1^{'} z_j \\ w_{j2} = &\quad \hat{u}_2^{'} z_j \\ \vdots &\quad \vdots \\ w_{jm} &\quad \hat{u}_m^{'} z_j, \end{aligned}$$

donde z_j es el vector columna de los valores estandarizados de las variables X para el j-ésimo individuo.

Se acostumbra a estandarizar los pesos asociados a cada CP. Si se emplea la matriz $\mathbf{R}$, los pesos estandarizados , digamos w^*, para la CP j-ésima se pueden obtener dividiendo los pesos no estandarizados entre $\sqrt{\gamma_i}$ (ya que la varianza de w_i es γ_i), o dividiendo el vector $\hat{u}_1^{'}$ entre $\sqrt{\gamma_i}$ y multiplicando por z_j. Para el individuo j, su peso en la CP 1 es:

$$w_{ji}^* = \frac{w_{j1}}{\sqrt{\gamma_1}} = \frac{\hat{u}_1^{'}}{\sqrt{\gamma_1}} z_j = \frac{vector\, de\, cargas\, CP\, 1}{\gamma_1} z_j,$$

estos pesos los calculan los paquetes estadísticos como MatLab, R o S-Plus.

1.4.6 El Teorema de la Descomposición Espectral

El TCE es la herramienta básica sobre la cual descansa el ACP, lo incluímos acá a efectos de hacer autocontenido nuestro trabajo, sin hacer las pincipales demostraciones. El lector interesado puede ver [6].

El teorema de la descomposición espectral del álgebra lineal establece que cualquier matriz simétrica real A de orden $p \times p$ se puede escribir como

$$A = P \Lambda P^{'},$$

donde Λ es una matriz diagonal de raíces características de A, y P es una matriz ortogonal cuyas columnas son los vectores característicos normalizados asociados a las raíces características de A.

Si se aplica este teorema a la matriz de covarianzas Σ, esta se puede factorizar como:

$$\Sigma = (\hat{P}\hat{\Lambda}^{\frac{1}{2}})(\hat{\Lambda}^{\frac{1}{2}}\hat{P}') = \hat{L}\hat{L}',$$

donde

$$\hat{\Lambda}^{\frac{1}{2}} = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p}),$$

λ_i raíz característica i-ésima de Σ

$$\hat{L} = \hat{P}\hat{\Lambda}^{\frac{1}{2}},$$

donde $\hat{P} = (V_1, \ldots, V_p)$ = matriz con los vectores característicos (columna) de Σ.

Definamos $y = \hat{P}^t x$. Entonces:

$$Var(y) = \hat{P}'\Sigma\hat{P} = \hat{P}'(\hat{P}\hat{\Lambda}\hat{P}')\hat{P} = \hat{\Lambda},$$

es decir, las variables y_i tienen varianzas λ_i, y cero covarianzas, es decir, están incorrelacionadas. Las y_i son entonces, las CP de Σ. Se observa entonces que el ACP es quivalente a una factorización de Σ como el producto de una matriz $\hat{L}$ y su traspuesta.

También, Σ se puede escribir como:

$$\Sigma = \lambda_1 V_1 V_1' + \cdots + \lambda_p V_p V_p'.$$

Conforme se extraen sucesivamente las CP de Σ, las matrices $\lambda_j V_j V_j'$ y su suma acumulada, pueden utilizarse para evaluar la aproximación de Σ mediante un número de variables componentes menor que el número original de variables.

1.5 Histogramas Bivariados

El histograma bivariado es una técnica de la MD que consiste en una tabla de contingencia que manifiesta las variaciones y correlaciones de los estadísticos de orden de dos variables. La idea subyacente al histograma de este estilo es poder hacer una representación a través de una tabla que exprese la relación funcional o estadística de dos variables asociadas a un fenómeno. [3].

1.5.1 Construcción

La construcción del histograma bivariado deberá constituir una serie de pasos de manera sistemática que permita generar una tabla que resuma los valores producidos por las variaciones de determinada característica, representando la frecuencia con que se presentan distintas categorías o clases dentro de dicho conjunto.

En general, dados dos vectores $I = (I_1, \ldots, I_n)$ de clases, y $H = (H_1, \ldots, H_m)$ de variaciones de determinada característica, se quiere construir una tabla de orden $(n \times m)$ que contenga el conteo de elementos de la forma (I_i, H_j) con $i = 1, \ldots, n$ y $j = 1, \ldots, m$ para cada casilla. Considerando que cada elemento debe cumplir con la condición de la clase I_i y pertenecer a la variación H_j.

En la siguiente tabla se observa la apariencia de un histograma bivariado con las características anteriores:

	H_1	H_2	$\ldots$	H_j	$\ldots$	H_m
I_1				$\vdots$		
I_2						
$\vdots$				$\vdots$		
I_i	$\ldots$		$\ldots$	Nro. de (I_i, H_j)	$\ldots$	
$\vdots$						
I_n						

Tabla 1.1: Histograma Bivariado.

En el siguiente diagrama de flujo se aprecia el proceso de construcción de un histograma bivariado mediante un algoritmo elaborado en MatLab:

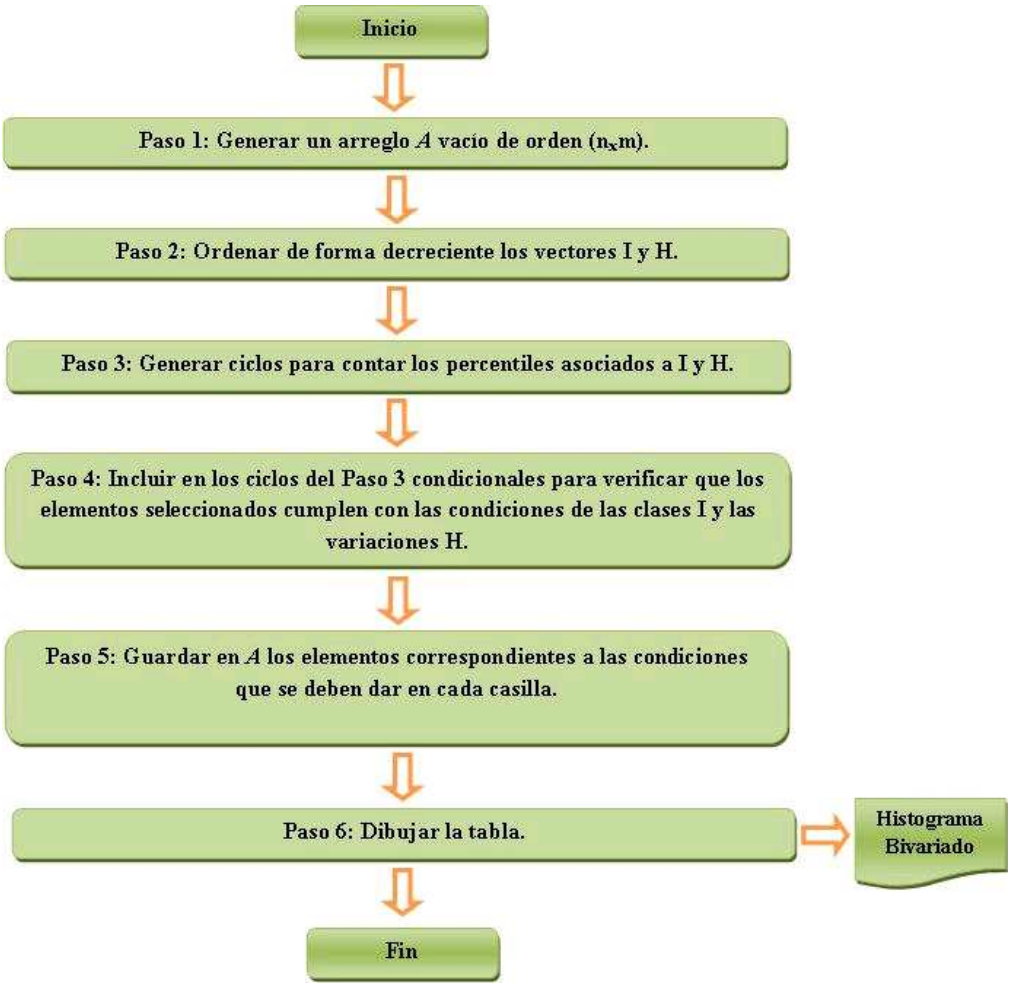

Figura 1.5: Diagrama de flujo para la construcción de un Histograma Bivariado.

Capítulo 2

Aplicaciones del KDD & MD

2.1 Introducción

En este capítulo vamos a aplicar las técnicas de KDD y MD sobre una criba o fusión horizontal (filas/registros) como vertical (columnas/atributos) que será nuestra matriz de datos, conformada por registros de tráfico de voz y datos por horas del día durante cinco meses del año 2007 para una compañía de telefonía móvil venezolana.

Comenzaremos haciendo una exploración elemental de las estructura de los datos con técnicas de estadística descriptiva. Seguiremos con la proyección de los datos para considerar únicamente aquellas variables o atributos que van a ser relevantes para ir afinado las técnicas que nos permitirán postular los modelos correspondientes y las restricciones que encontramos.

La clasificación constituye la revisión del sumario estadístico de cada factor para hacer el respectivo análisis de los histogramas y box plots para detectar alguna posible distribución de los factores y definir la importancia de los mismos, finalmente la transformación de los datos dependerá de la tarea a realizar en la Minería de Datos. Durante la redacción final de este TEG supimos de un software especilizado en DM llamado WEKA que no consideramos aquí.

2.2 Preparación de los datos

La base de datos a manejar está presentada en forma matricial donde las columnas son factores y las filas son registros (individuos) de dichos factores. Los datos fueron extraídos a través de medidores de tráfico de llamadas denominados radiobases (MTX), que tienen la función de procesar la información mensual del tráfico de datos de telefonía celular en una región determinada, distribuidos a lo largo de Venezuela.

Los registros de tráfico de voz y datos están hechos durante las 24 horas del día y se cuenta con una base de datos para los meses desde abril hasta agosto del año 2007, es importante mencionar que dichos registros se encuentran procesados en una misma unidad, que se usa en telefonía como una medida estadística del volumen del tráfico denominada ***Erlang***. El ***Erlang*** será una medida muy importante ya que sus intensidades nos permitirán establecer el diagnóstico de los patrones encontrados con la MD.

De acuerdo a [9] el tráfico de un Erlang corresponde a un recurso utilizado de forma continua, o dos canales utilizados al 50 %, y así sucesivamente, de manera proporcional. Un Erlang puede ser considerado como un multiplicador de utilización por unidad de tiempo, así un uso del 100 % corresponde a 1 Erlang, una utilización de 200 % son 2 Erlangs, y así sucesivamente. Por ejemplo, si el uso total del celular en un área por hora es de 180 minutos, esto representa $180/60 = 3$ Erlangs. En general, si la tasa de llamadas entrantes es de λ por unidad de tiempo y la duración media de una llamada es h, entonces el tráfico T en Erlangs es:

$$T = \lambda \times h.$$

Esto puede ser usado para determinar si un sistema está sobredimensionado o se queda corto (tiene demasiados o muy pocos recursos asignados). Para calcular el tráfico en Erlang existen diferentes fórmulas como por ejemplo: Erlang B, Erlang C y Engset, para mayor información sobre el funcionamiento de estas fórmulas revisar Apéndice A.

Para el estudio se requiere establecer las variables o factores que van a definir el tráfico de voz y datos, y son las siguientes:

- **Tráfico primario CDMA de voz (TP.3G.VOZ):** es el registro primario recibido por las radiobases (MTX) de llamadas telefónicas hechas a través del móvil.

- **Tráfico primario CDMA de datos (TP.3G.DATA):** es el registro primario recibido por las radiobases (MTX) de mensajería de texto, multimedia y conexión a internet.

- **Tráfico secundario CDMA de voz (TS.3G.VOZ):** es el registro compartido por las radiobases (MTX) de llamadas telefónicas hechas a través del móvil.

- **Tráfico secundario CDMA de datos (TS.3G.DATA):** es el registro compartido por las radiobases (MTX) de mensajería de texto, multimedia y conexión a internet.

La matriz estudiada contiene estos 4 factores y 563.904 registros por horas del día para los meses antes mencionados. Esta información se puede dividir en dos conjuntos, tráfico primario y secundario. El tráfico primario de una MTX, es aquel que corresponde al flujo de llamadas originadas en ella misma, el secundario corresponde al flujo de llamadas originadas en otras MTX que se vieron forzadas a liberar la información debido a factores como saturación o movimiento del equipo móvil. Además se cuenta con columnas de esa matriz que contienen la fecha, la hora y la región del país donde se tomaron los registros. Cabe destacar que para el estudio sólo se seleccionó los factores más significativos relacionados al tráfico, pero existen estudios profundos sobre la estimación del mismo considerando la ubicación espacial en el trabajo Series Temporales, Kriging y sus aplicaciones (J. Yerena, [16]).

Una vez definidas las variables asociadas a la inmensa matriz de datos se procederá a la aplicación de la primera fase del KDD que tiene que ver con la selección, clasificación y transformación de los datos. Para ello, se necesita de técnicas de estadística descriptiva clásica como un sumario estadístico de la matriz completa, histogramas de densidad y diagramas de cajas y colas. A continuación, se muestran los estadísticos empíricos clásicos y los gráficos para todas las variables.

2.2.1 Sumario estadístico para las variables originales

Para evitar problemas de sesgo se hizo un sondeo observando la base de datos y se encontró la presencia de una gran cantidad de datos " nulos" o " nulos camuflados", valores que tienden a cero, y, que de alguna manera impiden el trabajo de minería, ya que pueden introducir un sesgo en el conocimiento extraído. La matriz original era de 563.904 registros, luego se aplicó un algoritmo simple en S-Plus (ver S-Plus, [11]) de recorrido en esta matriz que permitió filtrar los datos nulos, conservando tan solo el 28% de los datos quedando asi 157.428 registros. En la siguiente tabla se presenta el sumario estadístico que se hizo con la matriz filtrada.

	TP3GVOZ	TP3GDATA	TS3GVOZ	TS3GDATA
Mínimo	**1**	**1**	**0**	**0**
1er Cuartil	39.46	1.27	18	0.52
Media	75.49	2	47.45	1
Mediana	64.46	2	37	0.8
3er Cuartil	102.39	2.37	69.3	1.35
Máximo	**510.10**	**13**	**357.37**	**9.6**
Varianza	**2351**	**1.13**	**1409.40**	**0.6**
Stdv	48.48	1	37.54	0.78
Sesgo	**1.14**	**2**	**1**	**2**
Curtosis	**2**	**5.55**	**2**	**6**

Tabla 2.1: Sumario estadístico para las variables originales.

En la tabla anterior, que resume los estadísticos empíricos más relevantes, tenemos información importante. En primer lugar, podemos observar que en una distribución simétrica de los datos los valores de la mediana y la media identifican al mismo punto, en este sentido, las variables TP.3G.DATA, y TS.3G.DATA coinciden en esta medida pero la presencia de una gran cola de atípicos genera un comportamiento asimétrico. Por otra parte, la varianza para estas tres variables es pequeña, lo que señala poca dispersión de los datos alrededor de la media, sin embargo, los coeficientes de sesgo y curtosis bastante alejados del 0 y 3 respectivamente terminarán generando una asimetría que se corrobora con los histogramas de frecuencia asociados a las variables.

En el caso de las variables asociadas a los datos de voz (TP.3G.VOZ y TS.3G.VOZ) observamos que la diferencia de valores entre la media y la mediana generará una distribución asimétrica de los datos y esto se puede explicar por los altos valores de las varianzas que como medida de dispersión

es suceptible a los valores atípicos, además, si observamos los valores de sesgo y curtosis se justifica de nuevo una distribución asimétrica de los datos.

En telecomunicaciones es más evidente la presencia de datos atípicos en aquellas variables asociadas al tráfico de voz, esto se puede explicar por los siguientes factores:

- Los usuarios realizan llamadas de larga duración.
- Se efectúan repiques entre usuarios o llamadas de poca duración.
- Uso extremo de un plan ofrecido por la compañía en telecomunicación.
- Caídas del sistema.
- Consumo en días extraordinarios (Feriados o de algún suceso importante).
- Errores con el registro o recolección de los datos.

Estas razones son las que posiblemente generen estos valores extremos que se reflejarán como atípicos en la muestra.

Finalmente, es importante resaltar que en este conjunto de variables el registro de tráfico es absorbido por dos de ellas TP.3G.VOZ y TS.3G.VOZ, esto se puede observar por los valores mínimos y máximos de la muestra, que en comparación con los de las variables de tipo DATA son extremadamente altos, esta situación tendrá consecuencias interesantes en el Análisis de Componentes Principales que se verán en la sección (2.5).

2.2.2 Histogramas de las variables originales

Los siguientes histogramas representan la distribución de los datos originales para cada variable, al observarlos podremos notar que son unimodales. La mayoría de los datos están concentrados alrededor de una media de valores muy pequeños. Observando la tabla (2.1) los valores para la media son mayores que los de la mediana lo que procura un histograma sesgado con cola hacia la derecha.

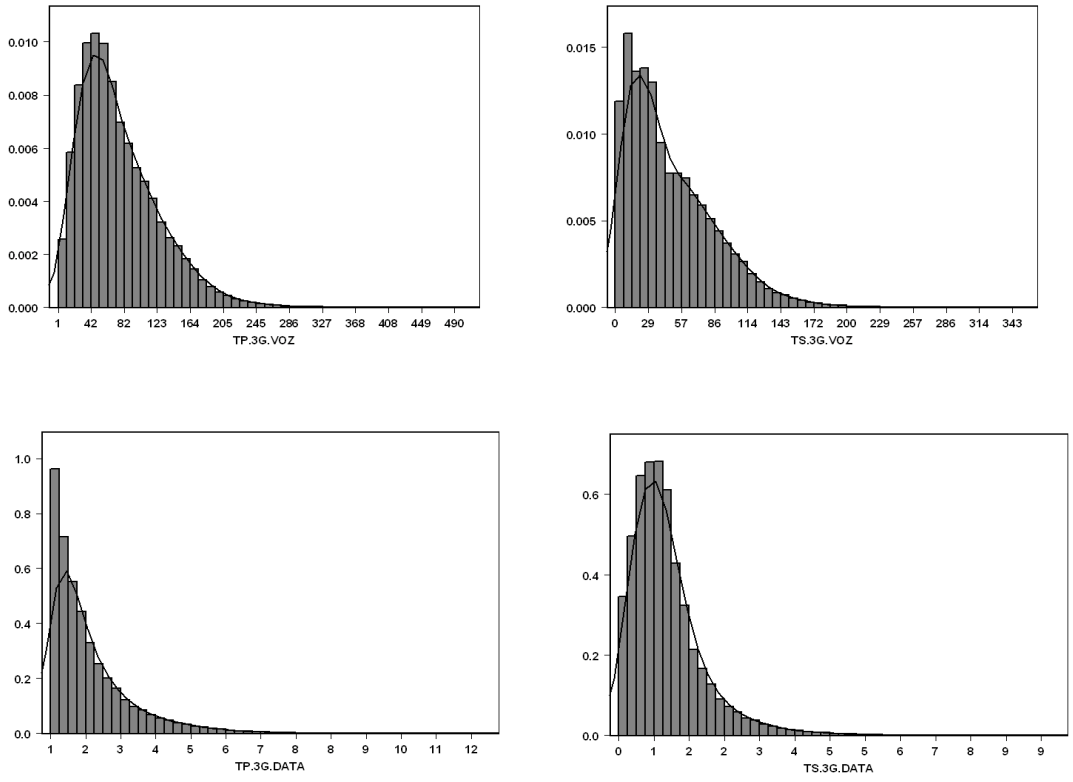

Figura 2.1: Histogramas de frecuencia para las variables originales.

Otro aspecto importante que resaltan estos histogramas es la ausencia de normalidad. La tabla (2.1) nos muestra unos valores de asimetría y curtosis alejados del 0 y 3 respectivamente que corroboran esto, además, los histogramas de frecuencia se ilustran muy alejados de la forma gaussiana con curtosis y asimetría positivas. Adicional se aplicó un test Kolmogorov-Smirnov que arrojó un p-valor igual a cero menor que el estadístico de la prueba, para rechazar la hipótesis nula de normalidad.

2.2.3 Diagramas de cajas y colas (Box Plots) para las variables originales

Los siguientes diagramas de caja determinan donde hay mayor o menor densidad de la muestra, además ilustran los posibles valores atípicos (outliers) que se puedan presentar.

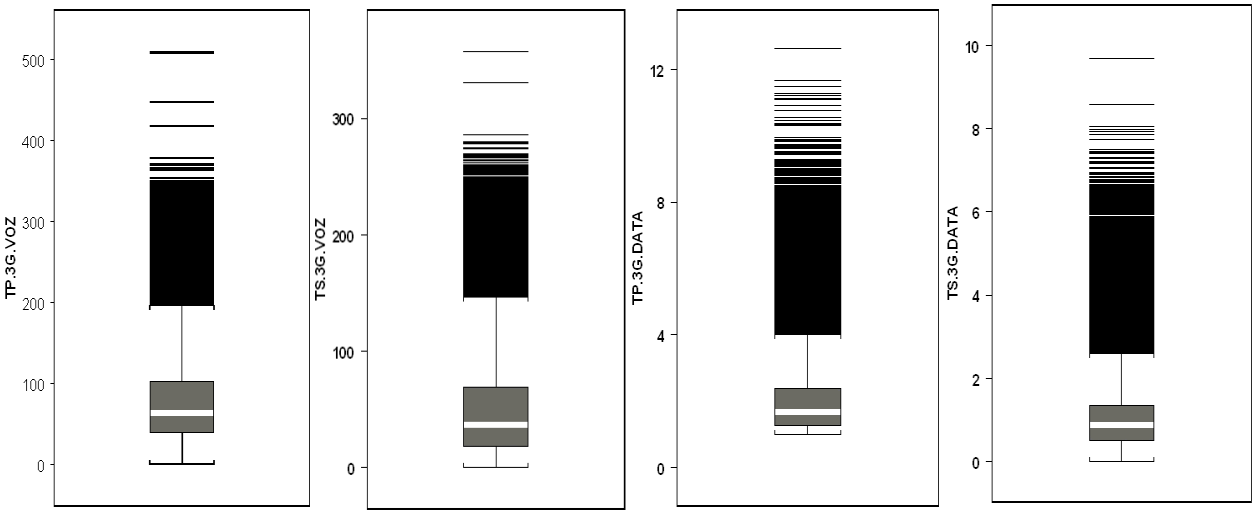

Figura 2.2: Boxplots para las variables originales.

Una manera de verificar si existen problemas de escala es evaluando simultáneamente los diagramas de caja de cada variable. Lo que se busca medir es si la diferencia entre los valores para la mediana (localizada en la linea central de la caja), el 1er cuartil (caja superior a la línea de la mediana) y 3er cuartil (caja inferior a la línea de la mediana) es notablemente significativa. En este sentido, se destacan dos grupos: el grupo de varibles de tipo DATA (TP.3G.DATA y TS.3G.DATA) y el grupo de variables de tipo VOZ (TP.3G.VOZ y TS.3G.VOZ), la diferencia entre los valores de la mediana para variables de un mismo grupo no es considerablemente alta, sin embargo, si comparamos variables de tipo DATA contra las de tipo VOZ las diferencias son muy marcadas. La parte sombreada en negro de los diagramas nos señala la presencia de datos extremos, por su gran cantidad pudiesen no ser atípicos sino datos fuera del rango de los cuartiles que definen ciertamente la asimetría en los histogramas de la figura (2.1), consideraremos estos datos “ delicados” para el análisis porque pudiesen representar algunos de los factores descritos en la sección (2.2.1).

2.3 Separación de los datos

Debido a la alta dimensionalidad la primera estrategia a emplear será crear submatrices de la matriz original. Dadas las limitaciones de S-Plus y MatLab para realizar los cálculos en la matriz principal se consideró implementar algoritmos sencillos que permitieran crear particiones por meses y días de los datos, a fin de garantizar una metodología y estudio más refinado (ver S-Plus [11]). Sin embargo, existen métodos y paquetes que manejan estas limitaciones: procesamiento paralelo, ORACLE y WEKA, [8].

Por otra parte, para pasar a la segunda fase del KDD es necesario hacer una transformación de la data. El Análisis de Componentes Principales será la primera herramienta usada por la Minería de Datos para abordar el problema de dimensión, en este sentido, se aplicó a los datos una ***estandarización***. Para ello se elaboró un algoritmo en MatLab que realizó esta instrucción para cada registro de todas las variables.

Una vez implementados los algoritmos y seleccionadas las submatrices tipificadas por meses se realizó un estudio exploratorio como el anterior para cada mes, acá sólo mostraremos los resultados obtenidos para el mes de abril que es una submatriz de 23.655 registros y 4 factores.

2.3.1 Sumario estadístico para las variables tipificadas

	TP3GVOZ	TP3GDATA	TS3GVOZ	TS3GDATA
Mínimo	-2	-1	-1	-1
1er Cuartil	-1	-1	-1	-1
Media	0	0	0	0
Mediana	0	0	0	0
3er Cuartil	0	0	1	0
Máximo	6	9	4	8
Varianza	1	1	1	1
Stdv	1	1	1	1
Sesgo	1	2	1	2
Curtosis	2	6	1	5

Tabla 2.2: Sumario estadístico para las variables tipificadas.

En esta tabla tenemos la estabilización para los datos. Por otra parte, los valores para el sesgo

y curtosis sufrieron un descenso poco significativo y se mantiene la distribución asimétrica con un resultado del test Kolmogorov-Smirnov de p-valor menor que el estadístico de la prueba para rechazar la hipótesis nula de normalidad de los datos.

2.3.2 Histogramas de las variables tipificadas

Los siguientes histogramas representan la distribución de los datos tipificados para cada variable del mes de abril, al observarlos podremos concluir como en los anteriores que son unimodales, con la diferencia de que los valores están estabilizados y la mayoría de los datos se centran con respecto a la media que esta vez es cero.

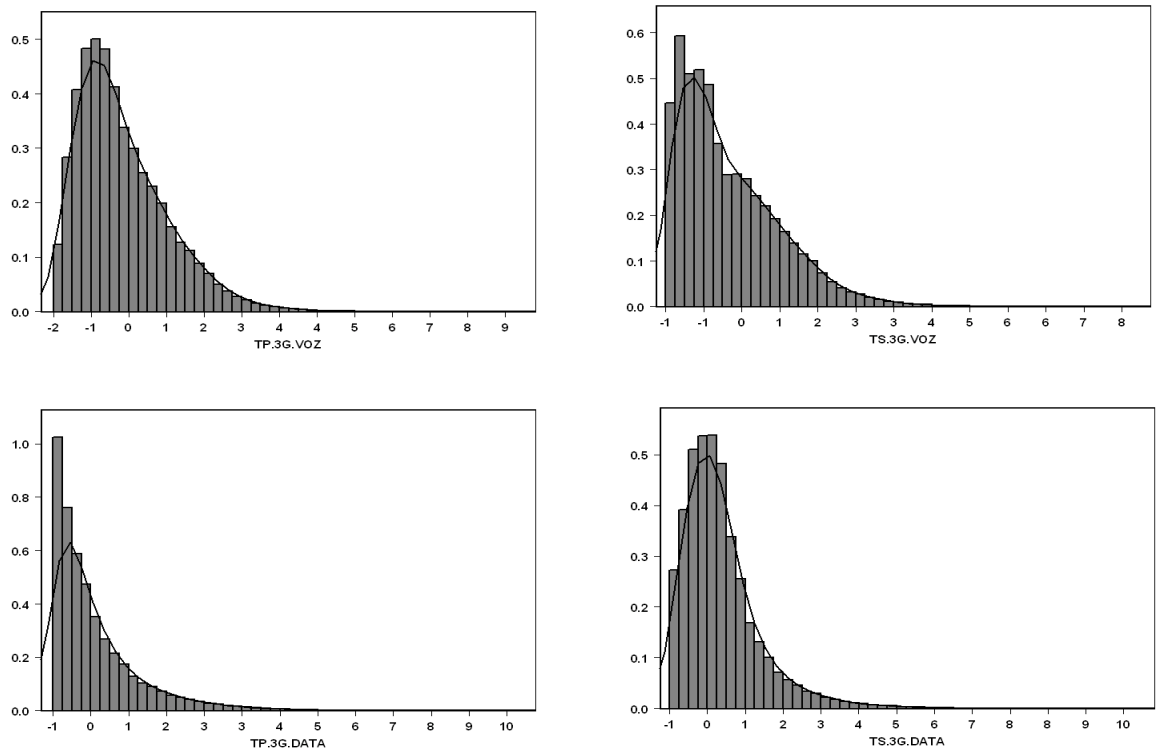

Figura 2.3: Histogramas de frecuencia para las variables tipificadas del mes de Abril.

En la tabla (2.2) los valores de sesgo y curtosis positivos se encuentran lejos del rango para una distribución normal explicando la ilustración de estos histogramas sesgados con cola hacia la derecha.

2.3.3 Diagramas de cajas y colas para las variables tipificadas

Los siguientes diagramas de cajas y colas determinan donde hay mayor o menor densidad de la muestra.

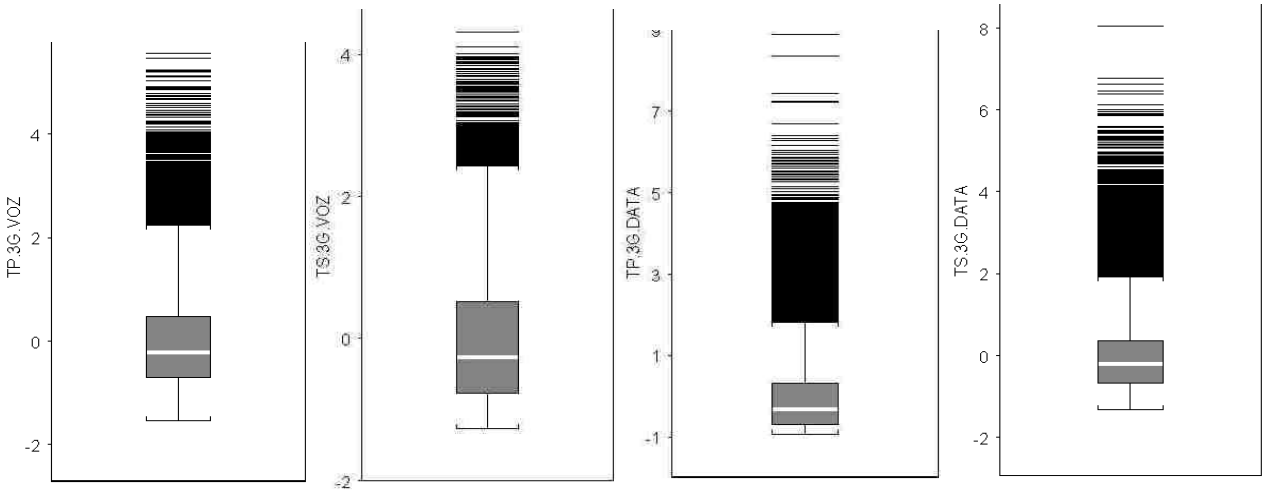

Figura 2.4: Boxplots para las variables tipificadas.

Tipificando los datos hemos resuelto el problema de escala ya que la diferencia entre la mediana, el 1er cuartil y el 3er cuartil no es tan significativa, se mantiene la forma de los diagramas pero cambian los valores. La parte sombreada en negro de los diagramas nos sigue señalando la presencia de una cola de datos que genera la asimetría y evitaremos eliminarlos para luego estudiar si juegan un papel importante en el proceso.

Una vez hecho los análisis para los gráficos, finalmente la fase de selección, clasificación y transformación nos asegura que nuestros datos están listos para la segunda fase del KDD denominada Minería de Datos.

2.4 Minería de Datos como fase del proceso de KDD

Anteriormente se estudió de manera descriptiva la base de datos y se consideró estar frente a un problema de dimensión, en este sentido, la técnica de minería de datos que nos garantiza la solución a esto es el Análisis de Componentes Principales (ACP).

El ACP será una técnica que nos permitirá, si es posible, describir con precisión los valores de p variables por un pequeño subconjunto de $m < p$ de ellas construidas como combinaciones lineales de las originales, si esto se logra, se habrá reducido la dimensión a costa de una pequeña pérdida de información.

El estudio de MD se realizará para verificar qué grupo de variables (VOZ o DATA) tiene mayor información incorporada, el análisis descriptivo indicó un contraste de tráfico bastante alto, además de la diferencia implicita que llevan ambos grupos por ser uno relacionado a mensajería multimedia, texto e internet y el otro de llamadas telefónicas, que se generan de maneras diferentes e independientes.

2.5 Aplicación de ACP para el grupo de variables por meses.

Debido a la alta dimensión de los datos, se realizó una separación por meses de los registros hechos para cada variable, por lo que la aplicación del ACP se realizará sobre muestras mensuales de la población de estudio. A continuación se presentan los resultados obtenidos para el mes de abril que corresponden a 4 variables de atributos en telecomunicaciones de data y voz por 23.655 registros obtenidos en tiempo; dichos resultados están basados en los autovalores de la matriz de covarianzas, estimada por el paquete S-PLUS a partir de la matriz tipificada.

2.5.1 Matriz de covarianzas de los datos

En la tabla (2.3) se refleja la matriz de covarianza de los datos, debido a la tipificación esta coincide con la matriz de correlación. En ella se puede apreciar que entre las variables TP.3G.VOZ y TS.3G.VOZ existe una alta correlación, con una unidad menos pero con la misma tendencia tenemos estos resultados para las variables TP.3G.DATA y TS.3G.DATA, esta alta correlación se debe a que las variables están en un mismo grupo. Lo interesante de este análisis será comprobar si se puede establecer cuál de las variables que están en contraposición es la que define el espacio de los datos, reduciendo así la dimensión del problema.

Variables	TP3GVOZ	TS3GVOZ	TP3GDATA	TS3GDATA
TP3GVOZ	1	**0,83**	0,56	0,48
TS3GVOZ	**0,83**	1	0,42	0,67
TP3GDATA	0,56	0,42	1	**0,71**
TS3GDATA	0,48	0,67	**0,71**	1

Tabla 2.3: Matriz de covarianzas de los datos.

Al calcular los vectores propios de la matriz de covarianza se obtienen direcciones que expresan la variabilidad de los datos, luego las CP serán aquellos vectores cuya dirección exprese la mayor variabilidad, es decir, las que incorporen la mayor información sobre los datos [13].

Variables	Comp 1	Comp 2	Comp 3	Comp 4
TP3GVOZ	**0,51**	-0,47	-0,48	0,54
TS3GVOZ	**0,52**	-0,49	0,34	-0,61
TP3GDATA	0,47	**0,61**	-0,52	-0,38
TS3GDATA	**0,5**	0,41	**0,62**	0,44

Tabla 2.4: Vectores propios de la matriz de covarianza.

La variabilidad de la nube de puntos entorno a la media empírica se mide considerando la suma de los autovalores asociados a los autovectores de la tabla (2.4). Esta suma representa la variabilidad de los datos, por eso aquel autovector cuyo valor propio asociado a él, contribuya en mayor grado a explicar esta variabilidad, será el que incorpore mayor información y definitivamente coincidirá con la CP.

2.5.2 Determinación del número de las CP

En la siguiente tabla tenemos los autovalores de las primeras 3 componentes principales, y sus respectivos porcentajes de explicabilidad (PE) o porcentajes de varianza y el porcentaje de explicabilidad acumulada (PEA) o porcentaje de varianza acumulada obtenidos a partir del ACP.

Aplicando el criterio de selección de la media aritmética visto en el Capítulo 1 para variables tipificadas, retendremos aquellas componentes para las cuales su autovalor sea mayor que 1, es decir, según la tabla (2.5) seleccionaríamos sólo la 1^a componente que explica el $71,16\%$ de la información original.

Nro. Componente	Autovalores	PE	PEA
1	2,85	71,16	71,16
2	0,71	17,64	88,81
3	0,40	9,97	98,79
4	0,05	1,20	100

Tabla 2.5: Valores para el estudio de la retención de componentes.

Para la CP 2 se observa en la figura (2.5) y la tabla (2.5) que aunque la CP 1 es mucho mayor en porcentaje comparado con el de la CP 2, entre ambas explican el 89% de la información original, por lo tanto a la hora de hacer la proyección de los datos en el espacio de las componentes junto con las variables se utilizarán estas dos.

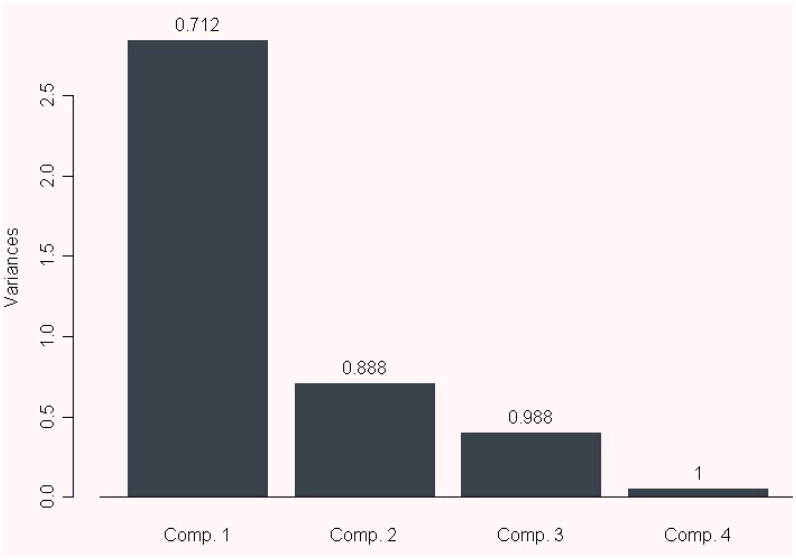

Figura 2.5: Importancia de las componentes principales.

2.5.3 Gráfico bidimensional de correlación

Para verificar la relación entre las variables originales y las CP se utilizará el gráfico bidimencional de correlación, que consiste en la proyección de las variables de estudio en un plano formado por las CP. La cercanía entre las proyecciones y los ejes de las componentes indicarán la relación existente entre las variables y las CP.

Observando la figura (2.6) podemos señalar dos aspectos: el primero es la estrecha relación que posee la CP 1 (Eje horizontal), con las variables TP.3G.VOZ y TS.3G.VOZ. Sin embargo, las variables TP.3G.DATA y TS.3G.DATA también tienen una correlación positiva con la CP 1 pero como el volúmen del tráfico que expresan es distinto, solo aquella que posea mayor flujo tendrá más correlación con la CP 1, en este caso las variables de tipo VOZ.

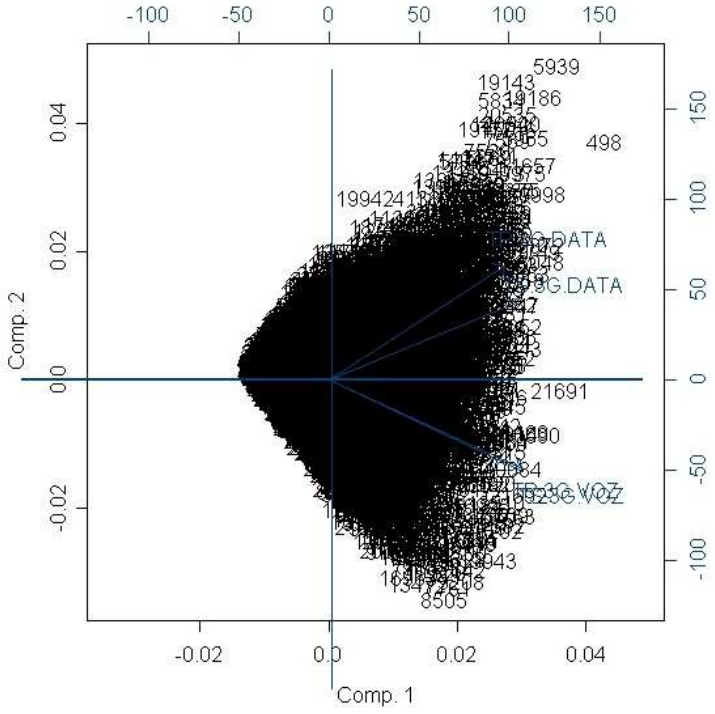

Figura 2.6: Proyección de las componentes principales para el mes de Abril.

Por otra parte, las variables de tipo VOZ se corresponden negativamente con la CP 2 (Eje vertical) y las de tipo DATA positivamente, lo que indica evidentemente que el comportamiento entre el flujo de llamadas y el flujo de datos, (multimedia, texto e internet), no necesariamente debe ser el mismo.

En la tabla (2.4) se aprecia también que las variables TP.3G.VOZ y TS.3G.VOZ presentan mayor correlación con el autovector de la CP 1 y la variable TP.3G.DATA con el de la CP 2, sin embargo la variable TS.3G.DATA posee una correlación de **0,62** en el autovector de la CP 3 que no está reflejada en el gráfico, pero podemos visualizar en la misma tabla que la correlación de esta variable con el

autovector asociado a la CP 1 es de **0,50** mostrando en el gráfico una cercanía de la proyección de dicha variable a la CP 1.

El análisis que ofrece este gráfico es sobre la relación contrapuesta de las variables de tipo VOZ y de tipo DATA. Por otra parte, la superposición de las variables de un mismo grupo indica que ellas están numéricamente correlacionadas, a pesar de que el tráfico primario se genera independientemente del secundario. Sin embargo, se aplicó un test Chi-Cuadrado para corroborar la dependencia de las variables TP.3G.VOZ y TS.3G.VOZ que son las de mayor flujo, y se obtuvieron los siguientes resultados para la matriz $A_{23655\times 2}$ que contiene los valores para las variables anteriores:

chisq.test(A)
Pearson's chi-square test without Yates' continuity correction.
data: A
X-square = 102208.8, df = 23654, p-value = 0

Tabla 2.6: Resultados para el test Chi-Cuadrado de las variables de tipo VOZ.

Considerando la hipóteis nula H_0: Los factores TP.3G.VOZ y TS.3G.VOZ son independientes, tenemos que la tabla (2.6) señala que el p valor es 0 entonces rechazamos H_0, es decir, este test indica claramente que hay asociación entre los factores anteriores.

Finalmente, si observamos la tabla (2.1) y el gráfico (2.6) el ACP sugiere reducir el problema de dimensión 4 a dimensión 2 considerando sólo las varibales TP.3G.VOZ y TS.3G.VOZ que poseen mayor volúmen de tráfico generando saturación de las radiobases MTX.

Es preciso resaltar que esta información resultó ser exactamente igual para cada uno de los meses representados por matrices de diferentes dimensiones de los datos originales. A continuación, se aplicará la técnica de Histogramas Bivariados para plantear solución al problema de saturación por tráfico de voz de las radiobases.

2.6 Histogramas Bivariados

Como se mencionó en el Capítulo 1, el histograma bivariado es una técnica de MD basada en conteos de frecuencia en una tabla de contingencia para chequear variaciones de los datos y encontrar patrones.

Después de la aplicación de la técnica ACP obtuvimos que las variables que definen el espacio de los datos son las relacionadas al tráfico de voz, debido a los altos registros de Erlang que estas proporcionan a las radiobases MTX en comparación con las de tipo DATA, por ello, se seleccionaron estas variables para aplicar la técnica a la cual hacemos referencia en esta sección.

Para el análisis se considerará el estudio de submatrices por mes, semanas y días del mes de abril. Los histogramas bivariados servirán para visualizar los patrones de consumo de tráfico de voz que realizan los usuarios, comportamiento o tendencias atípicas y para detectar las horas y días en las que el tráfico satura a las radiobases MTX.

2.6.1 Histograma Bivariado por mes

La interpretación de los histogramas bivariados viene dada por la intensidad de Erlangs (Eje vertical) y las horas de cada día de la semana (Eje horizontal). Las intensidades se representan en tres niveles: bajo, medio y alto, para cada día de la semana las 24 horas del día de todo el mes. La numeración descrita en el eje vertical del 0 al 100 indica la cantidad de registros encotrados en las matrices de Hora y Tráfico que coinciden con ese nivel de Erlang.

Si observamos la tabla (2.1) de la sección de sumario estadístico encontramos que para la variable TP.3G.VOZ el nivel bajo de Erlang se encuentra entre el mínimo y el primer cuartil, es decir, entre 1 y 39,46 Erlangs; el nivel medio de tráfico está entre 39,47 y 102,38 Erlangs; y el nivel alto entre el máximo y el tercer cuartil que corresponde a 510 y 102,39 Erlangs respectivamente. Para la variable TS.3G.VOZ tenemos que el nivel alto está entre 69,3 y 357,37; el nivel medio entre 18,1 y 69,2 y el nivel bajo entre 0 y 18 Erlangs.

A continuación se muestran las gráficas de los histogramas bivariados asociados a los datos del

mes de abril donde se establecen series estacionales con período cada 7 días de intensidad de Erlangs para el mes considerando días y horas de las variables anteriores.

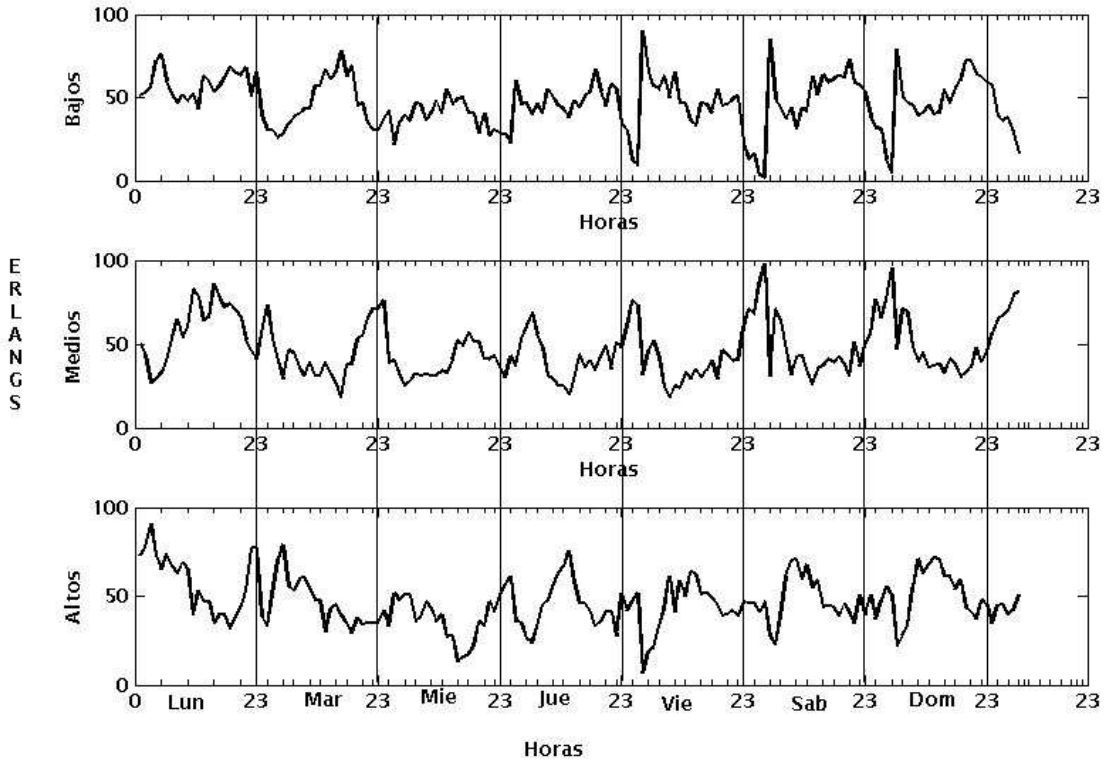

Figura 2.7: Histograma Bivariado para el mes de Abril de la variable TP.3G.VOZ.

Basados en lo anterior, las gráficas (2.7) y (2.8) nos brindan un panorama muy general de lo que sucede en el mes de abril con las variables TP.3G.VOZ y TS.3G.VOZ. En líneas generales, solo podemos apreciar que el comportamiento para las series de los días lunes, martes, miércoles y jueves es totalmente diferente. Sin embargo, para los días restantes se conserva una tendencia que se puede corroborar con la cantidad de registros encontrados para los tres niveles y las horas en que los mismos suceden, teniendo una división de la semana en dos grupos con diferentes comportamientos que no se puede apreciar con claridad.

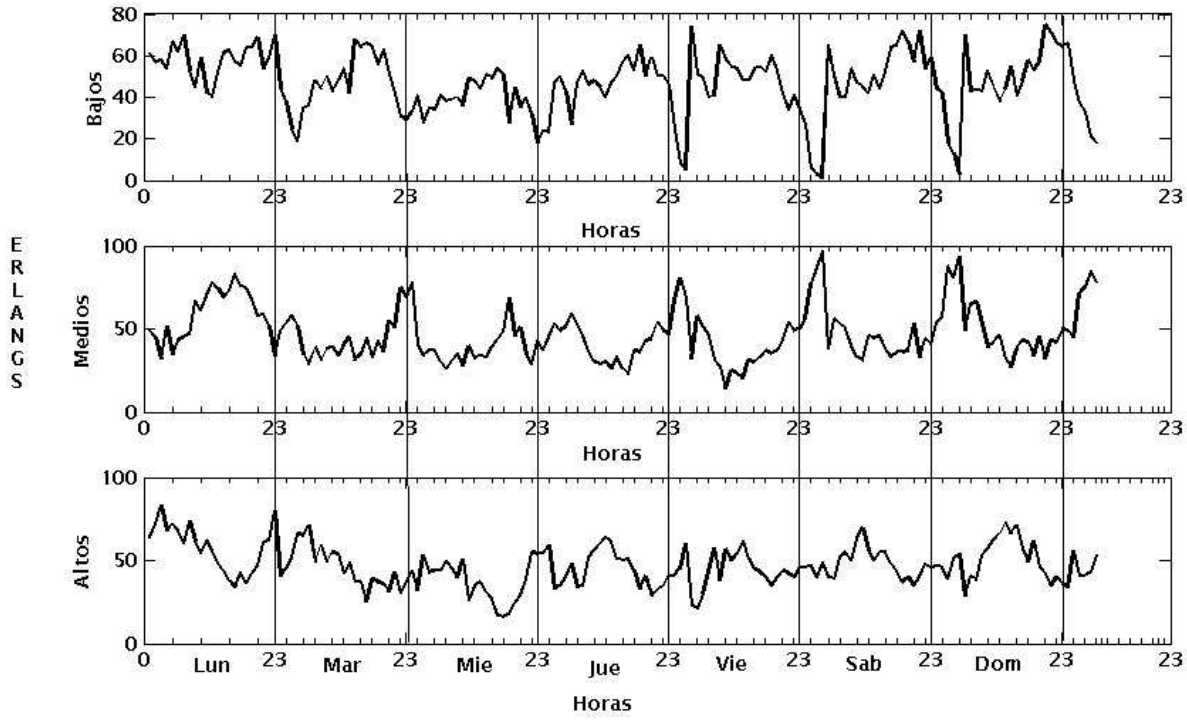

Figura 2.8: Histograma Bivariado para el mes de Abril de la variable TS.3G.VOZ.

Dados los resultados anteriores, este estudio sugiere una clasificación por semanas y días del mes para poder encontrar patrones que definan el comportamiento del tráfico y su incidencia en la saturación de las radiobases MTX. A continuación, se presentan los resultados de los histogramas para ambas variables durante las cuatro semanas del mes en estudio.

2.6.2 Histograma Bivariado por semanas y días

La lectura de estos histogramas será similar a la anterior, con la diferencia que se se mostrarán las series por semana del mes, ubicando, verticalmente en la retícula cada uno de los días de la semana que lo conforman. Además, se establecen bandas verticales punteadas sobre las horas del día para señalar los picos de tráfico según los niveles de Erlang.

Histogramas para la variable TP.3G.VOZ

En estos histogramas se tiene un panorama diferente, una visión microscópica de lo que sucede con el comportamiento del tráfico primario de voz por cada día. Haciendo una lectura de los gráficos (2.9), (2.10), (2.11) y (2.12) de manera vertical, observamos que los días lunes, martes y miércoles tienen un comportamiento similar, debido a que las horas de mayor flujo de tráfico para esos días se establecen entre la 1am y 8am, el menor flujo se da entre 8am y 4pm dejando un nivel medio al resto de las horas.

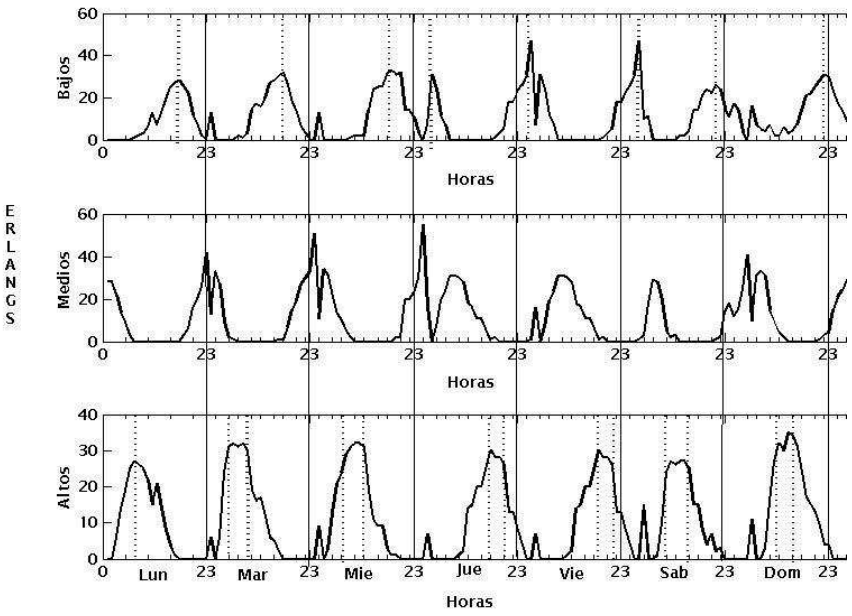

Figura 2.9: Histograma Bivariado 1era semana mes abril variable TP.3G.VOZ.

Por otra parte, el tráfico para los días jueves y viernes tiene características particulares. Si observamos los gráficos para cada semana notaremos que jueves y viernes poseen un flujo similar de tráfico, sin embargo, a medida que transcurren las semanas ese flujo se desplaza en diferentes horarios.

Para los días sábado y domingo el comportamiento es muy parecido, estableciéndose de 8am a 5pm un patrón de consumo alto de tráfico , observándose algunos picos de 2am a 3am.

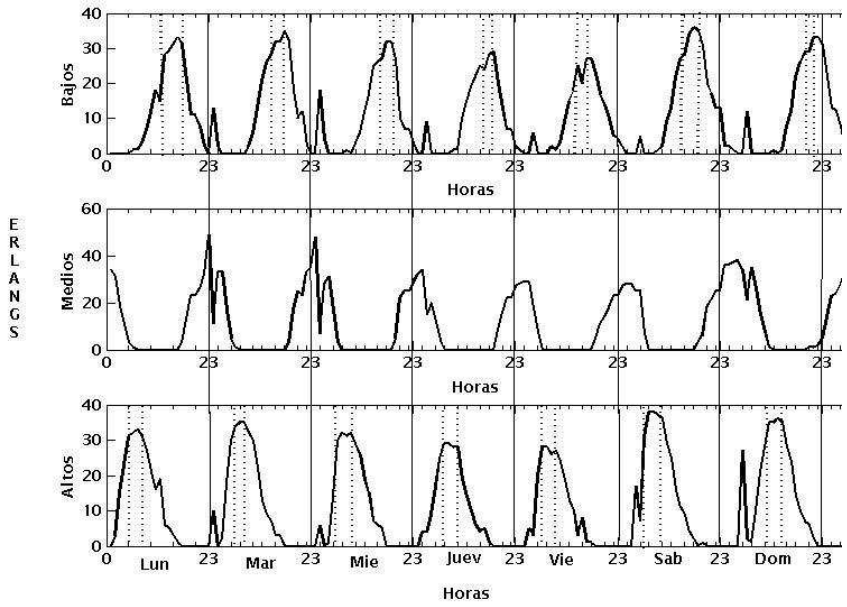

Figura 2.10: Histograma Bivariado 2da semana mes abril variable TP.3G.VOZ.

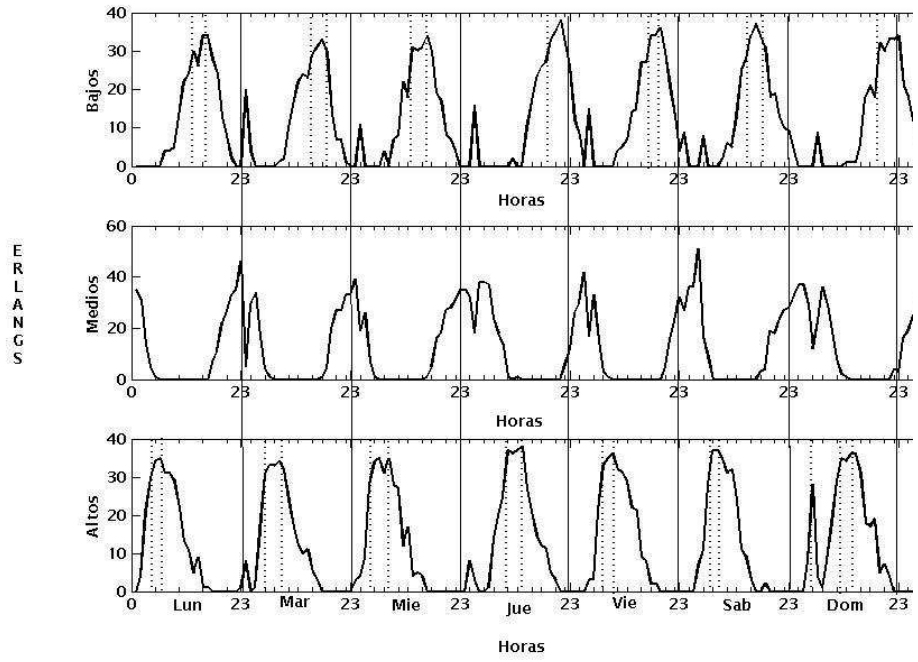

Figura 2.11: Histograma Bivariado 3ra semana mes abril variable TP.3G.VOZ.

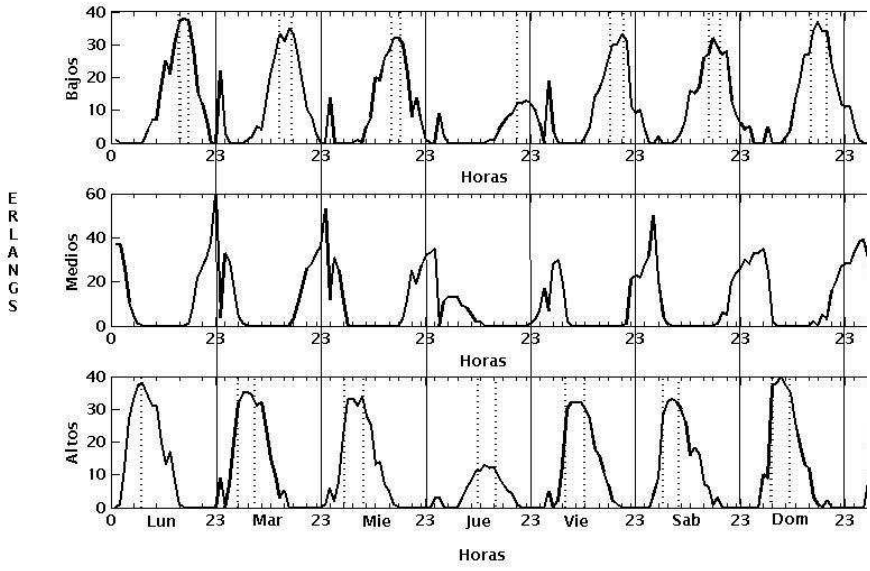

Figura 2.12: Histograma Bivariado 4ta semana mes abril variable TP.3G.VOZ.

En líneas generales, el uso de estos histogramas nos permiten decir que el mayor flujo de tráfico primario de voz (TP.3G.VOZ) se realiza en un horario promedio de 1am y 8am para los días lunes, martes y miércoles; para jueves y viernes oscilante y sábados y domingos hay tendencias desde las horas de la mañana, 8am apróximadamente, hasta horas de la tarde 3pm, apróximadamente. Estos patrones de horarios permiten saber las horas en las que una radiobase MTX puede sobredimensionarse o saturarse.

Histogramas para la variable TS.3G.VOZ

En esta oportunidad, mostraremos los histogramas por semanas y días de la variable TS.3G.VOZ que tiene que ver con el tráfico secundario de voz. En el apartado de ACP se obtuvo como resultado que el tráfico primario y secundario tenían una estrecha relación numérica a pesar de que se generaban de manera independiente. Es importante verificar si los patrones obtenidos con el tráfico primario tienen similitud con los del tráfico secundario, ya que ambas variables tienden a saturar las radiobases MTX.

A continuación se presentan los histogramas asociados a las cuatro semanas del mes de abril para

la variable TS.3G.VOZ:

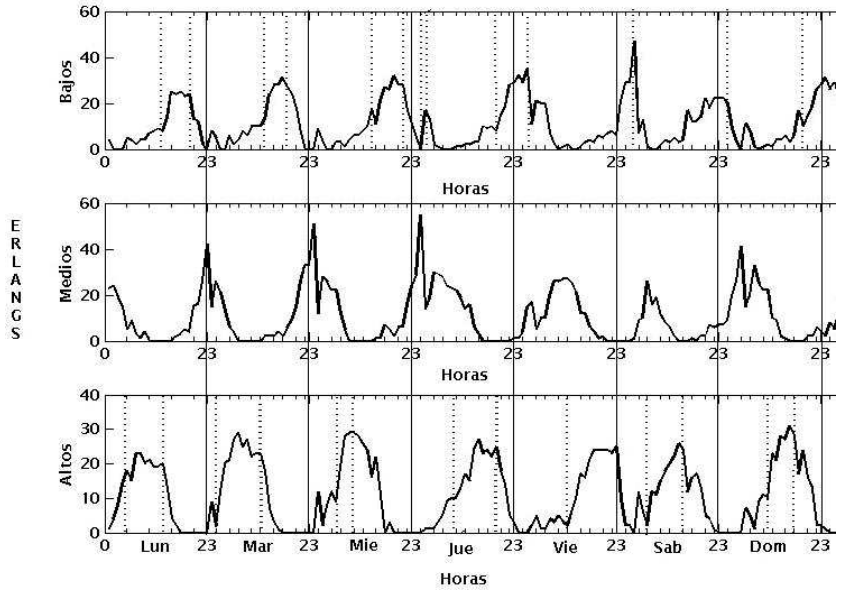

Figura 2.13: Histograma Bivariado 1era semana mes abril variable TS.3G.VOZ.

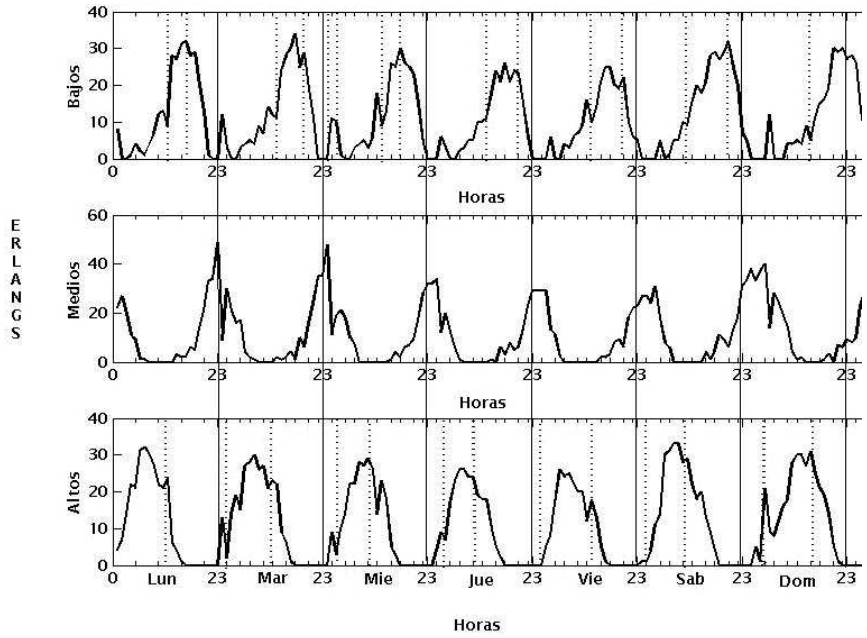

Figura 2.14: Histograma Bivariado 2da semana mes abril variable TS.3G.VOZ.

Estas gráficas evidencian el mismo patrón de días para ambas variables, sin embargo, establecen

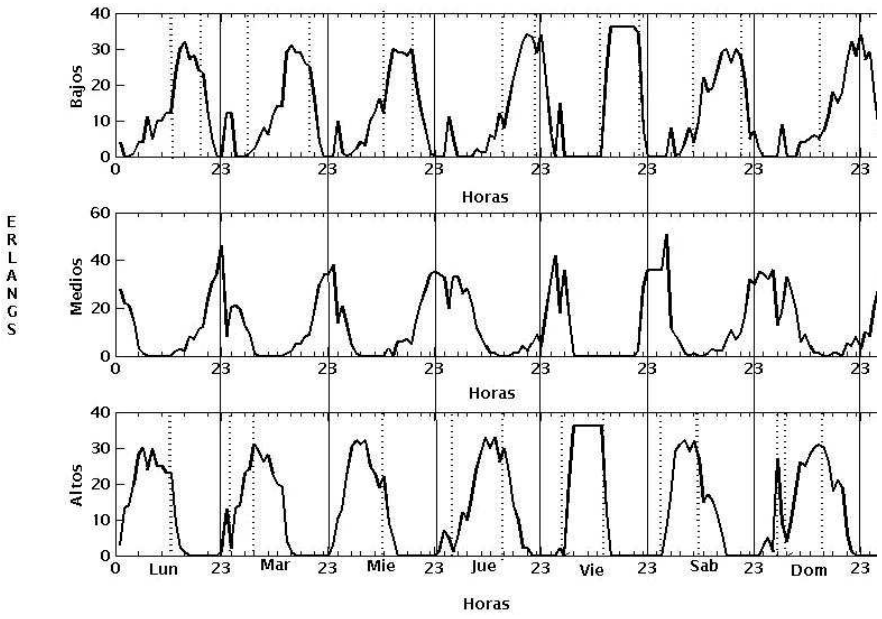

Figura 2.15: Histograma Bivariado 3ra semana mes abril variable TS.3G.VOZ.

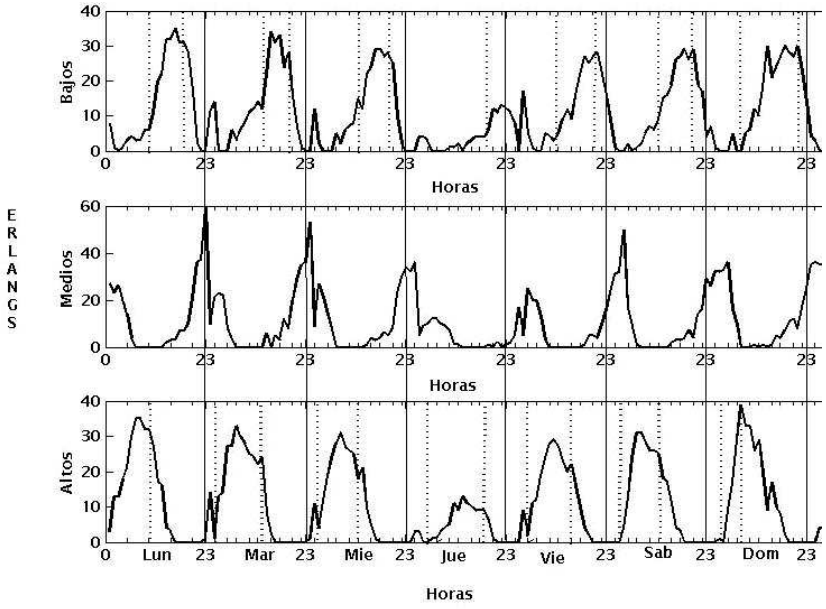

Figura 2.16: Histograma Bivariado 4ta semana mes abril variable TS.3G.VOZ.

algunas diferencias de horario. Los días lunes, martes y miércoles presentan flujo alto de tráfico entre la 1am y 12pm, un horario un poco más extendido que el de tráfico primario. Para los días jueves y viernes ya no oscila el horario sino que se establece un patrón entre las 7am y 12pm apróximadamente.

Finalmente, para los sábados y domingos la tendencia de consumo alto comienza unas horas antes que el primario, ubicándose entre 4am y 3pm.

Estos histogramas nos permiten establecer tres grupos durante la semana: el primero, de lunes a miércoles con horarios matutinos para el consumo alto, vespertino para el consumo bajo y nocturnos para el consumo medio. El segundo, jueves y viernes con horarios diferentes según el tipo de tráfico (primario o secundario). El tercero, sábados y domingos con horarios matutinos - vespertinos para el consumo alto, vespertinos - nocturnos para el consumo bajo, quedando los consumos medios para las horas de la madrugada.

Es importante resaltar, que los mismos resultados se obtuvieron para los meses mayo y junio. Sin embargo, las series asociadas a los meses julio y agosto tienen componente estacional distinta, generando modelos bastante complejos para la predicción.

Capítulo 3

Conclusiones y Recomendaciones

La aplicación de la MD, como fase del proceso de extracción de conocimiento sobre los datos de telefonía móvil, permitió desarrollar una metodología con la finalidad de obtener información relevante en el momento de tomar decisiones que involucren la calidad del servicio telefónico y la cobertura de la demanda del mismo.

En el Capítulo 2 se evidencia la aplicación de la metodología del KDD, obteniendo en su primera fase información muy importante, ya que a través de los sumarios apreciamos los estadísticos básicos de la muestra: media, varianza y coeficientes de curtosis y sesgo; que en conjunto con los histogramas y box-plots nos permitieron concluir la ausencia de normalidad de los datos.

Por otra parte, se garantizó que el mayor flujo de tráfico era generado por las variables de tipo VOZ debido a los valores máximos y mínimos expresados en la tabla de sumario. Con las gráficas de box-plot se detectó un problema de escala que luego se estabilizó con la estandarización de los datos para aplicar la segunda fase del proceso denominada Minería de Datos.

El problema a resolver en esta segunda fase fué la dimensión, para ello, se aplicó la técnica ACP. Con el ACP se obtuvo una reducción de dimensión 4 a 2 con una pérdida del 11 % de la información, considerando las variables de tipo VOZ como aquellas que incorporaban mayor información, por estar ubicadas en la dirección de máxima variabilidad y con mayor flujo de tráfico.

Otro resultado importante arrojado por el ACP fue la evidencia de contraposición entre las varia-

bles de tipo DATA y VOZ, afirmándose que el flujo de llamadas y de datos no necesariamente tiene que ser el mismo. Además, se corroboró la alta correlación numérica entre variables de un mismo grupo aunque las mismas se generen de forma independiente.

Una vez seleccionadas las variables de tipo VOZ se aplicó la técnica de Histogramas Bivariados, y se obtuvieron patrones de consumo en los usuarios que sirvieron para establecer bloques de días y horas de la semana donde se encontraban los consumos altos, medios y bajos de tráfico. Este resultado es muy importante ya que permite garantizar que las radiobases MTX tienden a saturarse por tráfico de voz de lunes a miércoles en horas de la mañana, jueves y viernes con horarios alternantes pero más hacia las horas de la tarde y sábados y domingos con horario matutino tendiendo hacia las horas del mediodía.

En conclusión, la metodología del KDD y las técnicas de MD utilizadas en este trabajo para el estudio de datos de telecomunicación ayudó a encontrar patrones útiles en la extracción de conocimiento para la toma de decisiones en las políticas varias de una compañía telefónica, además de poder colaborar en el desarrollo de la plataforma tecnológica de la misma, como por ejemplo, la asignación de recursos en caso de que una radiobase se encuentre saturada para evitar la pérdidad de información en el sistema.

Se recomienda hacer la estimacíon de la distribución de los datos para simulación de los mismos, asi como también la formulación de modelos autorregresivos funcionales para las predicciones basadas en clasificaciones previas y la utilización de herramientas computacionales para el manejo de bases de datos. También sería interesante poder hacer una fusión entre este trabajo y Series Temporales, Kriging y sus Aplicaciones (Yerena, [16]) para establecer un componente importante como lo es la estimación del volúmen del tráfico considerando ubicación espacial.

Apéndice A

Fórmulas para calcular el tráfico

El tráfico medido en Erlangs es usado para calcular el nivel de servicio. Hay diferentes fórmulas para calcular el tráfico, entre ellas: Erlang B, Erlang C y la fórmula de Engset. Esto será discutido a continuación, y cada una de ellas puede ser derivada como un caso espacial de procesos de tiempo contínuo de Markov conocido como *Birth-Death Process*. (Parkinson, [9])

A.1 Fórmula Erlang B

En telecomunicaciones se trabaja con una población infinita de orígenes (como usuarios de telefonía), la cual ofrece tráfico en conjunto a N servidores. La tasa de llegadas de nuevas llamadas (tasa de nacimiento) es igual a λ y es constante, no depende del número de recursos activos, porque se supone que el total de recursos es infinito. La tasa de abandono es igual al número de llamadas en progreso dividida por h, la media del tiempo de llamadas en espera. La fórmula calcula la probabilidad de bloqueo en una pérdida del sistema, si un requerimiento no es atendido inmediatamente cuando trata de utilizar un recurso, y este es abortado. Por lo tanto no son encolados. El bloqueo ocurre cuando hay un nuevo requerimiento de recursos, pero todos los servidores ya están ocupados. La fórmula asume que el tráfico que es bloqueado se libera inmediatamente.

$$B(N, A) = \frac{\frac{A^N}{N!}}{\sum_{i=0}^{N} \frac{A^i}{i!}}$$

A continuación se muestra cómo puede ser expresado recursivamente, en una forma que es usada para calcular tablas de la fórmula de Erlang B:

$B(0, A) = 1$.
$B(N, A) = \frac{AB(N-1,A)}{N+AB(N-1,A)}$.
donde:

- B es la probabilidad de bloqueo.
- N es el número de recursos como servidores o circuitos en un grupo.
- $A = \lambda h$ es la cantidad de tráfico entrante expresado en Erlangs.

La fórmula Erlang B se aplica a los sistemas con pérdidas, tales como sistemas telefónicos tanto fijos como móviles, que no ofrecen almacenamiento de llamadas (es decir, no permiten dejar la llamada en espera), y no se pretende que lo hagan. Se supone que las llegadas de llamadas pueden ser modeladas por un proceso de Poisson, pero es válida para cualquier distribución estadística de tiempos entre llamadas.

El Erlang B también es una herramienta para dimensionar tráfico entre centrales de comunicación de voz.

A.2 Fórmula Erlang C

La Fórmula de Erlang C también asume una infinita población de fuentes, las cuales ofrecen en conjunto, un tráfico de $\boldsymbol{A}$ Erlangs hacia $\boldsymbol{N}$ servidores. Sin embargo, si todos los servidores estan ocupados cuando una petición llega de una fuente, la petición es introducida en una cola. Un sin fin de números de peticiones podrían ir a la cola simultáneamente. Esta fórmula calcula la probabilidad de que una petición se sume a la cola dependiendo de la intensidad de tráfico, asumiendo que las llamadas que fueron bloqueadas se quedarán en el sistema hasta que se pueda atender. Esta fórmula es usada para determinar la cantidad de agentes o representantes de clientes, que se necesitará en un

centro de llamadas (*Call Center*) para después saber la probabilidad de entrar en cola. La fórmula de Erlang C tiene la siguiente estructura:

$$P_w = \frac{\frac{A^N}{N!}\frac{N}{N-A}}{\sum_{i=0}^{N-1}\frac{A^i}{i!} + \frac{A^N}{N!}\frac{N}{N-A}}.$$

Donde:

- A es la intensidad total del tráfico ofrecido en unidades de Erlangs.

- N es la cantidad de servidores.

- P_w es la probabilidad de que un cliente tenga que esperar para ser atendido.

Se asume que las llamadas entrantes puede ser modeladas usando una distribución de Poisson y que el tiempo de espera de las llamadas se describen por una distribución exponencial negativa.

A.3 Fórmula Engset

Esta fórmula está relacionada con las anteriores pero funciona con una población finita de $\boldsymbol{S}$ orígenes en lugar de la población infinita de orígenes que asume Erlang B y C:

$$E(N, A, S) = \frac{A^N \begin{pmatrix} S \\ N \end{pmatrix}}{\sum_{i=0}^{N} A^i \begin{pmatrix} S \\ i \end{pmatrix}}.$$

Esto puede ser expresado recursivamente del siguiente modo, en una forma que es usada para calcular las tablas de la fórmula Engset:

E(0, A, S)= 1.

$E(N, A, S) = \frac{A(S-N+1)E(N-1,A,S)}{N+A(S-N+1)E(N-1,A,S)}$.

Donde:

- E es la probabilidad de bloqueo.

- A es el tráfico en Erlangs generado por cada orígen cuando está desocupado.

- S es el número de orígenes.

- N es el número de servidores.

De nuevo, se asume que las llamadas que llegan pueden ser modeladas por una distribución Poisson y que los tiempos de espera se desciben por una distribución exponencial negativa. Sin embargo, debido a que hay un número finito de servidores, la tasa de llegada de las nuevas llamadas decrece a medida que nuevos orígenes se vuelven ocupados y por lo tanto no pueden originar nuevas llamadas. Cuando $N = S$, la fórmula se reduce a una distribución binomial.

Todas las fórmulas anteriores se pueden revisar con más detalle en [9].

Bibliografía

[1] Aggarwal, C, Han Jiawei and Yu, Philip. (2004). *On Demand Classification of Data Streams*. U.S. National Science Foundation Grant.

[2] Fayyad, U. et al. (1996). *Advanced in Knowledge Discovery and Data Mining*. MIT Press.

[3] FUNDIBEQ (2002, Septiembre 10). *Histogramas*. Recuperado (Marzo 27, 2009), de: http://www.fundibeq.org/metodologías/herramientas/histogramas.pdf.

[4] Golab and Özsu, T. (Junio 2003). *Issues in Stream Management*. Sigmod Record, Vol. 32, No. 2.

[5] Hernández José, Ramírez María y Ferri César. (2004). *Introducción a la Minería de Datos*. España. Pearson Prentice Hall.

[6] Hoffman, Kenneth. (1989). *Álgebra Lineal*. Prentice Hall Hispanoamericana.

[7] Jollife, I.T. (1986). *Principal Component Analysis*. U.S.A. Springer - Verlag.

[8] Molina José y García Jesús. (2004). *Técnicas de Análisis de Datos: Aplicaciones prácticas usando Microsoft Excel y WEKA*. Universidad Carlos III de Madrid, España.

[9] Parkinson, R. (2003). *Traffic Engineering Techniques in Telecomunications*. 2

[10] Solomonoff, R. (1996). *A Formal Theory of Inductive Inference, Part I*. Information and Control, Part I: Vol 7, No. 1, pp. 1-22.

[11] S-PLUS 2000. *Modern Statistics and Advanced Graphics*. Mathsoft, 1999, pp. 49-51, 427-431.

[12] Turaisingham, B. (1999). *Data Mining: Technologies, Techniques, Tools and Trends*. CRC Press.

[13] Uriel, E. (1995). *Análisis de Datos: Series Temporales y Análisis Multivariante*. Publicado por AC.

[14] Vera, Arturo (2006, Diciembre 7). *Sistemas Celulares de Tercera Generación*. Recuperado (Noviembre 5, 2008), de: http://www.monografías.com/trabajos15/telefonía-celular/telefonía-celular.shtml.

[15] Weiss and Indurkhya. (1998). *Predictive Data Mining*. San Francisco: Morgan Kaufmann.

[16] Yerena, J. (2008). *Series Temporales, Kriging y sus Aplicaciones*. Tesis de Grado, Escuela de Matemática de la UCV.