Una Aplicación de Clusters para Datos Correlacionados

José Ramón Domínguez Molina y Graciela Ma. González Farías
Centro de Investigación en Matemáticas A.C.

1 Introducción

La formación de grupos provenientes de muestras aleatorias de una distribución multivariada, es un tema que ha sido ampliamente estudiado en la literatura estadística. Estas observaciones aparecen en un orden estrictamente arbitrario, por lo tanto el orden de estas puede ser permutado sin alterar los resultados. Esta situación sin embargo, no es característica de todas las formas de datos multivariados.

Algunos tipos de observaciones aparecen en un orden específico y no es permisible intercambiar observaciones sin un cambio fundamental en el resultado. Esto es generalmente cierto para observaciones que son ordenadas en el tiempo, en el espacio o en ambas. En este tipo de datos el orden o localización induce dependencia entre los puntos vecinos, por lo que una técnica para agrupar observaciones correlacionadas deberá tomar en cuenta tal correlación, además de las similitudes o diferencias que existan entre las series.

Una sucesión de observaciones que son tomadas secuencialmente en el tiempo se le denomina "serie de tiempo". Muchos conjuntos de datos que aparecen en la vida diaria son series de tiempo, como por ejemplo: los reportes mensuales de tasas de interés de los bancos, ventas diarias, semanales, mensuales, anuales, de alguna compañía, la información que se origina cada año con la economía de México, por mencionar sólo unos ejemplos.

Un problemas clásico al analizar series de tiempo es el de agrupar tales series en categorías similares y la clasificación de nuevas series dentro de una de estas categorías. Estos dos

problemas han sido estudiados para vectores de observaciones convencionales y existe una amplia literatura (ver por ejemplo Johnson y Wichern (1992), McLachlan (1992) dedicada a discriminar y agrupar vectores normales multivariados. Tales metodologías usualmente dependen de las diferencias entre las medias de las subpoblaciones; las funciones lineales resultantes han sido adaptadas computacionalmente y son fácilmente aplicables para agrupar y discriminar y clasificar conjuntos de datos de alta dimensión y se encuentran en la paquetería de la mayor parte de los softwares estadísticos.

Métodos gráficos para la agrupación de datos multivariados, tales como, gráficas de las primeras componentes principales, gráficas de Andrews, caras de Chernoff entre otras, no toman en cuenta las correlaciones que existen dentro de las series y por lo tanto no son efectivas.

El objetivo de este trabajo es estudiar, diseñar e implementar un entorno adecuado para el estudio de un gran número de series de tiempo por ejemplo ventas de compañías refresqueras, consumo de bienes energéticos, datos de sismología, datos de estaciones meteorológicas, etc. de forma tal que, nos permita agrupar las series de tiempo en k grupos lo más homogéneos posibles y en cada grupo encontrar una familia de modelos que permita desarrollar una metodología automática de modelización Este problema ha sido estudiado en la literatura, ver Kakizawa et al.(1988) , Piccolo. D. (1990) , entre otros. Sin embargo en diferentes aplicaciones es necesario contar con una herramienta con capacidad de almacenamiento y procesamiento de datos, y que no requiera de un proceso de modelación detallado previo a la formación de los grupos por lo que se justifica un esfuerzo que tienda a un mejor conocimiento de las características esenciales de estas series para su agrupamiento y al desarrollo de técnicas de modelización que permitan un tratamiento simple, general y lo más automático posible. Todo esto, con el propósito de conseguir modelos de alta precisión en la predicción a corto plazo como ilustraremos más adelante en el caso de la industria refresquera.

En la Sección 1 se presenta una revisión de métodos para formación de conglomerados (clusters) con series de tiempo, en la Sección 2 se propone una medida de disimilitud entre

series de tiempo, la cual tomará en cuenta la estructura de dependencias dentro de las series basada en el trabajo de Piccolo (1990). Con esta medida definida, se podrán entonces aplicar los métodos clásicos de agrupación jerárquicos y no jerárquico. En la Sección 3 se lleva a cabo la comparación de la propuesta contra algunas metodologías que se encuentran en la literatura y por último, en la Sección 4, se presenta una aplicación con series de actividad económica diaria en la industria refresquera.

Cabe mencionar que se desarrolló el software necesario para su adaptación y las interfaces que lo hacen de fácil acceso a los usuarios incluyendo un módulo de detección de observaciones aberrantes para la modelización automática de pronósticos basado en la propuesta de la sección 2 y en la extensión de un resultado dado por Díaz-García y González-Farías (2001)[1]. El software fue desarrollado en S-plus para facilitar las interfaces pero es fácilmente transferible a lenguajes de programación básicos como C o Fortran, que agilizan su implementación. El software esta disponible bajo requisición en: jrdguez@cimat.mx.

2 Antecedentes

Una serie de tiempo de n observaciones sucesivas $\mathbf{z} = (z_1, z_2, ..., z_n)^T$ es considerada como una realización muestral de una población infinita, la cual podría ser generada por un proceso estocástico. Para hacer pronósticos, será necesario inferir la distribución de probabilidad de valores futuros de $\mathbf{z}$, dada una muestra de valores pasados de $\mathbf{z}$. Para hacer esto, es necesario caracterizar a $\mathbf{z}$ mediante clases de modelos estocásticos que puedan ser capaces de describir su comportamiento. Una excelente descripción de modelos para series de tiempo, frecuenciales o en el tiempo (ecuaciones en diferencias) puede encontrarse en : Fuller (1995) , Hamilton (1994) , Box y Jenkins (1994) , Wei (1990) , Shumway (1988) entre otros.

Se presentan dos propuestas de formación de conglomerados : Kakizawa, Shumway y Taniguchi (1998), y Piccolo (1990). La primera es válida para series estacionarias y la segunda

[1] Ver: Domínguez Molina J.R. (2001) .

esta diseñada para modelos ARIMA en general, pero es necesario contar con un modelo previo antes de llevar a cabo la formación de clusters.

2.1 Propuesta de Kakizawa

Kakizawa, Shumway y Taniguchi (1998) sugieren que las similitudes y diferencias entre series de tiempo pueden ser caracterizados en términos de la estructura de covarianza o equivalentemente por el espectro. Esto lo proponen debido a que muchos de los análisis de discriminación que se han hecho de series de tiempo involucran el espectro o una amplitud que es proporcional a la integral del espectro, por tanto sugieren que la información para discriminar entre este tipo de series puede estar contenida en el espectro y proponen dos medidas de disimilitud en función de estas.

Además, comentan que las similitudes y diferencias entre este tipo de series no pueden ser siempre caracterizadas por las diferencias entre las medias de las subpoblaciones, ya que una serie de tiempo frecuentemente envuelve miles de observaciones correlacionadas en el tiempo. La dimensión de las series de tiempo hace prohibitivo los cálculos computacionales usando métodos multivariados clásicos. La preponderancia de tales vectores en muchas disciplinas tales como sismología y nuevos estudios que coleccionan grandes cantidades de datos hacen que el estudio de análisis discriminantes y de agrupamiento de series de tiempo sea de gran interés.

Aunque en su artículos ellos trabajan con series de tiempo multivariadas las cuales son más generales, en este trabajo se adecuarán sus resultados para series univariadas, que permita la comparación inmediata con los otros procedimientos. Para ello, las medidas de información de Kullback-Leibler (1951) y Chernoff (1952) serán desarrolladas para su aplicación en series univariadas.

2.1.1 Medidas de disimilitud

Se asume que se tiene una serie de tiempo estacionaria, x_t. Las funciones de densidad de probabilidades de este vector serán denotadas por $p(x)$ y $q(x)$, donde estas dos funciones típicamente corresponden a dos diferentes hipótesis acerca de la serie observada x_t. En el caso estacionario, se usará $\mathbf{f}(\lambda)$ y $\mathbf{g}(\lambda)$ para la densidad espectral correspondientes a la función de autocovarianzas $\mathbf{R}_p(s-t)$ y $\mathbf{R}_q(s-t)$.

Una medida clásica de disimilitud entre dos densidades multivariadas es la Kullback-Leibler (KL), la cual esta dada por

$$I(p;q) = E_p\left\{\log\frac{p(x)}{q(x)}\right\}$$

donde E_p denota la esperanza bajo la densidad de $p(.)$.

La medida de KL toma la forma

$$I(p;q) = \frac{1}{2}\left(\mathbf{R}_p\mathbf{R}_q^{-1} - \log\frac{|R_p|}{|R_q|} - T\right)$$

Una medida simétrica de disimilitud, la J divergencia, es definida por

$$J(p;q) = I(p;q) + I(q;p)$$

Esta cumple todas las propiedades de una distancia excepto la desigualdad del triángulo y por lo tanto se le llama una quasi-distancia.

Parzen (1990) propone una medida usando Chernoff

$$B_\alpha = -\log E_p\left\{\left(\frac{p(x)}{q(x)}\right)^\alpha\right\}$$

como una medida de disimilitud entre dos densidades, donde la medida esta indexada por α, $0 < \alpha < 1$. Para $\alpha = 0.5$, la medida de Chernoff es la medida simétrica de divergencia propuesta por Bhattacharya (1943), y se mantendrá la notación B_α. Para dos vectores que difieren sólo en la estructura de covarianzas, la medida toma el valor de

$$B_\alpha(p;q) = \frac{1}{2}\left(\log\frac{|\alpha\mathbf{R}_p + (1-\alpha)\mathbf{R}_q|}{|\mathbf{R}_q|} - \alpha\log\frac{|\mathbf{R}_p|}{|\mathbf{R}_q|}\right)$$

De nuevo se definirá una quasi-distancia de la siguiente manera

$$JB_{\alpha}\left(p;q\right)=B_{\alpha}\left(p;q\right)+B_{\alpha}\left(q;p\right)$$

La medida de disimilitud entre espectros desarrolladas en la sección previa pueden ser usadas como medidas de quasi-distancias para agrupar series de tiempo. Por ejemplo, sean $\mathbf{f}_{T}\left(\lambda_{s}\right)$ y $\mathbf{g}_{T}\left(\lambda_{s}\right)$ espectros de dos diferentes series de tiempo, calculadas mediante el estimador espectral suavizado,

$$\begin{aligned}\hat{f}_{x}\left(v_{k}\right) &= \frac{1}{L}\sum_{l=-(L-1)/2}^{(L-1)/2}P_{x}\left(v_{k}+\frac{l}{T}\right)\\ &= \frac{1}{L}\sum_{l=-(L-1)/2}^{(L-1)/2}\left|X\left(k+l\right)\right|^{2}\end{aligned}$$

donde

$$X\left(k\right)=\frac{1}{\sqrt{T}}\sum_{t=0}^{T-1}x_{t}\exp\left(-2\pi iv_{k}t\right)$$

y

$$v_{k}=\frac{k}{T},\quad k=0,1,2,...,\frac{T}{2}$$

Entonces considerando la aproximación para las medidas de las distancias calculadas para la J divergencia, para el caso univariado tenemos,

$$J\left(\mathbf{f}_{T};\mathbf{g}_{T}\right)=\frac{1}{2T}\sum_{s}\left(\mathbf{f}_{T}\left(\lambda_{s}\right)\mathbf{g}_{T}^{-1}\left(\lambda_{s}\right)+\mathbf{g}_{T}\left(\lambda_{s}\right)\mathbf{f}_{T}^{-1}\left(\lambda_{s}\right)-2\right)$$

y la información simétrica de divergencia de Chernoff

$$JB_{\alpha}\left(\mathbf{f}_{T};\mathbf{g}_{T}\right)=\frac{1}{2T}\sum_{s}\left(\log\frac{\left|\alpha\mathbf{f}_{T}\left(\lambda_{s}\right)+\left(1-\alpha\right)\mathbf{g}_{T}\left(\lambda_{s}\right)\right|}{\left|\mathbf{g}_{T}\left(\lambda_{s}\right)\right|}+\log\frac{\left|\alpha\mathbf{g}_{T}\left(\lambda_{s}\right)+\left(1-\alpha\right)\mathbf{f}_{T}\left(\lambda_{s}\right)\right|}{\left|\mathbf{f}_{T}\left(\lambda_{s}\right)\right|}\right)$$

donde $\lambda_{s}=\frac{2\pi s}{T},\ s=1,2,...,T.$

Por tanto es natural proponer usar estas dos quasi-distancias para agrupar series de tiempo similares. Usando estas dos medidas sobre una muestra de serie de tiempo podemos producir una matriz de quasi-distancias que puede ser usada como entrada en métodos de agrupación jerárquicos y no jerárquicos.

2.2 Medida de Piccolo

Piccolo (1990), propone una métrica para series de tiempo donde un modelo $ARIMA$ es ajustado a un gran número de series de tiempo con el propósito de hacer predicciones y ajustes estacionales. Observando a simple vista los modelos ajustados, es posible encontrar similitudes entre ellos y por lo tanto puede ser útil clasificarlos para detectar unos pocos modelos representativos de ese gran número de series. Esto es lo que hace importante investigar medidas de disimilaridad entre modelos del tipo $ARIMA$.

Utilizando la metodología propuesta por Piccolo es necesario escoger primero un procedimiento estadístico para el ajuste de modelos $ARIMA$,

$$\phi_p(B)\,\Phi_P(B^s)\,(1-B)^d\,(1-B^s)^D\,Z_t = \theta_q(B)\,\Theta_Q(B^s)\,a_t \tag{1}$$

Algunos procedimientos ampliamente usados como el X11, X11-ARIMA y TRAMO-SEATS entre otros dan buenos resultados en el ajuste de modelos $ARIMA$.

2.2.1 Definición y propiedades de la medida de distancia entre series

Sea a_t ruido blanco gausiano y Z_t un proceso estocástico con media cero tal que

$$Z_t \sim ARIMA(p,d,q)\,(P,D,Q)_s \tag{2}$$

y por lo tanto, siguiendo la notación estándar de Box y Jenkins (2) se puede escribir de la siguiente manera

$$\varphi(B)\,Z_t = \theta(B)\,a_t \tag{3}$$

donde $\varphi(B) = \phi_p(B)\,\Phi_P(B^s)\,(1-B)^d\,(1-B^s)^D$ y $\theta(B) = \theta_q(B)\,\Theta_Q(B^s)$.

Asumiendo que el modelo (3) es invertible, esto es, que las raíces del polinomio $\theta(B)$ estén fuera del circulo unitario, entonces $Z_t \in \pounds$, donde $\pounds$ es de la clase de modelos invertibles $ARIMA$. Una declaración equivalente es que si $Z_t \in \pounds$, entonces $W_t = (1-B)^d\,(1-B^s)^D\,Z_t$ es un proceso gausiano autorregresivo de promedios móviles $ARMA$.

Ahora, se sabe que si $Z_t \in \pounds$,

$$\begin{aligned} Z_t &= \bar{Z}_{t-1} + a_t \\ \bar{Z}_{t-1} &= \sum_{j=1}^{\infty} \pi_j Z_{t-j} = \{1 - \pi(B)\} Z_t \end{aligned}$$

donde $\bar{Z}_{t-1}$ es independiente de a_t. Entonces

$$Z_t = \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \cdots + a_t \tag{4}$$

donde el operador $AR(\infty)$ es definido por $\pi(B) = \varphi(B)/\theta(B) = 1 - \pi_1 B - \pi_2 B^2 - \ldots$.

Por lo tanto, dados valores iniciales y ordenes conocidos, cualquier proceso $Z_t \in \pounds$ es completamente caracterizado por la secuencia $\boldsymbol{\pi}' = (\pi_1, \pi_2, ...)$ la cual especifica completamente la distribución del proceso W_t y Z_t cuando a_t es un proceso gausiano de ruido blanco. La secuencia de $\boldsymbol{\pi}$ contiene toda la información útil acerca de la estructura del proceso estocástico ya que cualquier otra información necesaria para especificar Z_t es justamente a_t, la cual es impredecible al tiempo $t-1$. Entonces una medida de la diversidad estructural entre $X_t \in \pounds$ y $Y_t \in \pounds$ puede ser obtenida comparando sus respectivas secuencias de $\boldsymbol{\pi}$.

Una métrica sobre $\pounds$ puede definirse por la distancia

$$d(X,Y) = \left\{ \sum_{j=1}^{\infty} (\pi_{j,x} - \pi_{j,y})^2 \right\}^{1/2} \tag{5}$$

ya que $\sum_{j=1}^{\infty} \pi_j$, $\sum_{j=1}^{\infty} |\pi_j|$ y $\sum_{j=1}^{\infty} \pi_j^2$, son cantidades bien definidas, es fácil mostrar que $d(X,Y)$ siempre existe para cualquier proceso en $\pounds$ y satisface las propiedades clásicas de distancias, a saber, no-negativa, simétrica y la desigualdad del triángulo.

La definición de una distancia entre modelos de series de tiempo, permite en forma inmediata la aplicación de algoritmos de agrupamiento.

Es útil enlistar algunas de las propiedades de la distancia (5).

1. La distancia propuesta es completamente general y puede ser calculada aún si algunos ordenes del modelo pueden ser comparados con cero. De hecho, la métrica depende de los coeficientes de un $AR(\infty)$ el cual siempre converge. Una interpretación útil de la secuencia única de $\boldsymbol{\pi}$ es que determina la función de pronósticos para valores futuros dado valores presentes y pasados.

2. La métrica (5) no toma en cuenta la varianza residual ya que este es un parámetro puramente de escala y no es relevante para la comparación que pretendemos hacer.

3. El elemento cero (origen) de $\pounds$ es cualquier proceso de ruido blanco el cual esta caracterizado por la secuencia nula $(0, 0, ...)$. Entonces, la distancia entre X_t y el origen 0 es la norma de X_t, por ejemplo

$$d(X, 0) = \left\{ \sum_{j=1}^{\infty} (\pi_{j,x})^2 \right\}^{1/2} = ||X||$$

4. Existe una isometría entre las clases de $ARIMA$ no estacionales y las correspondientes clases de modelos $ARIMA$ estacionales. Por lo tanto, si X_t y Y_t están asociados con los operadores $\pi_x(B)$ y $\pi_y(B)$ respectivamente y $X_t^{'}$ y $Y_t^{'}$ están asociados con los operadores $\pi_x(B^s)$ y $\pi_y(B^s)$, entonces $d(X, Y) = d(X', Y')$.

5. Si restringimos nuestra atención a procesos $ARMA$ admisibles, podemos definir una métrica dual como:

$$\delta(X, Y) = \left\{ \sum_{j=1}^{\infty} \left(\psi_{j,x} - \psi_{j,y}\right)^2 \right\}^{1/2}$$

donde ψ es una secuencia que define un $MA(\infty)$ con $\theta(B)\varphi^{-1}(B) = \pi^{-1}(B)$. Sin embargo esta métrica es inadecuada para procesos integrados y no puede aplicarse a procesos no estacionarios.

3 Modificación del procedimiento de Piccolo

La metodología propuesta por Piccolo requiere en primer lugar ajustar un modelo del tipo $ARIMA$ para cada una de las series y posteriormente calcular los valores de π_j, $j = 1, 2, \ldots$. Si se requiere hacer esto para un gran número de series, resulta complicada y pesada computacionalmente. Además aplicar esta metodología sería como hacer grupos a posteriori, porque primero modelamos cada serie individualmente y después las agrupamos en función de esos modelos. En algunas de las aplicaciones la formación de clusters ayuda a entender también características que no habían sido tomadas en cuenta y permite identificar grupos racionales para la creación de escenarios comunes, que pueden resultar muy útiles en el proceso de predicción.

Por lo anterior, surge la idea de proponer una medida entre series que no requiera modelarlas antes de construir una medida de disimilitud para formar clusters de modelos semejantes. Con esto lo que se pretende es primero formar grupos y después modelar cada grupo con una familia mucho más pequeña de modelos tipo $ARIMA$ que se puedan ajustar. Esto incluye el caso de no-estacionareidad en la media, la no inclusión de covariables para limpiar las series antes de formar los grupos, como por ejemplo, variables para modelar valores aberrantes, tendencias, etc.

3.1 Definición de la medida de distancia propuesta

Con la representación de Z_t en (4) , se tiene que el proceso puede aproximarse por

$$Z_t = \sum_{l=1}^{h} \pi_l Z_{t-l} + a_t$$

para algún h, ya que $\pi_j \to 0$ cuando $j \to \infty$.

La función de log-verosimilitud de la serie observada Z_t condicionada sobre los primeros h valores, esta dada por:

$$l\left(\pi, \sigma^2 \mid Z_1, ..., Z_h\right) \propto -\left(\frac{n-h}{2}\right) \ln\left(\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{t=h+1}^{n} \left(Z_t - \pi X_t\right)^2$$

donde σ^2 es la varianza del proceso de ruido blanco, a_t, $X_t = \left(Z_{t-1}, ..., Z_{t-h}\right)$, $\pi = \left(\pi_1, ..., \pi_h\right)$. El estimador de máxima verosimilitud condicional de π, es

$$\hat{\pi} = \left(X^T X\right)^{-1} X^T Z^* \tag{6}$$

donde

$$X = \begin{bmatrix} X_{h+1} \\ X_{h+2} \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} Z_h & Z_{h-1} & \cdots & Z_1 \\ Z_{h+1} & Z_h & \cdots & Z_2 \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n-1} & Z_{n-2} & \cdots & Z_{n-h} \end{bmatrix} \quad \text{y} \quad Z^* = \begin{bmatrix} Z_{h+1} \\ Z_{h+2} \\ \vdots \\ Z_n \end{bmatrix}$$

el cual coincide con el estimador de mínimos cuadrados.

Haciendo esto se obtendrá una aproximación de los coeficientes de $\boldsymbol{\pi}$. Por lo que una métrica sobre $\mathcal{L}$ puede definirse ahora por la distancia

$$d\left(X, Y\right) = \left\{ \sum_{j=1}^{h} \left(\hat{\pi}_{j,x} - \hat{\pi}_{j,y}\right)^2 \right\}^{1/2} \tag{7}$$

Se ilustrara mediante un ejemplo la bondad del ajuste. El problema que queda abierto es el número de lags a considerar, esto es, el valor de h. Se puede ver que este valor dependerá de que tan aberrantes son las series bajo estudio.

3.2 Ejemplo

Para ilustrar que la aproximación de los coeficientes de π por medio de $\hat{\pi}$, es bastante buena, se simula una serie bajo el modelo conocido como de Aerolíneas de Box y Jenkins, el cuál esta dado por la siguiente expresión

$$ARIMA\left(0,1,1\right)\times\left(0,1,1\right)_{12}$$

$$\left(1-B\right)\left(1-B^{12}\right)z_t=\left(1-\theta B\right)\left(1-\Theta B^{12}\right)a_t$$

Este modelo se puede escribir en la forma

$$z_t=\sum_{j=1}^{\infty}\pi_j z_{t-j}+a_t$$

Los pesos π pueden ser obtenidos al igualar los coeficientes en

$$\begin{aligned}\frac{\left(1-B\right)\left(1-B^{12}\right)}{\left(1-\theta B\right)\left(1-\Theta B^{12}\right)} &= 1-\pi_1 B-\pi_2 B^2-\cdots\\ \left(1-B\right)\left(1-B^{12}\right) &= \left(1-\theta B\right)\left(1-\Theta B^{12}\right)\left(1-\pi_1 B-\pi_2 B^2-\cdots\right)\end{aligned}$$

donde

$$\begin{aligned}\pi_j &= \theta^{j-1}\left(1-\theta\right), \quad j=1,2,...,11\\ \pi_{12} &= \theta^{11}\left(1-\theta\right)+\left(1-\Theta\right)\\ \pi_{13} &= \theta^{12}\left(1-\theta\right)-\left(1-\theta\right)\left(1-\Theta\right)\\ \pi_j &= \theta\pi_{j-1}+\Theta\pi_{j-12}-\theta\Theta\pi_{j-13}, \quad j=14,15,...\end{aligned}$$

En la figura 1.1 se grafican los pesos reales de $\boldsymbol{\pi}_j$ $(j=1,...,30)$ para los valores de los parámetros $\theta=0.4$ y $\Theta=0.6$ y los valores de los coeficientes estimados de π, $\hat{\pi}$ dado por (6), para una serie simulada con $n=1000$. Se observa en la gráfica que la aproximación es bastante buena.

El programa utilizado para hacer esta simulación fue elaborado en S-PLUS y el archivo fuente se encuentra en Domínguez-Molina, J.R. (2001).

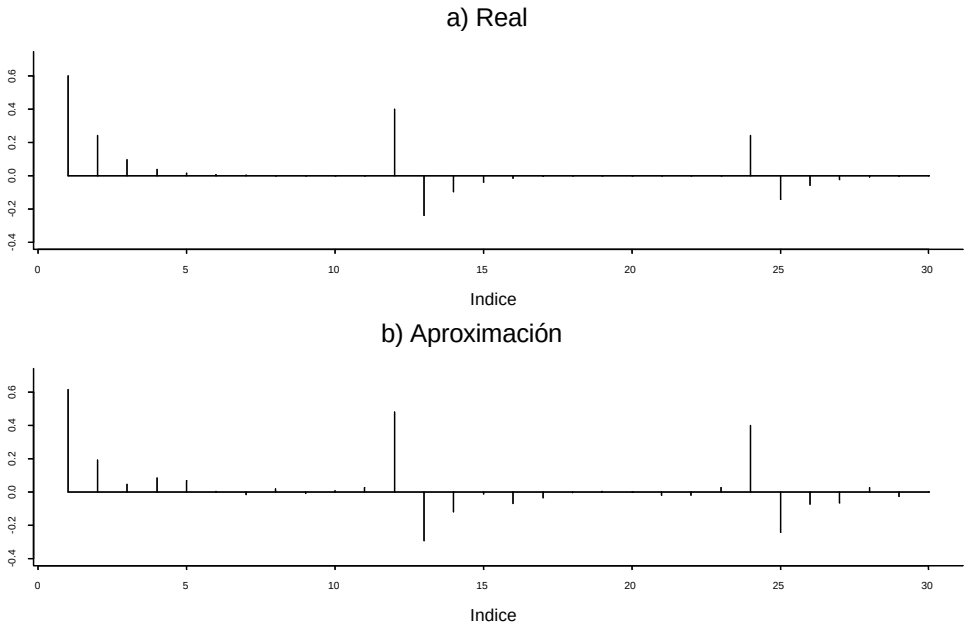

Figura 1.1. a) π_j (Reales), b) $\hat{\pi}_j$ (Estimados).

4 Comparación

Para comparar todas las técnicas descritas anteriormente se generan 10 series $(t = 1, ..., 250)$ de cada uno de los siguientes modelos:

Modelo 1.

$$Z_t \sim ARIMA\,(0,1,1)\,(0,1,1)_4$$

donde $\theta = 0.50,\ \Theta = 0.50.$

Modelo 2.

$$Z_t \sim ARIMA\,(0,1,3)\,(0,1,1)_4$$

donde $\theta = (0.50, 0.25, -0.50)\,,\ \Theta = 0.50.$

Modelo 3.

$$Z_t \sim ARIMA\left(3,1,0\right)\left(0,1,1\right)_4$$

donde $\phi = (0.50, 0.25, -0.50)$, $\Theta = 0.50$.

Para poder comparar las distintas propuestas utilizaremos en todos los casos una técnica jerárquica de agrupamiento por medio del método de liga completa (ver apéndice B). Como se sabe que las series generadas no son estacionarias, y el método propuesto Kakizawa, *et al* (1998) requiere que sean series estacionarias, es necesario entonces tomar las diferencias de orden 1 y 4 de las series, esto es

$$Y_t = (1 - B)\left(1 - B^4\right) Z_t.$$

En la figura 1.2, se tienen los dendogramas obtenidos a partir de utilizar una técnica de formación de clusters jerárquica y por el método de liga completa. La matriz de disimilitudes para

a), b) fue construida a partir de la metodología estudiada en la sección 1.1 sobre las series Y_t,

c) se utilizó la metodología propuesta por Piccolo sobre las series Z_t, aunque ya se conocían de antemano los parámetros de los modelos, estos se estimaron y se calculo π_j, $j = 1, ..., 30$.

d) se utilizó la metodología propuesta en la sección 2, estimando a $\hat{\pi}_j$, $j = 1, ..., 7$ a partir de Y_t. Debido a que la aproximación de π se hizo sobre una serie estacionaria, se tiene que sólo es necesario un orden de tamaño 7 para que la clasificación en este caso sea perfecta. Se observa además que la clasificación en todos los casos es perfecta. Si se deseara formar dos grupos, se tiene que todas las series generadas a partir de los modelos 1 y 2 se agruparían en uno sólo. Las medidas KL y Chernoff clasifican de una manera más clara los modelos 1 y 2 del 3.

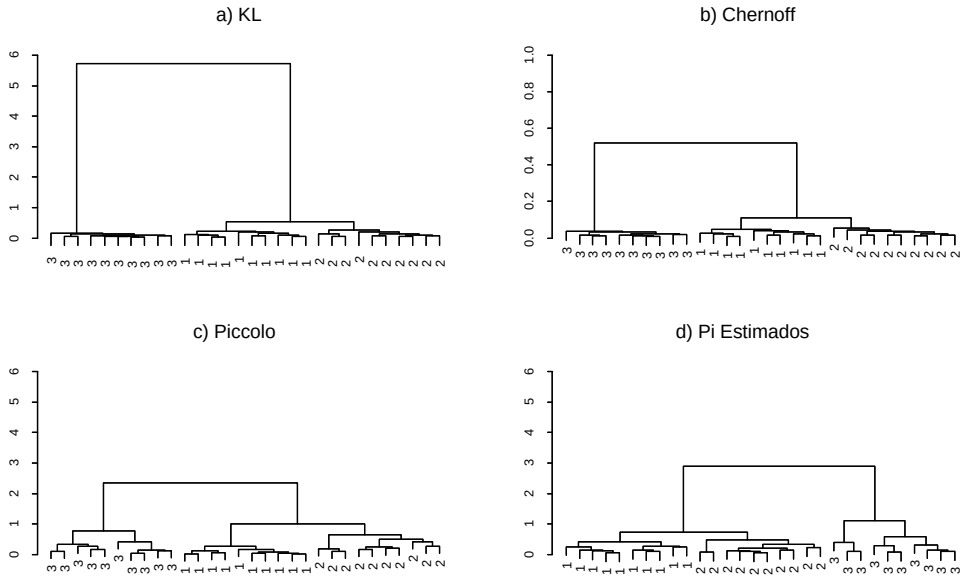

Figura 1.2. Comparación de las los distintos dendogramas para cada una de las distancias propuestas.

En la figura 1.3 se tiene ahora el dendograma utilizando la misma técnica de agrupamiento descrita en el ejemplo anterior, pero utilizando ahora como entrada la matriz de disimilitudes del método propuesto en este capítulo, esto es, estimar π a partir de la serie original z_t. Se observa que la clasificación también es perfecta. En este caso se uso un valor de $h = 30$.

Como se vio en este ejemplo, la metodología propuesta en este trabajo es mucho más simple de aplicar porque no es necesario estimar un modelo ni estacionalizar las series para poder construir una matriz de disimilitudes que sirva de entrada a técnicas de formación de clusters y que los forme con resultados equivalentes.

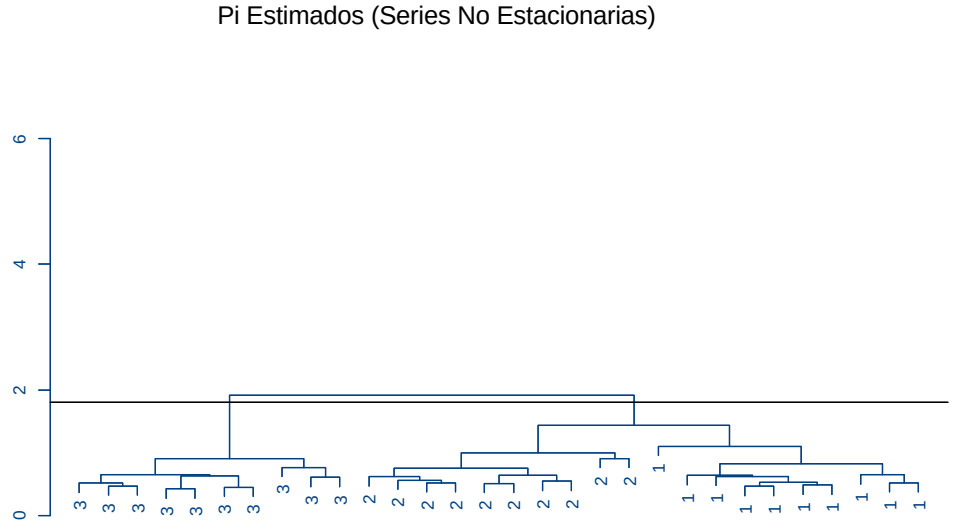

Figura 1.3. Dendograma utlizando la estimación de π sin estacionalizar las series.

5 Aplicación

5.1 Ejemplo 1

Los datos que se utilizarán a continuación son volúmenes de ventas diarias en cajas unidad de una compañía refresquera, registrados a los largo de dos años y medio (912 observaciones).

Se aplicará una transformación a las bases de datos con el propósito de guardar su confidencialidad, pero manteniendo su estructura original, además no se imprimirá en las gráficas el volumen de cajas unidad (eje y).

Utilizando TRAMO se observó que modelos del tipo

$$ARIMA\,(p,d,q)\,(P,D,Q)_7$$

se ajustaban bien. Los ordenes máximos de los modelos fueron

$$p = 3,\ d = 1,\ q = 3,\ P = 1,\ D = 1,\ Q = 1$$

y todos son modelos $ARIMA$ invertibles, por lo que la metodología presentada aquí es aplicable.

Debido a que se tienen series grandes y para mostrar que no es necesario estacionalizar las series, se utilizarán las series originales para aplicar la metodología propuesta. Resulta necesario utilizar un orden de h bastante grande ($h = 50$) para poder captar la estructura de estas series, debido a las raíces unitarias que contienen aunado a diferentes cambios estructurales presentes en cada serie. Este valor se determina empíricamente observando la consistencia de los grupos formados.

Para guardar la confidencialidad de los datos las series se identifican por : Gaseosas, Sabor A-E, Aguas , Mineral, Jugo, estas identificaciones representan las 5 grandes familias de productos que produce la compañía. Estos productos pueden ser vendidos en 11 empaques distintos, esto es, se puede vender un mismo producto pero en presentaciones de 6, 12, 24, etc., unidades por caja. Y por último se clasifican como Retornable (Ret) y No Retornable (NoRet). En los refrescos retornables el envase se recoge de los puntos de venta y se vuelve a utilizar, el refresco no retornable se utiliza una vez y después se desecha.

De la figura 1.4 a la 1.10 se observa que las series dentro de cada grupo tienen una estructura semejante por lo que podemos concluir que la técnica aplicada para formarlos funcionó adecuadamente. El nombre que tiene cada gráfica se refiere a la característica principal que tuvo cada grupo. En la figura 1.11, se tienen las series cuyo comportamiento fue muy distinto al de las demás series y cada una de ellas forman su propio grupo.

Es evidente al observar las gráficas que la mayoría de las series tienen observaciones aberrantes y algunas tienen cambios bruscos de nivel, es conocido que esto puede afectar gravemente la estimación de los parámetros de cualquier modelo que se intente ajustar. Por lo que es interesante repetir el análisis de clusters presentado en esta sección sobre las series

sin el efecto de los outliers. En la siguiente sección se utiliza una técnica para la detección de outliers, ver Domínguez Molina, J.R. (2001).

Realizando el análisis de clusters sobre las series sin el efecto de los outliers se tienen también grupos razonables desde un punto de vista gráfico, pero distintos y no tan satisfactorios como los que se obtuvieron con las series originales, lo que confirma que los outliers son una característica intrínseca para la descripción de la dinámica de las series.

Aplicando la técnica de componentes principales a la matriz de los coeficientes autorregresivos para cada serie, se observan agrupaciones que no resultan razonables, por lo que no se presentan aquí.

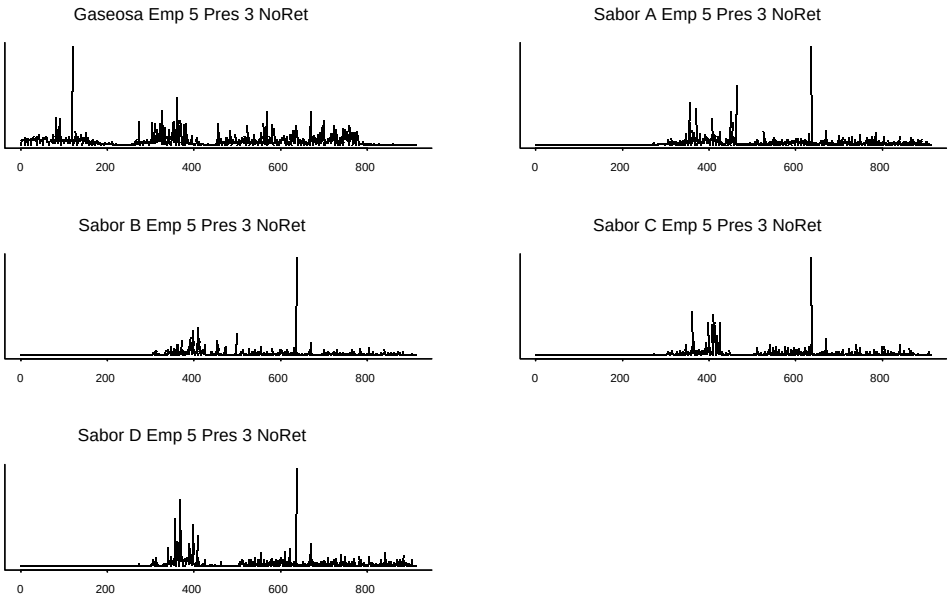

Figura 1.4. Grupo 1: Emp 5 Pres 3 NoRet.

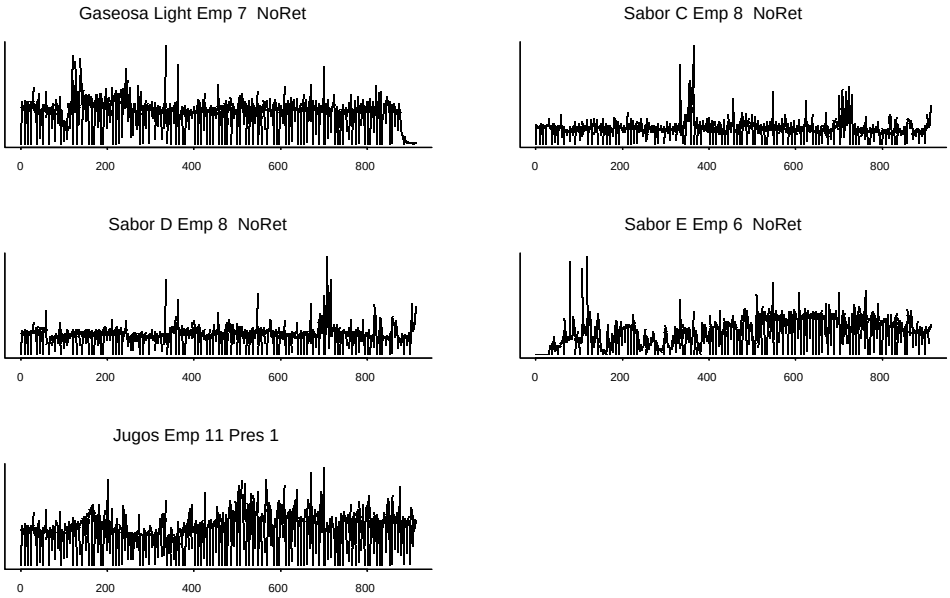

Figura 1.5. Grupo 2: No retornables y media constante.

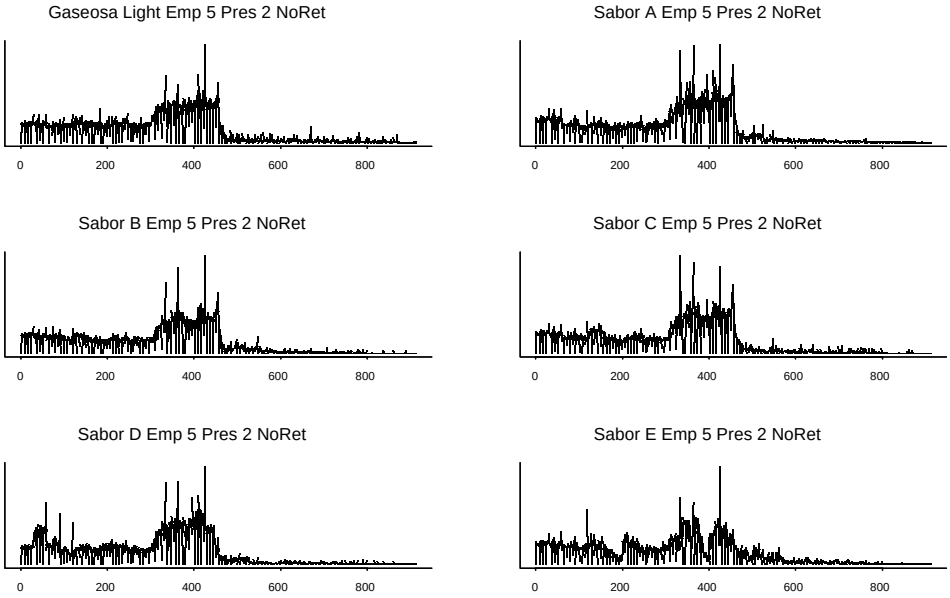

Figura 1.6. Grupo 3: Emp 5 Pres 2 NoRet.

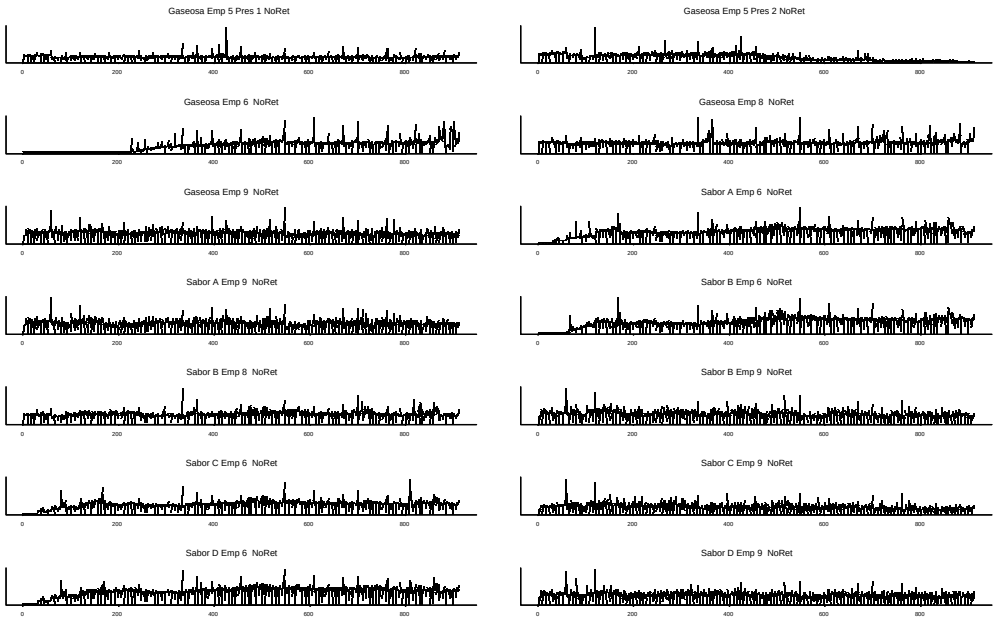

Figura 1.7. Grupo 4. No Retornable.

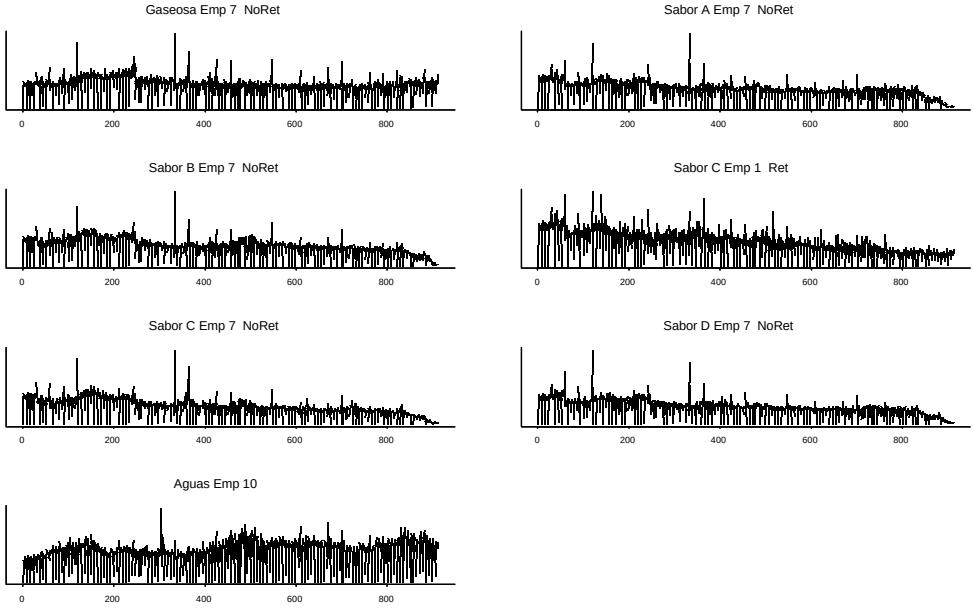

Figura 1.8. Grupo 5: Emp 7 NoRet.

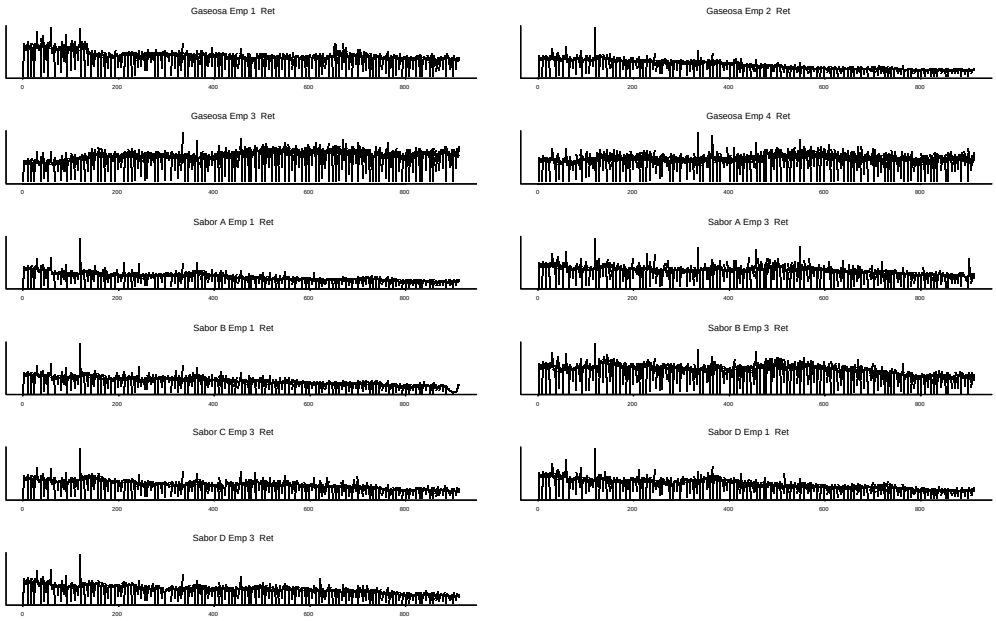

Figura 1.9. Grupo 6: Retornables.

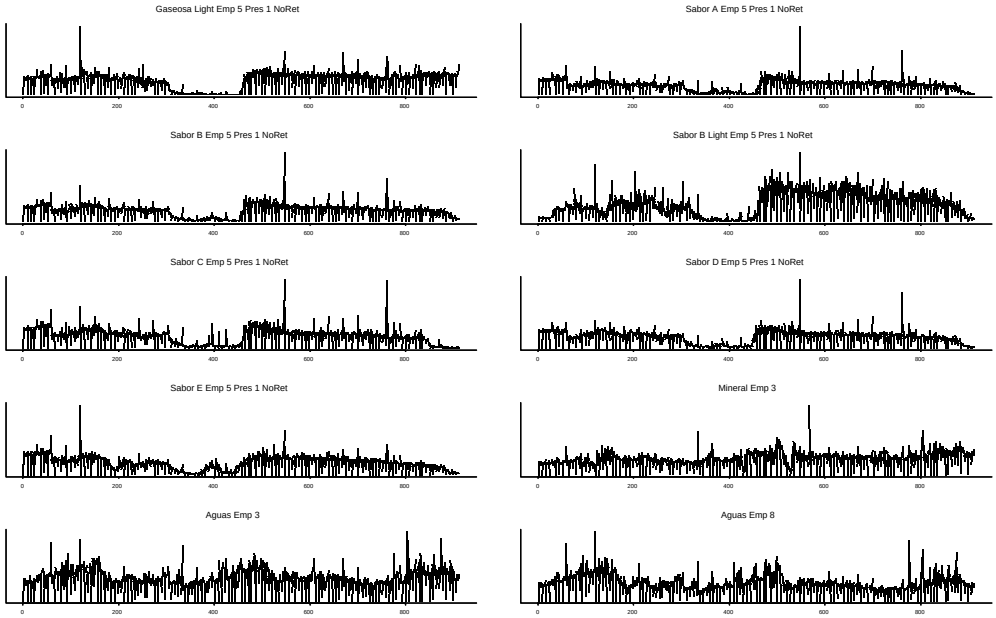

Figura 1.10. Grupo 7: Emp 5 Pres 1 NoRet.

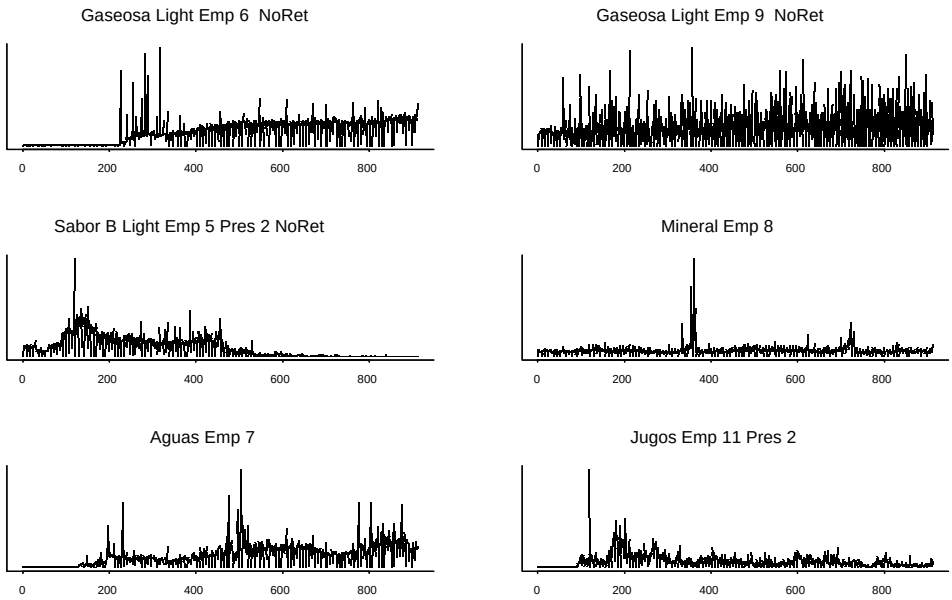

Figura 1.11. Series no agrupadas.

5.2 Ejemplo 2

Haslett, John y Raftery, A (1989) estudian datos de viento de 12 estaciones meteorológicas, tomados cada hora durante los años 1961-1978 en Irlanda. La velocidad del viento fue registrada en knots (1 knot = 0.5148 m/s). El objetivo de su artículo es desarrollar métodos para la evaluación del poder del viento en zonas de Irlanda donde no hay estaciones meteorológicas.

En este trabajo el objetivo sólo será encontrar grupos de ciudades donde las velocidades del viento sean semejantes. Para esto se utilizará un orden de $h = 50$ debido a que las series parecen provenir de un proceso de memoria larga, lo cual ocasiona dependencias en ordenes altos de rezagos. El método de agrupación fue el jerárquico y para el enlace se utilizó la técnica de la liga completa, el cálculo de la distancia entre los coeficientes $\pi's$ fue la Euclideana.

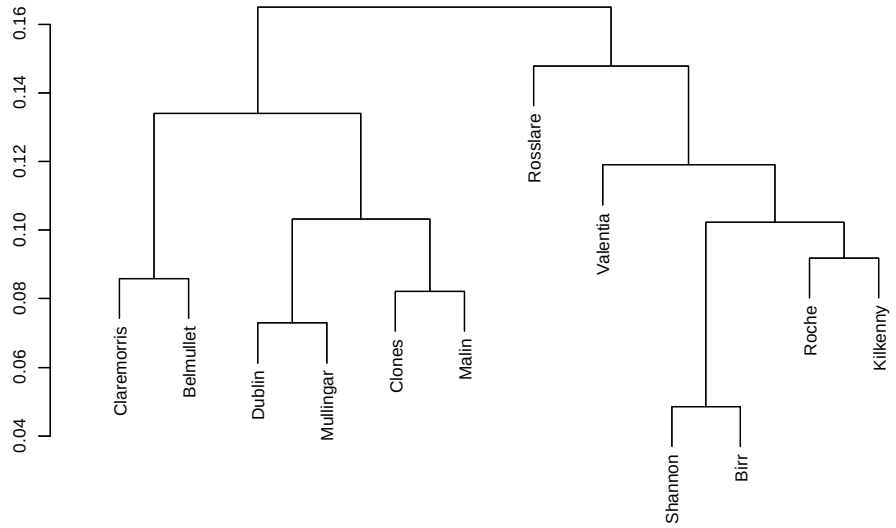

Figura 1.12. Dendograma de las estaciones meteorológicas de Irlanda.

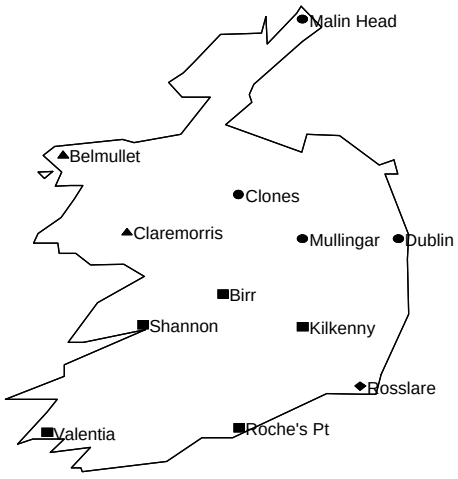

Figura 1.13. Estaciones meteorológicas de Irlanda agrupadas.

En las figuras 1.12 y 1.13 se tienen el dendograma y la ubicación de los grupos formados de las estaciones meteorológicas de Irlanda respectivamente. Se observa que las ciudades se

agrupan por su ubicación geográfica, esto es, Malin Head, Clones, Mullingar y Dublin están al norte; Valentia, Roche's Pt, Shannon, Birr y Kilkenny en el sur; Bellmullet y Claremorris al oeste; y por último Rosslare formo su propio grupo y esta ubicada al este y en la costa. Resulta interesante que Haslett, John y Raftery, A (1989) en su artículo eliminan de su análisis a la ciudad de Rosslare debido a que su correlación con las demás estaciones es muy baja y por lo tanto no tiene influencia sobre las demás estaciones, esto se ve reflejado en este análisis al no agruparse con las demás estaciones.

6 Conclusión

La medida de distancia entre series propuesta en este trabajo toma en cuenta la estructura de correlaciones entre las series, por lo tanto es común que los grupos formados no tengan semejanza a simple vista. Esto se debe a que dos series que provengan del mismo modelo $ARIMA$ pueden ser muy distintas cuando se grafican.

Es recomendable pero no necesario estacionalizar las series, ya que con series no estacionarias se requiere un orden de p muy grande para recoger toda la estructura de correlaciones mientras que, con series estacionarias, se pueden obtener modelos que cumplan con el principio de parsimonia.

El algoritmo empleado para estimar los coeficientes del modelo $AR(p)$ fue el de mínimos cuadrados, ya que comparado con otros métodos como el de máxima verosimilitud y Yule-Walker no se presentó ninguna diferencia en los resultados de los ejemplos anteriores. Se recomienda entonces utilizar MC por su sencillez cuando se tengan suficientes datos y cuando no sea así es recomendable entonces utilizar la metodología propuesta por Piccolo (1990), que consiste en obtener primero un modelo $ARIMA$ para una serie y después obtener los coeficientes de π.

El procedimiento propuesto en este trabajo es muy sencillo de implementar por que solamente es necesario tener una idea general del tipo de modelos $ARIMA$ con los cuales

las series serían modeladas de una manera adecuada, y con esto proponer un modelo $AR(p)$ que sea lo suficientemente grande para aproximar de una manera adecuada a $\boldsymbol{\pi}$.

Para aplicar esta técnica no es necesario que las series tengan la misma longitud, ya que sólo es necesario estimar un modelo $AR(p)$, por lo tanto lo único que importa es que contemos con suficientes datos para tener una buena estimación de ese modelo.

Habiendo decidido el orden p, a través de mínimos cuadrados ordinarios se pueden estimar los coeficientes del $AR(p)$, el cálculo de la matriz de distancias es trivial, entonces cualquier técnica de agrupamiento que utilice una matriz de disimilitudes es aplicable.

En este trabajo el objetivo fue definir una medida de disimilitud entre series de tiempo para así poder agruparlas, pero esta medida puede ser utilizada en otros procedimientos que necesiten el concepto de medidas entre series, por ejemplo, escalamiento multidimensional

7 Bibliografía

Bhattacharya (1943). "On a Measure of Divergence Between Two Statistical Populations" *Bulletin of the Calcutta Mathematical Society,* 35, 99-109.

Box. G. E. P., Jenkins. G. M. y Reinsel. G. C. (1994). *Time Series Analysis*, Prentice Hall.

Chernoff (1952). "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of the Observations", *Annals of Mathematical Statistics, 25*, 573-578.

Díaz-García J. A. y González-Farías, G. (2001). "A Note on the Cook's Distance", *Enviado para su publicación.*

Domínguez-Molina J.R (2001). "Formación de Conglomerados y valores atípicos en Series de Tiempo". *Tesis no publicada.* Maestría en Estadística. Programa conjunto CIMAT- Universidad de Guanajuato.

Fuller. W. A. (1996) . *Introduction to Statistical Time Series,* John Wiley & Sons.

Hamilton. J. D. (1994) . *Time Series Analysis,* Princenton University Press.

Haslett, J. y Raftery, A (1989). "Space-Time Modelling with Long-Memory Dependence: Assessing Ireland's Wind Power Resource" *Applied Statistics, 38, 1-50.*

Johnson. R. A. y Wichern. D. W. (1992) . *Applied Multivariate Statistical Analysis*, Prentice Hall.

Kakizawa, Y., Shumway. R. H. y Taniguchi. M. (1988) . "Discrimination and Clustering for Multivariate Time Series", *Journal of the American Statistical Association, 93, 328-339.*

Kullback, S. and Leibler, R. A (1951). "On Information and Sufficiency", *Annals of Mathematical Statistics, 22*, 79-86.

McLachlan (1992), *Discriminant Analysis and Statistical Pattern Recognition,* New York, Wiley.

Piccolo. D. (1990) . "A Distance Measure for Classifying ARIMA Models", *Journal of Time Series Analysis, 11, 154-164.*

Wei. W. W. S.(1990) . *Time Series Analysis*, Addison-Wesley.

Shumway. R. H. (1988) . *Applied Statistical Time Series Analysis,* Princenton Hall.