ÍNDICE DE COMOVIMIENTO ENTRE SERIES TEMPORALES: UNA APLICACIÓN

Por
Juan Carlos Herrera Órdenes

REQUIRIMIENTO PARA EL TITULO DE
INGENIERO EN ESTADÍSTICA
EN
UNIVERSIDAD DE VALPARAÍSO
VALPARAÍSO,CHILE
POR DEFINIR JPT DICIEMBRE 2008

UNIVERSIDAD DE VALPARAÍSO
DEPARTAMENTO DE ESTADÍSTICA

Por medio de esto certifico que han leído y recomiendan a la Facultad de de Ciencias para la aceptación de mi Proyecto de Título en el cumplimiento parcial "**Índice de comovimiento entre series temporales: Una Aplicación**" por **Juan Carlos Herrera Órdenes** los requisitos para el Título de **Ingeniero en Estadística**.

Fecha: Por definir jpt Diciembre 2008

Supervisor de Proyecto de Título: _____
Dr Ronny Vallejos

Comité Examinador: _____
Por Definir.

_____
Por Definir.

UNIVERSIDAD DE VALPARAÍSO

Fecha: **Por definir jpt Diciembre 2008**

Autor: **Juan Carlos Herrera Órdenes**

Título: **Índice de comovimiento entre series temporales: Una Aplicación**

Departamento: **de Estadística**

Grado: **Licenciado en Estadística** Convocación: **Por definir** Año: **2008**

Universidad de Valparaíso

Firma del Autor

Padre y Madre.

Tabla de Contenidos

Resumen

Para dos series temporales el Índice de Comovimiento es una medida de similaridad que contiene la información temporal entre ellas. Este índice es también llamado Coeficiente de Codispersión, el cual, es una adecuada normalización de suma de productos internos para dos secuencias temporales. De acuerdo a su definición, dos series conmueven (o se mueven conjuntamente), si sus conjuntos respectivos de pendientes son proporcionales entre sí. Este índice tiene como característica estar acotado entre 1 y -1. Si el coeficiente o índice es cercano a 1, se dice que las dos series se mueven juntas (comueven), en cualquier intervalo de tiempo $[t_i, t_{i+h}]$, donde h es el retardo del ndice en el proceso. Un coeficiente cercano a -1, se interpreta como un anti comovimiento. Ahora cuando este índice es cercano 0 se puede decir que no hay comovimiento entre las series, es decir, no hay relación entre las pendientes de las series en instantes sucesivos. Este trabajo se enfoca en dos puntos esenciales. Primero, representar varias situaciones con modelos paramétricos asociado a esta medida. Segundo, aplicar un algoritmo de clasificación para series temporales basado en una medida de asociación que contiene el Índice de Comovimiento, llamado índice de disimilaridad adaptativo.
El índice de disimilaridad adaptativo, es un producto entre dos funciones, que contiene una función de afinacin de balance entre el comportamiento respecto del comovimiento entre las series temporales y la cercanía de los valores de basados en distancias convencionales. De esta manera se introduce una nueva medida alternativa para la clasificación de series temporales utilizando los algoritmos clásicos de clasificación como son por ejemplo el método de agrupación jerárquico.
Ahora bien, estas medidas se aplicaran a 7 AFP del sistema de pensiones Chileno. Donde, este sistema de pensiones Chileno, es un modelo que ha sido exportado a otros países, donde su funcionalidad es velar por los ahorro de los trabajadores para tener una futura pensión cuando llegue su jubilación. La característica principal de las AFP es su rentabilidad, existen empresas especializadas para lograr la rentabilidad de los ahorros de los chilenos. No obstante, debido a la facilidad de información de los mercados bursátiles, las AFP buscan oportunidades en ella. Esto ha creado que la competencia de las empresas de AFP, no presente grandes variabilidades respecto a las otras AFP, lo que comúnmente se llama fenómeno manada.
Los resultados que se presentan en este trabajo, son la aplicación de estas medidas a estas 7 AFP. Para así, agruparlas considerando su información de comovimiento y comportamiento respecto a su cercanía simultáneamente.

Agradecimiento

Por definir

Introducción

Ha existido gran interés por estudiar el comportamiento de una variable aleatoria a través del tiempo como es el caso en series de tiempo, en que una función del pasado es predicha en el futuro. Las aplicaciones de series de tiempo o series temporales es muy amplia, esto se puede apreciar en la Economía, Medicina, Meteorología, etc. Pero es natural pensar o estudiar, como se comportan dos o más series de tiempo. En la actualidad existen varios modelos multivariados para modelar esta situación, un ejemplo a nombrar: Peña (2001) realiza simulaciones de Monte Carlo, donde compara los modelos ARIMA y VARMA, y muestra cuando la dependencia entre las componentes de un vector de series de tiempo, hace crecer la precisión de los pronósticos multivariados respecto a los univariados. También existe el coeficiente de correlación espúrea, Karl Pearson (1897) que dice que un alto coeficiente de correlación entre dos variables es espúreo si este se explica por la presencia de un tercer factor y no debido a la existencia de una relación con sentido entre las variables analizadas. En este caso, la correlación estadísticamente significativa entre las variables es una correlación espúrea o sin sentido. Por nombrar algunos. La motivación de este Proyecto de Titulación, es estudiar el comportamiento dos series de tiempos (Procesos estocásticos) $\{X_t\}$ e $\{Y_t\}$ con un coeficiente de codispersión o comovimiento introducido por (Rukhin y Vallejos, 2006). En esencia este coeficiente se implemento para procesos espaciales autoregresivos y de media móvil intrínsicamente estacionarios. En este Proyecto de Titulación se particularizará a modelos y/o procesos unilaterales autoregresivos, de media móvil y algunos modelos ARMA. Este coeficiente tiene la ventaja de captar el comportamiento de dos series temporales y es una versión corregida de otro coeficiente, usado en Estadística Espacial. Este coeficiente compara proporcionalmente las pendientes en común de pares de puntos, a través, del tiempo. A modo de contraejemplo, consideremos la covarianza muestral de dos variables en estudio

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

donde este es un estimador crudo, que depende de la suma de los productos cruzados. La covarianza muestral permite identificar la dirección o sentido de la relación lineal entre variables, a través de su signo y esto nos permite establecer en que cuadrante se encuentran los datos. Esta es la única información relevante que proporciona la covarianza muestral.

En la literatura se pueden encontrar otras medidas de asociación, un ejemplo es el Coeficiente de correlación de Spearman, como una versión no paramétrica. Sin embargo, este Coeficiente no da la información sobre el comportamiento temporal entre las series, más bien, están orientadas a proveer la independencia de dos series (Ver Yong y Schreckengost, 1981).
Más adelante en el Capítulo I se harán definiciones formales y fundamentos de este coeficiente, se hará una introducción a algunas medidas de similaridad, algunos métodos no parámetritos de este coeficiente, se profundizará más sobre este coeficiente de codispersión, se mostraran algunas propiedades de este coeficiente, resultados importantes y limitaciones teóricas del mismo . En el Capítulo II, Se hará una reseña a los métodos de agrupación para series temporales introduciendo un índice de disimilaridad adaptativo, además se estudiara este índice, el cual es una función de balance, que contiene información del comovimiento entre las series temporales y el comportamiento respecto a la distancia, introduciendo así una nueva medida para la clasificación de las series temporales. En el Capítulo III se realizaran simulaciones, para entender y gráficas de manera más clara, las características y uso de este índice de disimilaridad adaptativo. se compararan los coeficientes anteriores para modelos ARMA. Se comparara el índice de similaridad adaptativo $D(X_t\ ,\ Y_t)$ con diferentes funciones. Se estudiará las desviaciones de la estacionalidad, se buscará analizar la robustez del coeficiente. Por último, en el Capítulo IV se hará una aplicación a datos reales, sacados del Sistema de Pensión Chileno. En esta parte se hará una breve intrroducción a Sistema Chileno de AFP, se hablara de la genesis de este sistema y algo sobre la nueva reforma de Previsión Social, también ver como este modelo se a exportado a otros países, se mencionara las ventajas, por ejemplo como una forma de ahorro a futuro y desventajas de este sistema de Pensión como el **Efecto Manada**,estos datos fueron fueron sacados de la base de datos"PONER DIRECCION", la unidad de análisis de esta base de datos es la rentabilidad mensual de 7 AFP en estudio y finalmente se hará uso y aplicará toda la metodología mencionada con sus respectivas conclusiones.

0.1. Objetivos del Proyecto

0.1.1. Objetivos Generales

El coeficiente de codispersin fue introducido por Matheron en el ao 1965, como una extensión del semivariograma para procesos espaciales intrnsecamente estacionarios.
Los avances de este coeficiente se pueden encontrar en la minería, procesamiento de imágenes y geoestadística entre otras. En este Proyecto de titulación se particularizará esta teoría para modelos unilaterales, autoregresivos, de media móvil y ARMA. Basado en fundamentos matemáticos de probabilidades, Inferencia y Series temporales, que ayudaran a sustentar este proyecto.
Para esto es necesario tener medidas o índices que resuman toda esta información en un solo nmero. Problema que abordara este proyecto de titulo.
Ahora bien, dependiendo de la perspectiva que se plantee, en la literatura se puede encontrar muchas

de clasificar y medir la similitud, por ejemplo la distancia Euclidiana. Por otra parte. Warren Liao (2005), hace una resea de varias medidas de asociación y medidas de similaridad para secuencias temporales y algoritmos, para aplicar en Cluster.
Por otra parte Ahlame Douzal Chouakria y Panduranga Naidu Nagabhushan(2007) proponen un Índice de Disimilaridad Adaptativo para medidas de proximidad en series temporales, la cual se llama automatic adaptive tuning function.
Rukhin y Vallejos (2006) introducen un coeficiente de similaridad para secuencias Espaciales y Temporales, donde este coeficiente, es una normalización de suma de incrementos para secuencias de tiempo o espacio.
También, revisaremos algunas medidas ms usadas de asociación y similaridad. De la misma forma, se verán algunas definiciones básicas de procesos estocástico, con algunas hipótesis que sustentan este Proyecto de Titulaci´n, como es la estacionalidad de las series temporales.
Planteando también la lógica del Coeficiente de Codispersin o Índice de Comovimiento y su interpretacin.
Seguidamente, se hará una conjunción entre el Índice de Disimilaridad Adaptativo y el Coeficiente de Codispersión, para así aplicar este Índice de Disimilaridad en algunos métodos de clasificación, con el cual se trabajara para la clasificación de series temporales, o Cluster el cual se aplicara al sistema chileno de AFP.

0.1.2. Objetivos Específicos

1. Estudiar el tema del coeficiente de Codispersión.
2. Aplicar el coeficiente de codispersión al sistema de AFP Chileno.
3. Implementar un algoritmo de clasificacin para un conjunto de series de tiempo, de datos reales del sistema de AFP chileno..

0.2. Hipótesis

1. Los modelos ARMA son modelos expresivos para representar una diversidad de escenarios reales.
2. Las series que se analizaran deben ser estacionarias.
3. La normalidad asinttica del coeficiente de similaridad se cumple.
4. Es una herramienta til para la clasificacin de series temporales.

Capítulo 1

Teoría y Fundamentos

Para dos o más secuencias temporales, un aspecto importante a considerar, nace de la siguiente pregunta Cómo resumir esta información de manera simple y compacta?. Para esto es necesario tener medidas o índices que resuman toda esta información en un sólo número. Dependiendo de la perspectiva que se planteé, en la literatura se puede encontrar muchas formas de plantear, por ejemplo la distancia Euclidiana. Por otra parte,T. Warren Liao (2005), hace una reseña de varias medidas de asociación y medidas de similaridad para secuencias temporales y algoritmos, para aplicar en Cluster. Por otra parte Ahlame Douzal Chouakria y Panduranga Naidu Nagabhushan(2007) proponen un índice de disimilaridad adaptativo para medidas de proximidad en series temporales, la cúal se llama *automatic adaptive tuning function*. Rukhin y Vallejos (2006) introducen un coeficiente de similaridad para secuencias Espaciales y Temporales, donde este coeficiente, es una normalización de suma de incrementos para secuencias de tiempo o espacio. En este Cápitulo revisaremos algunas medidas más usadas de asociación y similaridad. También, se verán algunas definiciones básicas de procesos estocástico,con algunas hipótesis que sustentan este Proyecto de Titulación, como es la estacionalidad de las series temporales. Planteando también la lógica del coeficiente de codispersión para casos continuo diferenciable y su interpretación. Además se hará una una conjunción entre el índice de disimilaridad adaptativo y el coeficiente de codispersión donde este es un caso general de la estructura principal de este índice adaptativo, para así implementar un nuevo índice de disimilaridad con el cual se trabajará para la clasificación de series temporales, el cual se verá con más detalles en el Capítulo II.

1.1. Definición de procesos Estocásticos.

El coeficiente de Codispersión o de Comovimiento tiene la capacidad de captar el comportamiento respecto a si dos series se mueven juntas en el tiempo, en varias ocasiones esto se puede modelar a través de modelos paramétricos que están asociados a esta medida de Codispersión. Primero es

necesario definir algunos objetivos y estructuras que nos ayudarán a describir diferentes escenarios. Comenzaremos por la defínción de Procesos Estocásticos.
Sea $\{X_t\}$ una función,

$$\begin{array}{rccc} X: & \Omega \times T & \longrightarrow & \mathbb{R} \\ & (\omega, t) & \mapsto & X(\omega, t) \end{array}$$

tal que para cada $t \in T$, $X_t(\omega)$ es una variable aleatoria. Llamamos al conjuno $\{X_t, t \in T\}$ Proceso Estocástico sobre el espacio Ω.
Una secuencia $\{X_t, t \in T\}$ es fuertemente o estrictamente estacionario si $\{X_{t_1}, \ldots, X_{t_n}\}$= $\{X_{t_1+h}, \ldots, X_{t_n+h}\}$ en distribución, para toda colección $t_1, ..., t_n$ y $h \in T$. Es decir, es invariante bajo traslación.
Por otra parte, una secuencia es débilmente, o de segundo orden estacionario si:

1. $\mathbb{E}(X_t) = \mu$.
2. $\mathbb{V}(X_t) = \sigma^2$.
3. $\mathbb{C}ov(X_t, X_{t+k}) = \gamma_k$. Es una función que depende únicamente de la distancia entre t y $t + k$.

Entonces la secuencia $\{\gamma_k, k \in T\}$ es llamada función de autocovarianza.
Seguidamente se define:
$\rho_k = \gamma_k / \gamma_0$ y $\{\rho_k, k \in T\}$ es llamada función de autocorrelación (ACF).

Como caso particular cuando $T = \mathbb{Z}$, enteros, se está en presencia de Series temporales.
Entonces las series de tiempo es un caso particular de los Procesos Estocásticos tema que ayudará como base para el índice de comovimiento. Ahora, en esta sección se darán algunas definiciones de modelos y/o procesos autoregresivos, de media movíl y Procesos ARMA.

1.2. Ruido blanco

La secuencia ϵ_t, consiste de variables aleatorias independientes con media 0 y varianza σ^2. Es llamada ruido blanco. Esta serie es estacionario de segundo orden con γ_0=σ^2 y γ_k=0 ,k $\neq$ 0.

1.3. Procesos Autoregresivos.

Sea $\{X_t, t \in Z\}$ un Proceso Estocástico. Se dice que $\{X_t\}$ es un modelo autoregresivo de orden p denotado por AR(p),si puede ser escrito por la ecuación:

$$X_t = \sum_{r=1}^{p} \phi_r X_{t-r} + \epsilon_t, \tag{1.3.1}$$

donde ϕ_r son los parámetros del modelo y ϵ_t es un ruido blanco independiente de media 0 y varianza σ^2. Equivalentemente, X_t puede ser denotado por

$$\Phi(B)X_t = \epsilon_t \tag{1.3.2}$$

donde

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p, \text{ con } B^k X_t = X_{t-k}. \tag{1.3.3}$$

Un proceso AR(p) es estacionario si y solo si, las raíces de $\Phi(B) = 0$, están fuera del disco unitario. En tal caso, el proceso X_t puede ser representado por un modelo de media movil infinito (MA(∞)).

$$X_t = \sum_{j=0}^{\infty} \Psi_j \epsilon_{t-j}. \tag{1.3.4}$$

Ejemplo: El proceso AR(1) está definido:

$$X_t = \phi_1 X_{t-1} + \epsilon_t,$$

donde $|\phi| < 1$ garantiza la estacionalidad del proceso.

1.4. Procesos de Media Móvil.

Sea $\{X_t, t \in Z\}$ un Proceso Estocástico. Se dice que X_t es un modelo de media móvil de orden q denotado por MA(q) si puede ser descrito por la ecuación:

$$X_t = \sum_{s=0}^{q} \theta_r \epsilon_{t-s}, \tag{1.4.1}$$

donde los θ_r son los parámetros del modelo y ϵ_t es un ruido blanco independiente con media 0 y varianza σ^2. Equivalentemente, X_t puede ser denotado por

$$\Theta(B)\epsilon_t = X_t, \tag{1.4.2}$$

donde,

$$\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q, \text{ con } B^k \epsilon_t = \epsilon_{t-k}. \tag{1.4.3}$$

Si las raíces de $\Theta(B) = 0$, están fuera del disco unitario, entonces el proceso es invertible. En tal caso:

$$\epsilon_t = \sum_{j=0}^{\infty} \Psi'_j X_{t-j}, \text{ es un proceso AR}(\infty). \tag{1.4.4}$$

1.5. Procesos ARMA

Los procesos Autoregresivos de media móvil, ARMA(p,q) están definidos por:

$$X_t - \sum_{r=1}^{p} \phi_r X_{t-r} = \sum_{s=0}^{q} \theta_r \epsilon_{t-s} \Leftrightarrow \Phi X_t = \Theta \epsilon_t. \tag{1.5.1}$$

donde ϵ_t es un ruido blanco. El proceso X_t es estacionario para apropiados valores de ϕ y θ.

1. Teorema: Sea $\{X_t, t \in Z\}$ un modelo ARMA. $\{X_t, t \in Z\}$ es estacionario si y solo si las raíces del polinomio $\Phi(B) = 0$ estan fuera del disco unitario. En tal caso.

   $$X_t = \Phi(B)^{-1}\Theta(B)\epsilon_t.$$

   Representación causal de $\{X_t\}$ (No depende del futuro). Representación $MA(\infty)$ asociada al proceso $\{X_t\}$.
   Obs: $\epsilon_t = \Theta(B)^{-1}\Phi(B)$ invertible con las raíces del polinomio $\Theta(B)$ estan fuera del disco unitario.

Entonces un proceso ARMA(p,q) es un proceso más general. Con esto, se puede dejar más claro cuales pueden ser los valores apropiados de ϕ y θ

1.6. Coeficientes de Codispersión para series temporales.

El coeficiente de codispersión fue introducido por Matheron en el año 1965, como una extensión del semivariograma para procesos espaciales intrínsecamente estacionarios.
Sean $\{X_t\}$ y $\{Y_t\}$ dos procesos débilmente estacionarios. El variograma cruzado está definido por:

$$\gamma(h) \quad = \quad \mathbb{E}[X_{s+h} - X_s][Y_{s+h} - Y_s] \tag{1.6.1}$$

tal que s y $s + h \in Z$. Entonces el coeficiente de codispersión es

$$\rho_{X_t,Y_t}(h) \quad = \quad \gamma(h)/\sqrt{\mathbb{V}_x(h)\mathbb{V}_y(h)} \tag{1.6.2}$$

donde

$$\mathbb{V}_x(h) = \mathbb{E}[X_{s+h} - X_s]^2$$

Se puede apreciar que (1.6.2) es una normalización del semivariograma y como consecuencia este coeficiente esta acotado en valor absoluto por 1,es decir, $|\rho_{X_tY_t}(h)| < 1$. La demostracion consiste en usar la desigualdad de Cauchy-Shwartz,

$$\mathbb{E}[XY] \leq \left[\mathbb{E}[X^2]\right]^{1/2} \left[\mathbb{E}[Y^2]\right]^{1/2}. \tag{1.6.3}$$

Ahora, consideremos dos procesos lineales generales de la forma:

$$X_t = \sum_{j=0}^{\infty} \phi_j \epsilon_1(t-j) \tag{1.6.4}$$

$$Y_t = \sum_{k=0}^{\infty} \theta_k \epsilon_2(t-k) \tag{1.6.5}$$

tal que $\sum_{j=0}^{\infty} \phi_j < \infty$, $\sum_{k=0}^{\infty} \theta_k < \infty$ y ϵ_1 y ϵ_2 son ruidos blancos con media 0 y varianza σ^2 y τ^2 respectivamente.
Además, consideremos la estructura de covarianza para los errores dada por

$$\mathbb{C}ov(\epsilon_1(t), \epsilon_2(s)) = \begin{cases} \rho\tau\sigma & \text{si } t = s \\ 0 & \text{eoc} \end{cases} . \tag{1.6.6}$$

En base a estos modelos se trabajara más en la sección de propiedades, donde estos modelos estarán asociados a nuestro Índice de Comovimiento, encontrando una forma explícita de nuestro coeficiente.
Hay varios procesos que satisfacen la forma lineal general, por ejemplo para procesos autoregresivos de orden 1 y procesos de media móvil de orden 1. Sin embargo, para procesos ARMA de orden superior no hay una forma explícita (análitica para $\rho_{X_t,Y_t}(h)$).
Un estimador de momentos natural para $\rho_{X_t,Y_t}(h)$ es el coeficiente de codispersión muestral para dos procesos como en (1.6.2).

$$\widehat{\rho}_{X_t,Y_t}(h) = \frac{\sum_{i=1}^{n-h}[X_{i+h} - X_i][Y_{i+h} - Y_i]}{\sqrt{\sum_{i=1}^{n-h}[X_{i+h} - X_i]^2 \sum_{i=1}^{n-h}[Y_{i+h} - Y_i]^2}} \tag{1.6.7}$$

Este coeficiente tiene la ventaja de captar el comportamiento de dos series temporales, en donde se compara proporcionalmente las pendientes en común de pares de puntos. Además, la estructura de correlación depende únicamente de los datos, dejando de lado tener que depender de algún estadístico. La codispersión definida en (1.6.2), corresponde a productos internos normalizados de primera diferencia para las secuencias $\{X_t\}$ e $\{Y_t\}$.

$$cm_{X_t,Y_t} = \rho(1,0) = \frac{\sum \Delta X \Delta Y}{\sqrt{[\sum \Delta X]^2 [\sum \Delta Y]^2}}$$

El coeficiente cm_{X_t,Y_t}, es un coeficiente geométricamente natural de comovimiento en que compara proporcionalmente pendientes de pares de en puntos en común.
En esta sección se examinará las bondades y características de las series de tiempo que se mueven conjuntamente, intuitivamente podemos decir que dos curvas comueven si sus conjunto de pendientes son proporcionales o casi proporcionales.

Para dos procesos diferenciables estacionarios X_t e Y_t el coeficiente de comovimiento esta definido por:

$$cm_{X_t,Y_t} = \frac{\mathbb{E}(X_t{}')\mathbb{E}(Y_t{}')}{\sqrt{\mathbb{V}(X_t{}')\mathbb{V}(Y_t{}')}} \tag{1.6.8}$$

donde se asume que $\mathbb{E}(X_t{}')<\infty$ y $\mathbb{E}(Y_t{}')<\infty$.
Para dos secuencias X_t y Y_t, el coeficiente muestral es el siguiente.

$$\widehat{cm}_{X_t,Y_t} = \frac{\sum_{t=1}^{n-1}[X_{t+1}-X_t][Y_{t+1}-Y_t]}{\sqrt{\sum_{t=1}^{n-1}[X_{t+1}-X_t]^2\sum_{t=1}^{n-1}[Y_{t+1}-Y_t]^2}} \tag{1.6.9}$$

En la forma integral para curvas suavizadas X_t y Y_t,el estadístico muestral es definido como:

$$\widehat{cm}_{X_t,Y_t} = \frac{\int X_t{}'dt \int Y_t{}'dt}{\sqrt{[\int X_t{}'dt]^2[\int Y_t{}'dt]^2}} \tag{1.6.10}$$

Como esta definido el coeficiente de comovimiento, no es el coeficiente de correlación entre las primeras derivadas. Debido a la naturaleza, local del comoviemiento en el numerador y denominador del coeficiente , la $\mathbb{E}(X_t{}')$ y $\mathbb{E}(Y_t{}')$ no son retardos, as para eliminarla. De hecho seria indeseable la resta de estos valores esperados (Como es el hecho, por instancias en el caso de la covarianza y correlación típica). Considere el ejemplo de dos lneas rectas con pendientes positivas,donde el valor del coeficiente es igual a 1, pero donde la recta de las medias lleva una expresión indeterminada de la forma 0/0. Una manera alternativa es pensar que la primera diferenciación, ya a logrado la restas de las medias.
Mientras que el coeficiente de comovimiento es definida, aqu no es correlacionado (De la primera diferenciación o tangentes,secuencias) comparte un número de propiedades estndar de coeficiente de correlación.
Es claro mostrar que el coeficiente de comovimiento y sus formas muestrales son invariante bajo traslación, positivamente homogneo, simétrico en sus argumentos, definido positivo para una secuencia y versiones con log para si mismo, e interpretable como el coseno del ngulo entre vectores formados por la primera diferencia de las series muestrales mostradas.
Note que, no hay nada en la definición en el estadístico de comovimiento que requiere que las secuencias sean estimadas, sean muestreadas equiespaciademente, como es usualmente en el caso del estadístico de serie de tiempo. La primera diferencia se puede lograr sin la uniformidad o continuidad de espacio entre observaciones. Lo que es requerido por la definición, es que todas los observaciones de las secuencias sean iguales en tiempo. Es decir, los datos muestreados en los procesos tienen que estar en los instantes $t_1, \ldots, t_n$.

1.7. Algunos ejemplos.

Conforme con las definiciones anteriores, se presentara gráficamente como se comportan las series para que el Coeficiente de Codispersión capte su grado de comovimiento.
Si bien el Coeficiente de Codispersión fluctúa entre -1 y 1, en esta sección se mostrarán los casos mas significativos, como cuando dos series temporales tienen comovimiento 1, -1y 0.

1.7.1. Cuando $\hat{\rho}_{X_t,Y_t} \approx 1$

En este gráfico se puede apreciar como las dos series presentan un comportamiento parecido. En la practica cuando uno observa este tipo comportamientos entre dos series, es natural pensar que las series se mueven juntas, pero no se sabe cuanto.
Por eso, es necesario cuantificar este grado de comovimiento, entonces si utilizamos el estimador muestral definido anteriormente se puede observar un $\rho = 0{,}999...$

Figura 1.1: Valores de $\hat{\rho}_{X_t,Y_t} \approx 1$.

Cuando se habla del comportamiento de las series de tiempo, se está en presencia de estructuras dinámicas que no están fijas, como en el caso de las variables aleatorias, ya que se ha calculado que grado de comovimiento tienen, entonces de acuerdo a su definición un $\hat{\rho}_{X_t,Y_t}$ cercano a 1, se puede interpretar que las series se mueven juntas en cualquier período de tiempo $[t_i, t_{i+h}]$.

1.7.2. Cuando $\hat{\rho}_{X_t,Y_t} \approx -1$

En este gráfico se puede apreciar un anti comovimiento, es decir, que sus conjuntos de pares de pendientes son inversamente proporcional.
Se puede ver como las series tienen un comportamiento inverso respecto del otro. Si calculamos su grado de comovimiento con el estimador muestral este es -0.999. Es decir, estas series se mueven distintas en cualquier periodo de tiempo $[t_i, t_{i+h}]$.

Figura 1.2: Valores de $\hat{\rho}_{X_t,Y_t} \approx -1$.

1.7.3. Cuando $\hat{\rho}_{X_t,Y_t} \approx 0$

En este gráfico se puede observar que el conjunto de pares de pendientes no tiene nada en común. Es decir, la series en cualquier périodo de tiempo se mueve distintas.

Figura 1.3: Valores de $\hat{\rho}_{X_t,Y_t} \approx 0$.

1.8. Propiedades.

En esta sección se darań a conocer algunos resultados principales para calcular ρ_{X_t,Y_t} para dos procesos lineales generales y la propiedad asintotica. Sean:

$$\begin{aligned} X_t &= \sum_{j=0}^{\infty} \phi_j \epsilon_1(t-j) \\ Y_t &= \sum_{k=0}^{\infty} \theta_k \epsilon_2(t-k) \end{aligned} \tag{1.8.1}$$

donde $\sum_{j=0}^{\infty} \phi_j^2 < \infty$, $\sum_{k=0}^{\infty} \theta_k^2 < \infty$. ϵ_1 y ϵ_2 son ruidos blancos con media 0 y varianza σ^2 y θ^2 respectivamente. Además, $\mathbb{C}ov(\epsilon_1(t), \epsilon_2(s)) = \begin{cases} \rho\tau\sigma & \text{si } t = s \\ 0 & \text{eoc} \end{cases}$.

Entonces

$$\mathbb{E}[X_{t+h} - X_t][Y_{t+h} - Y_t] = \mathbb{E}[X_{t+h}Y_{t+h}] - \mathbb{E}[X_{t+h}Y_t] - \mathbb{E}[X_tY_{t+h}] + \mathbb{E}[X_tY_t]$$

$$\begin{aligned} \mathbb{E}[X_{t+h}Y_{t+h}] &= \mathbb{E}\left[\left(\sum_{j=0}^{\infty} \phi_j \epsilon_1(t+h-j)\right)\left(\sum_{k=0}^{\infty} \theta_k \epsilon_2(t+h-k)\right)\right] \\ &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \phi_j\theta_k \mathbb{E}[\epsilon_1(t+h-j)\epsilon_2(t+h-k)] \end{aligned}$$

Si $t+h-j = t+h-k$, entonces $j = k$

$$\text{Por lo tanto } \mathbb{E}[X_{t+h}Y_{t+h}] = \sum_{j=0}^{\infty} \phi_j\theta_j\rho\tau\sigma$$

$$\begin{aligned} \mathbb{E}[X_{t+h}Y_t] &= \mathbb{E}\left[\left(\sum_{j=0}^{\infty} \phi_j \epsilon_1(t+h-j)\right)\left(\sum_{k=0}^{\infty} \theta_k \epsilon_2(t-k)\right)\right] \\ &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \phi_j\theta_k \mathbb{E}[\epsilon_1(t+h-j)\epsilon_2(t-k)] \end{aligned}$$

Si $t+h-j = t-k$, entonces $j = k+h$

$$\text{Luego } \mathbb{E}[X_{t+h}Y_t] = \sum_{j=0}^{\infty} \phi_{k+h}\theta_k\rho\tau\sigma.$$

$$\begin{aligned}
\text{Además,}\mathbb{E}[X_{t+h}Y_t] &= \sum_{j=0}^{\infty}\phi_{j+h}\theta_j\rho\tau\sigma \\
\text{Similarmente,} & \\
\mathbb{E}[X_tY_{t+h}] &= \mathbb{E}\left[\left(\sum_{j=0}^{\infty}\phi_j\epsilon_1(t-j)\right)\left(\sum_{k=0}^{\infty}\theta_k\epsilon_2(t+h-k)\right)\right] \\
&= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\phi_j\theta_k\mathbb{E}[\epsilon_1(t-j)\epsilon_2(t+h-k)] \\
\text{Si } t-j=t+h-k\text{, entonces } k=j+h & \\
\mathbb{E}[X_tY_{t+h}] &= \sum_{j=0}^{\infty}\phi_j\theta_{j+h}\rho\tau\sigma \\
\text{Finalmente,} & \\
\mathbb{E}[X_tY_t] &= \mathbb{E}\left[\left(\sum_{j=0}^{\infty}\phi_j\epsilon_1(t-j)\right)\left(\sum_{k=0}^{\infty}\theta_k\epsilon_2(t-k)\right)\right] \\
&= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\phi_j\theta_k\mathbb{E}[\epsilon_1(t-j)\epsilon_2(t-k)] \\
\text{Si } t-j=t-k\text{, entonces } k=j & \\
\therefore \mathbb{E}[X_tY_t] &= \sum_{j=0}^{\infty}\phi_j\theta_j\rho\tau\sigma \\
\text{Note que } \mathbb{E}[X_{t+h}Y_{t+h}] &= \mathbb{E}[X_tY_t]
\end{aligned}$$

Para el cálculo del denominador del coeficiente se tiene que

$$\begin{aligned}
\mathbb{E}[X_{t+h}-X_t]^2 &= \mathbb{V}ar[X_{t+h}-X_t] \\
&= \mathbb{V}ar[X_{t+h}]+\mathbb{V}ar[X_t]-2\mathbb{C}ov[X_{t+h}X_t]
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}ar[X_{t+h}] &= \mathbb{V}ar\left[\sum_{j=0}^{\infty}\phi_j\epsilon_1(t+h-j)\right] \\
&= \sum_{j=0}^{\infty}\phi_j^2\mathbb{V}ar[\epsilon_1(t+h-j)] \\
&= \sum_{j=0}^{\infty}\phi_j^2\sigma^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}ar[X_t] &= \mathbb{V}ar\left[\sum_{j=0}^{\infty}\phi_j\epsilon_1(t-j)\right] \\
&= \sum_{j=0}^{\infty}\phi_j^2\mathbb{V}ar[\epsilon_1(t-j)] \\
&= \sum_{j=0}^{\infty}\phi_j^2\sigma^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{C}ov[X_{t+h}X_t] &= \mathbb{E}[X_{t+h}X_t] - \mathbb{E}[X_{t+h}]\mathbb{E}[X_t] \\
&= \mathbb{E}[X_{t+h}X_t] \\
\mathbb{E}[X_{t+h}X_t] &= \mathbb{E}\left[\left(\sum_{j=0}^{\infty}\phi_j\epsilon_1(t+h-j)\right)\left(\sum_{k=0}^{\infty}\phi_k\epsilon_1(t-k)\right)\right] \\
&= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\phi_j\phi_k\mathbb{E}[\epsilon_1(t+h-j)\epsilon_1(t-k)]
\end{aligned}$$

Si $t+h-j=t-k$, entonces, $k=j+h$

$$= \sum_{j=0}^{\infty}\phi_j\phi_{j+h}\sigma^2$$

Similarmente, se tiene:

$$\begin{aligned}
\mathbb{E}[Y_{t+h}-Y_t]^2 &= \mathbb{V}ar[Y_{t+h}-Y_t] \\
&= \mathbb{V}ar[Y_{t+h}] + \mathbb{V}ar[Y_t] - 2\mathbb{C}ov[Y_{t+h}Y_t]
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}ar[Y_{t+h}] &= \mathbb{V}ar\left[\sum_{k=0}^{\infty}\theta_k\epsilon_2(t+h-k)\right] \\
&= \sum_{k=0}^{\infty}\theta_k^2\mathbb{V}ar[\epsilon_2(t+h-k)] \\
&= \sum_{k=0}^{\infty}\theta_k^2\tau^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}ar[Y_t] &= \mathbb{V}ar\left[\sum_{k=0}^{\infty}\theta_k\epsilon_2(t-k)\right] \\
&= \sum_{j=0}^{\infty}\theta_k^2\mathbb{V}ar[\epsilon_2(t-k)] \\
&= \sum_{k=0}^{\infty}\theta_k^2\tau^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{C}ov[Y_{t+h}Y_t] &= \mathbb{E}[Y_{t+h}Y_t] - \mathbb{E}[Y_{t+h}]\mathbb{E}[Y_t] \\
&= \mathbb{E}[Y_{t+h}Y_t] \\
\mathbb{E}[Y_{t+h}Y_t] &= \mathbb{E}\left[\left(\sum_{k=0}^{\infty}\theta_k\epsilon_2(t+h-k)\right)\left(\sum_{k=0}^{\infty}\theta_k\epsilon_2(t-k)\right)\right] \\
&= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\theta_k\theta_k\mathbb{E}[\epsilon_1(t+h-j)\epsilon_1(t-k)]
\end{aligned}$$

Si $t+h-j=t-k$, entonces, $k=j+h$

$$= \sum_{j=0}^{\infty}\theta_k\theta_{k+h}\tau^2$$

Ahora si juntamos los términos y factorizamos nos queda.

$$\rho_{X_t,Y_t}(h) = \frac{\rho\sum_{j=0}^{\infty}(2\phi_j\theta_j - \phi_{j+h}\theta_j - \phi_j\theta_{j+h})}{2\sqrt{\sum_{j=0}^{\infty}(\phi_j^2 - \phi_j\phi_{j+h})\sum_{j=0}^{\infty}(\theta_j^2 - \theta_j\theta_{j+h})}}.$$

Como caso particular:
Cuando $X_t \sim$ AR(1) y $Y_t \sim$ AR(1), los paramétros ϕ y θ satisfacen
Quedan como $\phi_j = \phi^j$,$|\phi| < 1$
y $\theta_j = \theta^j$, $|\theta| < 1$,$|\theta\phi| < 1$

Usando los siguientes identidades.

$$\begin{aligned}
\sum_{k=0}^{\infty}\theta^k\phi^k = \sum_{k=0}^{\infty}(\theta\phi)^k &= \frac{1}{1-\theta\phi}, \\
\sum_{k=0}^{\infty}\theta^k\phi^{k+h} = \phi^h\sum_{k=0}^{\infty}(\theta\phi)^k &= \frac{\phi^h}{1-\theta\phi}, \\
\sum_{k=0}^{\infty}\theta^k\phi^{k+h} = \phi^h\sum_{k=0}^{\infty}(\theta\phi)^k &= \frac{\phi^h}{1-\theta\phi}, \\
\sum_{k=0}^{\infty}\phi^{2k} &= \frac{1}{1-\phi^2}, \\
\sum_{k=0}^{\infty}\theta^{2k} &= \frac{1}{1-\theta^2}, \\
\sum_{k=0}^{\infty}\phi^k\phi^{k+h} = \phi^h\sum_{k=0}^{\infty}(\phi)^{2k} &= \frac{\phi^h}{1-\phi^2}, \\
\sum_{k=0}^{\infty}\theta^k\theta^{k+h} = \phi^h\sum_{k=0}^{\infty}(\theta)^{2k} &= \frac{\phi^h}{1-\theta^2},
\end{aligned}$$

tenemos

$$\rho_{X_t,Y_t}(h=1) = \frac{\rho(2-\phi-\theta)\sqrt{(1+\phi)(1+\theta)}}{1-\phi\theta}. \tag{1.8.2}$$

Este resultado fue obtenido por Rhukin y Vallejos (2008). Las propiedades asintótica de $\widehat{\rho}(h)$ fueron establecidos para el proceso $\mathbb{Z}_s = (\mathbb{X}_s, \mathbb{Y}_s)^t$ admitiendo la siguiente estructura,

$$\mathbb{Z}_{s+h} - \mathbb{Z}_s = \sum_l \mathbb{A}_l \epsilon_{s-l} \tag{1.8.3}$$

Donde $\mathbb{A}_l = \mathbb{A}_h(l)$ son matrices $2x2$ definida para todo l tal que, $\sum \|\mathbb{A}(l)\|^2 < \infty$. Donde $\|\cdot\|$ denota cualquier norma matricial y ϵ_t son vectores independientes con media cero y matriz de covarianza Σ.

- Teorema 1:
  Si el valor observado admite la representación del proceso $Z_t = (X_t, Y_t)^t$ (1,8,2) con matrices $A(k,l) = diag(\alpha_{kl}, \beta_{kl})$, la distribución limitante de $M[\rho_{X,Y}(\hat{h}) - \rho]$ es normalmente con media 0 y varianza

$$\nu^2 \quad = \quad \left(1 - \frac{\rho^2 \sum_{s=0}^{\infty} (\alpha_s \beta_s)^2}{\sum_{s=0}^{\infty} \alpha_s^2 \sum_{s=0}^{\infty} \beta_s^2}\right) \tag{1.8.4}$$

Con este resultado, se han realizados trabajos con Dócimas de hipótesis, intervalos de Confianza y Bandas de confianzas para el variograma, (Ver Rukhin y Vallejos, 2006).
Por otra parte se ha podido ver como las series temporales están asociadas a esta medida de Codispersión. La ventaja de encontrar una forma explicita para la varianza del coeficiente facilita los cálculos para dócimar y encontrar intervalos de Confianza ya que en caso contrario, habría que calcular la varianza, a través, de métodos de re-muestreo como Boostrap.

1.9. Limitaciones teorícas.

Dentro de las limitaciones teorícas se puede mencionar que el coeficiente de Codispersión, para modelos con muchos paramétros no existe una forma análitica para el coeficiente lo que lleva tener que implementar técnicas computacionales. Es obvio, mencionar que para series de tiempo no estacionarias el índice de codispersión no ha sido estudiado. Cuando una de las series presenta outliers, el coeficiente de codispersión es muy sensible. La definición de una version robusta del coeficiente de Codispesión aún es un problema abierto.

1.10. Métodos no parámetricos

En la literatura no existe un coeficiente de codispersión no parámetrico, debido a que este coeficiente es el estimador de momentos de $\delta_{X,Y}$, pero con esto, no se dice que no se pueda implementar un coeficiente de codispersión no paramétrico, pero esto motivaría otro proyecto que escapa a lo que se quiere mostrar en este Proyecto de Titulación. No obstante, si planteamos este coeficiente de codispersión para series espaciales, podemos encontrar que Tjϕstheim, quien propone una medida

de asociación para variables espaciales, en donde esta medida esta basada en los rangos de las observaciones y la localización de las coordenadas de los puntos medios. Además, en la literatura se pueden encontrar técnicas para estimar y predecir series de tiempo, los cuales dependen de una función de kernel y regresión no paramétrica.

Capítulo 2

Agrupamiento de series temporales.

2.1. Introducción.

El análisis por agrupación (Cluster analysis) es una técnica muy utilizada en problemas multivariados donde se quiere agrupar unidades experimentales con ciertas características o generalizar grupo de variables de interés. El objetivo de este Capítulo es definir algunas técnicas de agrupación. Las técnicas más usadas en este contexto son:

1. Métodos de agrupación no jerárquicas.

2. Métodos de agrupación jerárquicas.

En particular, los metodos de semejanza que usaremos están basadas en el Índice de Disimilaridad adaptativo mencionado en el Capítulo I. Si bien, el Índice de Disimilaridad es una medida alternativa a las medidas de semejanzas clásicas en análisis de grupos, la medida de semejanza que usaremos es una composición entre el coeficiente de Codispersión y alguna de las medidas usuales basado en alguna distancia.

2.2. Métodos de Agrupación.

1. **Métodos de agrupación no jerárquicas.** Se usan para agrupar objetos en un conjunto de k-cluster predeterminados, se parte de un conjunto inicial de cluster elegidos al azar luego se van cambiando de modo iterativo en la que habitualmente se usa el metodo de las k-medias.
   Métodos de las k-medias.
   Es un método que permite asignar a cada observación el cluster que se encuentra más proximo en términos el centroide(Valor medio de las observaciones de las variables en el valor del conglomerado).
   i) Se eligen al azar k-cluster iniciales.
   ii) Para el conjunto de observaciones se vuelve a calcular la distancia al centroide de los cluster y se reasignan a los que están mas próximos. Se vuelven a calcular los centroides de

los k-cluster después de las reasignaciones de los elementos.
iii) Se repiten los pasos anteriores hasta que no se produzca ninguna reasignación.

2. **Métodos de agrupación jerárquicas.**
   Metodos Divisivos.
   Se comienza con un gran conglomerado que contiene todas las observaciones u objeto en los pasos sucesivos en las observaciones que son más diferentes, se dividen y se construyen conglomerados más pequeños. Este proceso continua hasta que cada oservación es un mismo conglomerado,
   2.1 Método de encadenamiento simple (Vecino más cercano)
   Se basa en la distancia mínima, encuentra los dos objetos separados por la distancia más corta y las colocas en el primer conglomerado, a continuación se encuentra la distancia más corta o bien un tercer objeto. Se une a los dos primeros para formar un conglomerado o se forma un nuevo conglomerado de dos miembros. El proceso continua hasta que todos los objetos se encuentren en un conglomerado.
   2.2 Método de encadenamiento completo (Vecino más lejano)
   Este método es similar al interior excepto que el criterio de aglomeramiento se basa en la distancia máxima.

También, el uso de dendogramas es una herramienta útil cuando se quiere gráficar grupos. El método que se utilizará para clasificación es el de agrupación jerárquico, ya que, no se conocen la cantidad de grupos iniciales. En nuestro análisis usaremos el software R

2.3. Índice de similaridad adaptativo

En esta sección se presenta definiciones de algunas distancias más utilizadas en la literatura se pueden encontrar en T. Warren Liao (2005) en Clustering of Time Series data- an survey. Por otra parte, Chouakria y Nagabbhusham (2007) estudiaron un método para detectar la interdependencia de las series, dado que con medidas convencianales, no logran captar su dependencia temporal y es por eso que en esta sección se hará una pequeña descripción de estos métodos.

2.3.1. Distancia Euclidiana.

Sean $X_t = (x_1, \ldots, x_n)$ e $Y_t = (y_1, \ldots, y_n)$ dos series de tiempo de n valores observados en el tiempo en los instantes $t_1, \ldots, t_n$. La distancia Euclidiana δ_E entre X_t y X_t está definida como:

$$\delta_E = \sqrt{\sum_{i=i}^{n}(x_i - y_i)^2}. \tag{2.3.1}$$

2.3.2. Distnacia Minkowski

Para dos secuencias X_t e Y_t la distancia de Minkowski δ_M entre X_t e Y_t está definida como:

$$\delta_M = \sqrt[n]{\sum_{i=i}^{n}(x_i - y_i)^n}. \tag{2.3.2}$$

2.3.3. Distancia de Frechét

Se define una rejilla $r \in M$ entre dos series de tiempo $X_t = (x_1, x_2, \ldots, x_n)$ y $Y_t = (y_1, y_2, \ldots, y_k)$ como la secuencia de pares preservando el orden de la información.

$$r = ((x_{a_1}, y_{b_1}), (x_{a_2}, y_{b_2}), \ldots, (x_{a_n}, y_{b_m}))$$

con $a_i, i \in \{1, \ldots, n\}$,$b_j, j \in \{1, \ldots, m\}$ y satisfaciendo para $i \in \{1, \ldots, m-1\}$ la siguiente condición: $a_1 = 1$,$b_m = k$ $a_i = (a_i$ ó $a_{i+1})$ y $b_j = (b_j$ ó $b_{j+1})$.
Se define $|r| = \text{máx}_{i=1,\ldots,m} |x_{a_i} - y_{b_i}|$, la rejilla $|r|$ está representada por la máxima distancia entre dos observaciones apareadas.
La distancia $\delta_F(X_t, Y_t)$ se define como:

$$\delta_F(X_t, Y_t) = \min_{r \in M} |r| = \min_{r \in M} \left(\max_{i=1,\ldots,m} |x_{a_i} - y_{b_i}| \right) \tag{2.3.3}$$

2.4. De alineamiento de tiempo distorsionado (Dynamic time warping DTW)

El alineamiento de tiempo distorsionado (DTW) de tiempos es una técnica que encuentra la alineación óptima entre dos serie temporal si una vez la serie puede ser linealmente corrompido acelerando o desacelerando a lo largo de su linea de tiempo.
El alineamiento de tiempo distorsionado es a menudo usado en el reconocimiento del lenguaje hablado para determinar si dos formas de onda representan la de la misma forma locución hablada. Adentro una forma de onda de discurso, la duración de cada sonido hablado y el intervalo de en medio suena es admitido disentir, pero el mono las formas de onda de discurso deben ser similares. El alineamiento de tiempo distorsionado ha sido útil dentro de muchas otras disciplinas, incluyendo datos extrayendo de la cantera, la robtica, la manufactura, y la medicina. El alineamiento de tiempo distorsionado comúnmente usado en la minera de datos como una distancia que Mida entre la serie temporal. Un ejemplo de cómo una serie temporal está distorsionada para serie es mostrado en Figura.

□

Figura 2.1: Alineamiento de tiempo distorsionado

En Figura , cada línea vertical conecta un punto en una serie temporal para su punto correspondiente, similar en la otra serie temporal. Las líneas realmente tienen valores similares en el eje vertical pero han sido separado, así es que las lneas verticales entre ellos pueden ser miradas más fácilmente.
Si ambas series temporales en la figura fueran idénticas, todas las líneas serían líneas directamente verticales porque cada punto sería similar entre ellos.
La distancia de Alineamiento de tiempo distorsionado es una medida de la diferencia entre el dos puntos a la vez de las series después que ha sido distorsionada conjuntamente, lo cual está medido por la suma de las distancias entre cada par de puntos conectados.

2.4.1. Formulación del problema

El problema del Alineamiento de tiempo distorsionado es indicado como sigue:
Dado dos Series de tiempo $X_t = (x_1, \ldots, x_i, \ldots, x_m)$ e $Y_t = (y_1, \ldots, y_j, \ldots, y_n)$ de largo m y n respectivamente.
Note: que esta distancia puede ser calculada para series de distinta longitud.
Con estas series se construye el camino distorsionado que llamaremos W, esta es una secuencia de pares de puntos ordenados, basado en los índices de las series X_t e Y_t, donde:
$|r| = W = \{w_1, \ldots, w_K\}$, máx$\{m, n\} \leq K \leq n \cdot m$

Además,$K = m \cdot n$ es la longitud máxima del camino distorsionado. El camino distorsionado se define de la siguiente forma:
$w_k = (i, j)$, $k = 1, \ldots, K$
Donde i es el índice de la Serie X_t y j es el índice de la serie Y_t.
El camino distorsionado debe empezar desde el principio de cada tiempo la serie en $w_1 = (1, 1)$ y el final ambas series terminan en $w_K = (m, n)$.
Esto asegura que cada ndice de ambas serie temporales es usado en el camino distorsionado.
Hay tambin una restricción en el camino distorsionado obliga i y j a ser monotonamente creciente. Se define mas formalmente:

$$r = ((x_{a_1}, y_{b_1}), (x_{a_2}, y_{b_2}), \ldots, (x_{a_m}, y_{b_n})) = (w_1, w_2, \ldots, w_K)$$

con $a_i, i \in \{1, \ldots, n\}$,$b_j, j \in \{1, \ldots, n\}$ y satisfaciendo para $i \in \{1, \ldots, m-1\}$ la siguiente condición: $a_1 = 1$,$b_n = k$ $a_i = (a_i$ ó $a_{i+1})$ y $b_j = (b_j$ ó $b_{j+1})$.

Para poner en lnea usando DTW para dos secuencias de tiempo, construimos una matriz de $m \cdot n$ donde (i^{th}, j^{th}) es el elemento de la matriz contiene la distancia $d(x_{a_i}, y_{b_i})$ entre los dos puntos x_{a_i} e y_{b_i}, es decir,$d(x_{a_i}, y_{b_i}) = |x_{a_i} - y_{b_i}|$. Cada elemento matricial (i, j) corresponde para La alineacin entre el punto de x_{a_i} e y_{b_i}.

Este camino puede ser encontrado usando programación dinámica para evaluar la siguiente Recurrencia, que defina la distancia acumulativa $\gamma(i, j)$ como la distancia $d(i, j)$ encontrada en la celda actual y el mínimo de las distancias acumulativas de lo adyacente los elementos:

$$\gamma(i,j) \quad = \quad d(i,j) + min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\} \tag{2.4.1}$$

2.4.2. Hacia el Alineamiento de tiempo distorsionado

El Alineamiento de tiempo distorsionado es un algoritmo para medir la similitud entre dos secuencias que pueden variar en el tiempo o la velocidad. Por ejemplo, las similitudes en las pautas para caminar sería detectado, incluso si en un video la persona que caminaba lentamente y en otro si él o ella se camina más rápidamente, o incluso si hay aceleraciones y deceleraciones en el transcurso de una observación. DTW se ha aplicado a video, audio y grficos de hecho, cualquier dato que pueda ser convertido en una representación lineal se puede analizar con DTW. Una conocida ha sido la aplicación automática de reconocimiento de voz, para hacer frente a diferentes velocidades haciendo uso de la palabra.

En general, DTW es un método que permite a un ordenador para encontrar una óptima adecuación entre dos secuencias dada (por ejemplo, series de tiempo), con ciertas restricciones. Las secuencias son deformado no lineal en la dimensión de tiempo para determinar una medida de su similitud de algunos independientes no lineales variaciones en la dimensión temporal. Esta secuencia alineación método se usa a menudo en el contexto de modelos ocultos de Markov.

Un ejemplo de las restricciones impuestas a la adecuación de las secuencias se encuentra en la monotocidad de la cartografía en la dimensión temporal. La continuidad es menos importante en DTW que en otros algoritmos de Coincidencia de patrones; DTW es un algoritmo especialmente adecuado para las secuencias de concordancia con la información que falta, siempre hay tiempo suficiente para equiparar los segmentos que se produzca. El proceso de optimización se realiza utilizando programacin dinámica, de ahí el nombre.

La extensin del problema de dos dimensiones series como las imágenes (warping plana) es NP-completo, mientras que el problema de una dimensión como las seales de series de tiempo se pueden resolver en tiempo polinómico.

Ejemplo de una de las muchas formas del algoritmo

```
##########################################################################
####RUTINA PARA DETERMINAR EL CAMINO DISTROSIONADO#########################
  int DTWDistance (char s [1 .. n], char t [1 .. m], int d [1 .. n, 1 .. m]) (
      DTW declarar int [0 .. n, 0 .. m]
      declarar int i, j, el costo

      for i: = 1 to m
          DTW [0, i]: = infinito
```

```
    for i: = 1 hasta n
        DTW [i, 0]: = infinito
    DTW [0,0]: = 0

    for i: = 1 hasta n
        for j: = 1 to m
            coste: = d [s [i], t [j]]
            DTW [i, j]: = + costo mnimo (DTW [i-1, j], / / insercin
                                         DTW [i, j-1], / / supresin
                                         DTW [i-1, j-1]) / / partido

    DTW retorno [n, m]
)
##############################################################################
```

□

Figura 2.2: Camino distorsionado

2.4.3. Definición DTW

Considere una nueva definición de la rejilla propuesta en (2.3.3) como la suma de las distancias de todas las observaciones apareadas.

$$|r| \quad = \quad \sum_{i=1}^{m} |x_{a_i} - y_{b_i}| \tag{2.4.2}$$

Desde la definición anterior distancia de Fréchet se presenta la definición de Alineamiento de tiempo Distorsionado (Dynamic Time Warping) como una variante de la distancia de Fréchet, entonces se define la distancia $\delta_D TW(X_t, Y_t)$ como sigue:

$$\delta_{DTW}(X_t, Y_t) \quad = \quad \min |r| = \min_{r \in M} \left(\sum_{i=1}^{m} |x_{a_i} - y_{b_i}| \right). \tag{2.4.3}$$

2.4.4. Ejemplo

Sea $X_t = (1,0{,}34, 0{,}65, 2, 3, 3{,}4, 1, 1{,}2, 0{,}88, 5{,}9, 7)$ e $Y_t = (2, 5{,}5, 5{,}67, 3{,}45, 7, 3{,}3, 2{,}43, 1{,}34)$.
Calcular $\delta_{DTW}(X_t, Y_t)$.
Considere la siguiente rutina.

```
##############################################################################

library(dtw)
```

```
x<-c(1,0.34,0.65,2,3,3.4,1,1.2,0.88,5.9,7)
y<-c(2,5.5,5.67,3.45,7,3.3,2.43,1.34)
a<-dtw(x,y)
b<-a$distance
plot(a,xlab="indice de x",ylab="indice de y",main="Camino Distorcionado")
> length(x)
[1] 11
> length(y)
[1] 8
> b
[1] 26.45

##############################################################################
```

En la salida de esta rutina se puede apreciar que la longitud de vector X_t es de 11 elementos y la longitud del vector Y_t es 8 elementos. Por otra parte la distancia es 26.45.
La opción `plot(a)` en R-project muestra el siguiente gráfico

□

Figura 2.3: Camino distorsionado de $\delta_{DTW}(X_t, Y_t)$

Una de las característica de DTW, es la representación gráfica de su similitud entre series temporales y en este caso se puede apreciar que las series no presentan un gran grado de similitud, ya que estos puntos no definen una linea recta.
Desde aquí, se puede apreciar y entender la definición de camino distorsionado.
En estas sección se han presentado las distancias δ_E, δ_M, δ_F, y δ_{DTW} pero ellas ignoran la estructura temporal de los valores como proximidad, ya que, están basada sobre las diferencias entre los valores $|x_{a_i} - y_{b_i}|$.
En la siguiente sección se propondrá un Índice que considere la estructura de comovimiento que ayude a solucionar el problema de clasificación series temporales.

2.5. Índice de disimilaridad para medidas de proximidad en series de tiempo.

Las medidas convencionales de proximidad en series temporales, están basadas sobre la cercanía de los valores observados de las series de tiempo, la idea de está sección es presentar un Índice de Disimilaridad, que contenga la información de la codipersión y la cercanía de las series temporales. En otras palabras se define la similaridad entre dos series de tiempo, considerando estas dos característica, uno el comportamiento respecto a su comovimiento y la otra sobre su cercana. Generalmente,se asume que esta medida entre las secuencias X_t e Y_t toman valores positivos.
Uno puede cuantificar este concepto de similaridad considerando el coeficiente clásico de correlacioń de Pearson. Sin embargo, esta correlación guía una sobre estimación de la correlación. El coeficiente de correlación de Spearman puede también ser usado como una medida de similaridad. No obstante, se ha visto que los rangos de la correlación esta principalmente basado sobre los rangos y no sobre los valores observados. Dejando de lado la informacin respecto a su comovimiento.

2.5.1. Correlación temporal para medidas de proximidad.

Sea $X_t = (x_1, \ldots, x_n)$ una colección de n numeros reales independientes. La varianza clásica de X_t puede ser escrita en dos vias equivalentes.

$$Var(X_t) = \frac{1}{n-1}\sum_{i=i}^{n}(x_i - m)^2 = \frac{1}{n-1}\sum_{i,i'}(x_i - m)^2, \tag{2.5.1}$$

donde m es la media de los valores de X_t. Similarmente el coeficiente clásico de correlación entre $X_t = (x_1, \ldots, x_n)$ y $Y_t = (y_1, \ldots, y_n)$ puede ser escrita de dos formas:

$$Corr_{X_t,Y_t} = \frac{\sum_{i=i}^{n}(x_i - m_1)(y_i - m_2)}{\sqrt{Var(X_t)Var(Y_t)}} = \frac{\sum_{i,i'}(x_i - x_{i'})(y_i - y_{i'})}{\sqrt{\sum_{i,i'}(x_i - x_{i'})^2 \sum_{i,i'}(y_i - y_{i'})^2}}. \tag{2.5.2}$$

En el caso general de medidas independientes varianza/covarianza y coeficiente de correlación son medidas basadas sobre la contribución de todos los pares de medición. A la inversa en el caso de las interdependencias definiendo una relación entre vecino. La expresión de la varianza puede ser descompuesta como sigue.

$$Var(X_t) = \frac{1}{2n-1}\sum_{i,i'}(x_i - x_{i'})^2 \tag{2.5.3}$$

$$Var(S) = \frac{1}{2n-1}\sum_{i \text{ es vecino de } i'}(x_i - x_{i'})^2 + \frac{1}{2n-1}\sum_{i \text{ no es vecino de } i'}(x_i - x_{i'})^2 \tag{2.5.4}$$

La idea principal para incluir la informción sobre la dependencia, es una restricción a la expresión de la varianza/covarianza a los pares de valores dependientes (Vecinos).

$$VarT(X_t) = \frac{1}{2n-1}\sum_{i \text{ no es vecino de } i'}(x_i - x_{i'})^2 \tag{2.5.5}$$

Para la similaridad, se considerara una relación temporal de vecinos de primer orden medidas en el périodo $[t_i, t_{i+1}]$. Este coeficiente de correlación temporal es definido como:

$$\rho_{X_t,Y_t}(1) = \frac{\sum_{i=i}^{n-1}(x_{i+1} - x_i)(y_{i+1} - y_i)}{\sqrt{\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2 \sum_{i=i}^{n-1}(y_{i+1} - y_i)^2}}, \tag{2.5.6}$$

donde $\rho_{X_t,Y_t}(1)$ pertenece al intervalo $[-1, 1]$. Si valor $\rho_{X_t,Y_t}(1) = 1$, en el périodo observado $[t_i, t_{i+1}]$, las series X_t e Y_t se incrementan o decrementan simultámeamente con el misma razón de pendientes, o del punto de vista del coeficiente de codispersión con un paso atrás, se puede interpretar, las secuencias de tiempo se mueven iguales. El valor $\rho_{X_t,Y_t}(1) = -1$, medida en cualquier périodo observado $[t_i, t_{i+1}]$, la serie X_t e Y_t donde la serie X_t incrementan e Y_t decrementan o vice-versa, respecto a sus pendientes. Finalmente, el valor $\rho_{X_t,Y_t}(1) = 0$, expresa que aquí no hay comovimiento entre la serie X_t y Y_t.

2.5.2. Índice de disimilaridad entre series de tiempo.

El propósito de esta capítulo es apuntar aun nuevo Índice de Disimilaridad Adaptativo con aspiración a cubrir ambas medidas convencionales para la proximidad de valores y la correlación temporal para medir el comportamiento.
El resultado de la medida de disimilaridad debe también permitir ajustar el peso de la contribución entre ambos cuantiles. La función $f(x)$ modula la distancia incrementando la medida convencional. Entonces, si la correlación temporal decrese a -1. La proximidad resultante debe aproximarse a la medida convencional, si la correlación temporal es 0, entonces cuando incrementa la correlación

temporal esta desde 0 a 1.
Conforme con las especificaciones anteriores, se propone un índice de disimilaridad D basado sobre una función automática.
Ahora bien, en el contexto de anlisis de grupos (Cluster) Chouakria y Nagabhushan (2007) propusieron un Índice de Disimilaridad Adaptativo para medidas de proximidad en series temporales, como una nueva medida de distancia, la cual definieron:

$$D(X_t, Y_t) \quad = \quad f(\widehat{\rho}_{X_t,Y_t}(h))\delta_C(X_t, Y_t) \tag{2.5.7}$$

Donde $f(x)$ es una función exponencial adaptativo (tuning)

$$f(x) \quad = \quad \frac{2}{1 + exp(kx)},\ k \geq 0 \tag{2.5.8}$$

y $\delta_C(X_t, Y_t)$ es cualquier distancia convencional sobre X_t y Y_t. En la siguiente figura se puede apreciar la función de afinamiento adaptativo con distintos valores de k.

□

Figura 2.4: Función exponencial adaptativo

A continuación se muestra la siguiente tabla con la contribución en % de las dos funciones que compone este Índice de Disimilaridad.

Contribución en % de k con respecto a		
k	$f(x)$	$D(X_t, Y_t)$
0	0	100
1	46.2	53.7
2	76.2	23.8
3	90.5	9.4
≥ 5	$\approx$100	$\approx$0

Por ejemplo, cuando $k = 0$, $f(0) = 1$, el Índice de Disimilaridad está representado solamente por la distancia convencional y estamos en presencia de una medida de semejanza clásica, para los algoritmos de clasificación en Cluster.
Ahora si consideramos nuestro coeficiente de codispersión muestral y reemplazamos $\rho_{X_t,Y_t}(1)$ por $\rho_{X_t,Y_t}(h)$ tenemos un nuevo Índice de Disimilaridad el cual nos ayuadará a clasificar series de tiempo, por lo tanto esto será:

$$D(X_t, Y_t) = f(\rho_{X_t,Y_t}(h))\delta_C(X_t, Y_t)$$

ya que, $\rho_{X_t,Y_t}(1)$ es un caso particular con h=1 de $\rho_{X_t,Y_t}(h)$, con el cual se puede calcular para más retardos. Con este nuevo ińdice de disimilaridad se espera que capte mejor la correlación temporal que las medidas convencionales de distancias para clasificación de series temporales.
Este ińdice de disimilaridad es una medida alternativa para clasificar series temporales de los métodos clásicos que se conocen. El cual será de ayuda para implementar en el algoritmo de clasificación que sigue en el cápitulo II y III.

Capítulo 3

Simulación

3.1. Introducción

Para enterder mejor la funcionalidad del Índice de Disimlariad, es necesario realizar algunas simulaciones en donde se ponga aprueba como funciona teóricamente y desde ahí, explicar mejor esta nueva medida para clasificar series temporales. Como se menciono anteriormente el Índice de Disimilaridad considera la distancia entre las series ponderada por una función adaptable de afinación que balancea la proximidad con relación a los valores y la proximidad con relación al comportamiento. La contribución del comportamiento y componentes de valores para el Índice de Disimilaridad es comparada con las medidas convencionales usando dos conjuntos de datos.
La idea es sensibilizar esta medida y compararla con las distancias convencionales que fueron definidas anteriormente y ver cuales son sus diferencias.
El objetivo de este Capítulo es poner a prueba este Índice y dejar claro su importancia.
Se espera que el Índice de Disimilaridad al considerar la información del comovimiento agrupe las series, que las medidas convencionales no agrupan, es decir, este Índice agrupa series temporales con una justificación mayor que las distancias convencionales. Se espera que este Capitulo pueda dejar más claro su característica principal y su importancia.

3.2. Simulación entre series correlacionadas.

Antes de representar esta situación es necesario explicar, cuales son las condiciones de esta simulación, detallar los modelos que serán analizados con sus respectivas características.
Los elementos necesario para esta simulación es la estructura de correlación de los errores.
Por ejemplo, considere la siguiente relación:

$$\mathbb{C}ov(\epsilon_c(t), \epsilon_b(s)) = \begin{cases} \rho\tau\sigma & \text{si } t = s \\ 0 & \text{eoc} \end{cases} . \tag{3.2.1}$$

Donde $\epsilon_c(t)$ y $\epsilon_b(t)$, son errores correlacionados con media 0 y varianza τ^2, σ^2, respectivamente.
Seguidamente, esta estructura se extenderá para nuestro caso que será considerar 7 modelos AR(1), cada uno con errores $\{\epsilon_i, i = 1, \ldots, 7\}$ variables aleatorias independientes normales con media 0 y varianza σ^2.
Para verificar las bondades del Índice de Disimilaridad antes mencionadas, intencionalmente crearemos errores o variables $\epsilon_2(t)$ y $\epsilon_3(t)$ que estarán correlacionados con $\epsilon_1(t)$. Entonces, dependiendo de que valor se le de a cada correlación ρ_1 y ρ_2 es cuan correlacionadas estarán los errores. Estos errores estarán incluido en los modelos para poder diferenciar de los otros, se esperaría que el Índice de Disimilaridad agrupara primero las series con los errores que fueron intencionalmente correlacionadas.
Ahora, es claro mostrar que:

$$\begin{aligned} corr(\epsilon_1(t), \epsilon_3(s)) &= \rho_1,\ t = s \\ corr(\epsilon_1(t), \epsilon_4(s)) &= \rho_2,\ t = s \end{aligned}$$

Los valores de ρ_1 y ρ_2, ayudaran a diferencias los otros 4 modelos que sus errores serán variables independientes normal con media 0 y varianza 1.
Otra forma de considerar estos errores correlacionados, es a través de una matriz de correlación para el vector $\underline{v} = (\epsilon_1, \epsilon_2, \epsilon_3)^t$, donde cada una tiene varianza σ^2, τ^2 y η^2, respectivamente, entonces esta estructura es simulada a través, de la matriz de covarianza Σ.

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & \rho_1\tau\sigma & \rho_2\sigma\eta \\ \rho_1\tau\sigma & \tau^2 & \rho\tau\eta \\ \rho_2\sigma\eta & \rho\tau\eta & \eta^2 \end{pmatrix}$$

Y como consecuencia la matriz de correlación es:

$$\mathbb{C}\mathbf{orr}(\underline{v}) = \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho \\ \rho_2 & \rho & 1 \end{pmatrix}$$

Luego, de la estructura de correlación se incorpora otro elemento que ayudará entender mejor esta nueva medida. Se definen 7 modelos AR(1), los cuales 3 de ellos tendrán una característica en común, difiriendo por el grado de interdependencia, que se establecerá de la forma mencionada anteriormente esto es:

$$\begin{aligned} corr(\epsilon_1(t), \epsilon_2(s)) &= \rho_1,\ t = s \\ corr(\epsilon_1(t), \epsilon_3(s)) &= \rho_2,\ t = s \end{aligned}$$

Ahora, ya creados estas variables (errores) ϵ_2 y ϵ_3, se incorporan en los modelos Y_t y W_t. Finalmente, considere los siguientes modelos AR(1):

$$\begin{aligned} X_t &= \phi_1 X_{t-1} + \epsilon_1(t) \\ Y_t &= \phi_2 Y_{t-1} + \epsilon_2(t) \\ W_t &= \phi_3 W_{t-1} + \epsilon_3(t) \\ V_t &= \phi_4 V_{t-1} + \epsilon_4(t) \\ T_t &= \phi_5 T_{t-1} + \epsilon_5(t) \\ U_t &= \phi_6 U_{t-1} + \epsilon_6(t) \\ Z_t &= \phi_7 Z_{t-1} + \epsilon_7(t) \end{aligned}$$

El criterio para seleccionar los parámetros, se estableció a través de lo visto en el Capítulo 1, donde $|\phi| < 1$. Eso nos asegura de estar en presencia de series estacionarias.
Para nuestro caso los parámetros de los modelos son:$\phi_1 = -0{,}5$; $\phi_2 = 0{,}3$; $\phi_3 = -0{,}8$; $\phi_4 = 0{,}7$; $\phi_5 = 0{,}1$; $\phi_6 = -0{,}9$ y $\phi_7 = 0{,}2$.
Ademaś, se tiene que:
$\rho_1 = 0{,}9$ y $\rho_2 = 0{,}7, n = 200$.
Donde $\{\epsilon_i\}$ con $i = 4, 5, 6, 7$ son variables aleatorias independientes normal con media 0 y varianza 1.
Si gráficamos las series simuladas nos queda:

□

Figura 3.1: Series simuladas

En este gráafico se puede apreciar un conjunto de series estacionarias.
Además,si aplicamos el algoritmo de clasificación con las distancias convencionales, se pueden apreciar los siguientes resultados.

▭

Figura 3.2: Dendogramas.

Se puede apreciar que la distancia euclidiana agrupa 2 de las 3 series correlacionadas, en cambio las distancia de Frechét y DTW no agrupa ninguna de las series correlacionadas intencionalmente. Por otra parte, si se aplica el algoritmo de clasificación con el Índice de Disimilaridad Adaptativo con $h = 1$ y $k = 3$ se puede observar:

▭

Figura 3.3: Dendogramas.

En los dendogramas se pueden apreciar que el Índice de Disimilaridad con distancia euclidiana y DTW agrupa las series que están correlacionadas intencionalmente.

Si se observa el Índice con la distancia de Frechét este presenta una variación en la agrupación, reconociendo primero la serie Y_t y W_t.
No obstante, se puede apreciar que este dendograma presenta un grupo con las series que fue intencionalmente en un orden distinto, pero tiene un cluster con las series que se esperaba.
La simulación a mostrado que el Índice de Disimilaridad agrupa series temporales considerando su estructura de comovimiento y el comportamiento a su cercanía, es decir, el Índice agrupa las series que fueron intencionalmente correlacionadas fuertemente.
Se Demostró que las medidas convencionales ignoran la relacin de interdependencia entre las medidas, ya que, son primordialmente basadas en la proximidad con relacin a los valores.

Capítulo 4

Aplicación con datos reales

4.1. Introducción

En este Capítulo se hará un paréntesis al estudio del coeficiente de codispersión y al índice de disimilaridad, para introducir algunos conceptos del Sistema de Pensiones, se hablará de su función, sus génesis y la reforma de Previsión Social.
Estos conceptos nos ayudara para entender y analisar 7 AFP distintas en el periodo de Enero 1990-Febrero 2004 en su rentabilidad mensual.
Las AFP en Estudio son:

1. Cuprum.
2. Habitat.
3. Provida.
4. Planvital.
5. ING Santa María.
6. Summa Bansander.
7. Magister.

El ahorro de los trabajadores ha sido fundamental para el crecimiento y desarrollo económico de Chile, ya que ha sido mayoritariamente invertido en actividades productivas o crediticias haciendo posible el acceso a financiamiento de miles de chilenos, sea para comprar la casa propia o acceder a productos a los que nunca antes tuvo oportunidad.
El sistema de AFP ha operado en estos 26 años bajo el principio de Giro Único o exclusivo lo cual ha resguardado los fondos previsionales de problemas como las ventas atadas y los conflictos de interés, que surgen cuando un administrador de ahorros distribuye conjuntamente otros productos y además gestiona simultáneamente dineros propios y de terceros.

4.2. Sistema de AFP.

En noviembre de 1980 se publicó el D.L. 3.500 que estableció un nuevo sistema de pensiones de Vejez, Invalidez y Sobrevivencia, sobre la base del ahorro de los trabajadores y la capitalización en cuentas individuales.
El nuevo Sistema de Pensiones incorpora el concepto de propiedad de los ahorros previsionales por parte de los trabajadores afiliados, enfatizando la estrecha correspondencia entre el esfuerzo de ahorro realizado a lo largo de la vida activa de una persona y los beneficios en pensiones de vejez, invalidez y sobrevivencia que ésta recibe.
Así también, se instaura la administración de los ahorros por empresas privadas (AFP), con giro único y con el rol de otorgar beneficios y prestaciones previsionales. La normativa que regula a estas instituciones, además de la ley, es dictada por la Superintendencia de AFP, la que además fiscaliza el adecuado funcionamiento de estas sociedades.
El objetivo del Sistema de Pensiones es proveer ingresos de reemplazo para los trabajadores que dejan la vida activa o laboral y cubrir los riesgos de invalidez (total o parcial) y de muerte del trabajador (sobrevivencia), de manera de proteger al afiliado y a su grupo familiar.
Se basa en el ahorro y la Capitalización Individual. Los trabajadores dependientes cotizan obligatoriamente en las AFP y los independientes lo hacen en forma voluntaria. Los trabajadores son dueos de su ahorro previsional y en ellos recae la responsabilidad de preocuparse de su pensión, sin perjuicio que el Estado garantice pensiones mínimas.
Otorga libertad de elección a los afiliados. De este modo el trabajador puede elegir la administradora que gestione sus ahorros previsionales y cambiarse cuando lo desee, así como la edad a la que quiere pensionarse (jubilación por vejez o anticipada) y la modalidad de pago de pensión (retiro programado, renta vitalicia o retiro programado con renta vitalicia diferida). Asimismo, puede elegir el Tipo de Fondo en donde invertir sus ahorros.
Es uniforme en la aplicación de las normas para todos los afiliados y establece directa relación entre las contribuciones de los trabajadores y los beneficios obtenidos.
La administración es privada y está a cargo de sociedades anónimas especializadas

4.3. Rentabilidad de los Fondos de Pensiones.

La rentabilidad de los fondos es un factor determinante en un sistema de capitalización y ahorro individual, donde el saldo acumulado durante la vida activa del trabajador incidirá significativamente en el monto de las pensiones. De esta forma, un punto porcentual de mayor ganancia anual puede elevar el monto de la pensión al cabo de 30 ó 40 años de ahorro entre un 25 % y 30 %. Tras 26 años del sistema de pensiones la rentabilidad promedio real anual alcanza a UF +10 %, un resultado muy superior al 4 a 5 % de la rentabilidad estimada como suficiente para financiar pensiones de reemplazo con una tasa del orden del 70 % del promedio de las remuneraciones.

4.4. Contribución al Desarrollo Económico.

Los fondos de pensiones han hecho un aporte significativo al desarrollo económico de Chile, otorgando, asimismo, acceso a amplios sectores de la población a servicios financieros a los que nunca habían podido acceder. Para un país en desarrollo como Chile, hoy economía emergente, disponer de ahorro fue siempre un problema y una aspiración, ya que es sabido que el círculo virtuoso del progreso comienza con el ahorro y sigue con la inversión y el crecimiento. En la actualidad un 25 % del total de activos financieros nacionales pertenece a los fondos de pensiones de los trabajadores, lo cual constituye un aporte tangible al desarrollo económico.

El mercado de capitales chileno se ha expandido en forma significativa los últimos 26 años, progreso en el cual el ahorro de los fondos de pensiones ha jugado un rol clave. Se han convertido en una fuente de financiamiento para el Estado, para los bancos y para las empresas. El sistema de pensiones ha introducido competencia en el mercado de capitales.

Casi el 100 % del financiamiento de letras hipotecarias para la vivienda y bonos de bancos pertenece a los fondos de pensiones; también un 45 % de los bonos de empresas y un 55 % de los títulos estatales.

En materia de inversión en infraestructura los fondos de pensiones han contribuido, mediante la inversión en bonos, al financiamiento de los principales proyectos de infraestructura vial y de transporte, dentro de los que se puede mencionar la expansión del Metro, construcción de aeropuertos y autopistas.

4.5. Antiguo Sistema Previsional.

Chile fue el primer país de América Latina que creó un Sistema de Seguridad Social, a comienzos del siglo XX. A lo largo de los años se fueron creando diversos regímenes de pensiones diferenciados por el tipo de actividad o grupos ocupacionales con reglas y beneficios distintos.

Es así como llegar a coexistir 52 "Cajas" o Instituciones de Previsión que operaban bajo el esquema de reparto. Esto significa que los aportes de los afiliados activos financiaban las pensiones de los pasivos y, por tanto, la subsistencia del Sistema estaba supeditada a la relación trabajador/pensionado existente en la poblacin en cada momento del tiempo.

Durante los primeros aos de existencia del sistema, la proporción de trabajadores fue suficiente para financiar los beneficios de los pensionados. Sin embargo, los cambios demográficos, que fueron reflejando una permanente disminución de la natalidad y un aumento en las expectativas de vida revirtieron esta relación, provocando un fuerte desfinanciamiento del Sistema.

Mientras que en el año 1955, por cada 12,2 trabajadores cotizantes haba 1 pensionado, en 1980 por cada 2,5 trabajadores cotizantes había 1 pensionado. Es decir, sólo en 25 aos el costo de los trabajadores cotizantes, se incrementó casi 5 veces.

Un agravante del problema del financiamiento lo constituyó la fuerte evasin previsional, ya que a

trabajadores y empleadores les resultaba más económico hacer imposiciones por el mínimo legal, preocupndose sólo de imponer por valores reales los últimos años de la vida activa del trabajador, cuando las imposiciones eran consideradas para la jubilación. Esta situación obligaba al Estado a elevar las imposiciones lo que, a su vez, incentivaba una mayor evasión previsional y así sucesivamente. A ellos se agrega que el Estado fue proclive a otorgar beneficios sin el adecuado financiamiento, lo que acentuó el problema reseñado, ocasionando un déficit fiscal creciente, equivalente a un 28 % del gasto en la década 1970 - 1980.

El Antiguo Sistema se caracterizó también por su falta de equidad. Dado que no existía una relación directa entre los aportes de los trabajadores y los beneficios percibidos, se apreciaban notables desigualdades entre los múltiples grupos cubiertos. Esta situación se sustentaba en la facultad del poder político para definir quién se beneficiaba y cuánto, quedando de manifiesto el otorgamiento de mayores concesiones a los grupos que ejercían mayor presión. En efecto, en el año 1965 (1), los obreros chilenos, que representaban el 70 % de los cotizantes y cuyos ingresos eran los más bajos, percibían en términos absolutos, pensiones equivalentes a la mitad de lo que obtenían los empleados privados y a 1/14 de lo obtenido por los empleados públicos, los cuales eran grupos de mayores ingresos. Al mismo tiempo, el aporte efectuado por estos trabajadores en el mismo período (2), era equivalente a más del doble del que realizaban los empleados públicos y sólo un 10 % inferior al de los privados.

Desde hacía mucho tiempo en nuestro país se presentaba la necesidad de introducir cambios al sistema de Seguridad Social. Ya en la década del '60 se elaboraron diversos informes sobre las falencias del antiguo sistema de Seguridad Social Chileno en los que se proponían cambios profundos.

El desfinanciamiento y la inequidad del esquema de reparto dieron origen a una reforma Previsional que creó, mediante el D.L. 3.500 de 1980, un Nuevo Sistema de Pensiones basado en la Capitalización Individual y administrado por entidades privadas denominadas Administradoras de Fondos de Pensiones (AFP). El Antiguo Sistema continuó funcionando, principalmente a través de un ente único, denominado Instituto de Normalización Previsional (INP), el cual fusionó a las principales Cajas de Previsión.

El Estado se hizo responsable del financiamiento de las cotizaciones pagadas en el Antiguo Sistema por aquellas personas que se cambiaron al Nuevo Sistema. Ello se materializa a través de unos instrumentos financieros denominados Bonos de Reconocimiento, los cuales son representativos de dichos períodos de cotizaciones y que el trabajador hace efectivo al instante de pensionarse o fallecer.

El Bono de Reconocimiento se reajusta de acuerdo a la variación de la inflación y devenga un interés del 4 % real anual, el cual se capitaliza cada año.

4.6. Efecto Manada.

Uno de los vicios que se le han imputado al sistema de AFP es que si todas no invirtieran en los mismos activos, quizás inyectarían competencia al mercado por la va de ofrecer mejores retornos a los afiliados.
Es por eso que las distintas AFP invierten en acciones en las mismas empresa, revelando que la rentabilidad en las distintas AFP son similares. Este fenómeno se llama Efecto Manada.

4.6.1. Contexto y consecuencias de la competencia en rentabilidad vía ranking

Para entender el contexto de esta competencia, es útil tener en mente dos hechos. Primero, las magnitudes de las diferencias en rentabilidad tienden a ser pequeas. De hecho, 10 puntos base de mayor o menor rentabilidad pueden hacer la diferencia entre el primero y el último en el ranking en un ao determinado.
Segundo, hay extrema transparencia en las decisiones de inversión. Las carteras de todos los fondos de pensiones son conocidas 10 días después de cerrado el mes. Si se imagina que una buena decisión de inversión es una especie de invento, entonces, en este caso, no habría un período de protección o patente de uso; todos podrían conocerlo y utilizarlo en forma gratuita pocos días después de haber sido creado.
Lo anterior, en forma absolutamente natural, genera incentivos a situarse muy cerca del benchmark (o vara de medida), que en este caso es la cartera de la competencia relevante, y a tomar decisiones en el margen, relativamente pequeas en magnitud. Nótese que la cantidad de trabajo o estudio necesarios para tomar una buena decisión de inversin es la misma, independientemente de la magnitud de la decisión de inversión tomada. El riesgo pertinente pasa a ser el comercial, que consiste en quedar en una mala posición en el ranking de rentabilidad. Una medida que refleja adecuadamente que el riesgo se mide en términos relativos es el llamado tracking error o variabilidad de la diferencia de retorno con el sistema. El fenmeno descrito genera el llamado efecto manada o herding (lo que, en todo caso, no es exclusividad chilensis ya que se da en casi toda la industria de administracin de carteras).
Algo que sí es particular del caso chileno es que existe por ley una rentabilidad mnima que se calcula con respecto al promedio del sistema de AFP. Esta norma tiende a reafirmar el comportamiento descrito en el párrafo anterior, pero es probable que incluso sin rentabilidad mínima la forma de competir fuera similar.

4.7. Acciones favoritas de las AFP.

Según Econsul,en un Articulo publicado el 8 de Abril de 2008, revisado el Miércoles 23 de Julio 2008,16:45 hrs, mostró que Provida, apuesta más que el resto en CMPC, donde mantiene una sobreesposición de 1.81 % respecto de la Industria. Su segunda apuesta más fuerte comparativamente hablando es Endesa y AntarChile. La segunda AFP por fondo administrativo, Habitat, tiene como favorita a Telefónica CTC, seguida por Endesa y AntarChile.
Planvital e ING Santa María,por ejemplo, están más jugadas por Soquimich que el sistema como un todo, aunque luego sus preferencias difieran. Cuprum, por su parte, favorece a Colbún, las acciones de Telefonica CTC y a Masisa; y al cierre de Junio Summa Bansander mostraba una mayor incliniación por Endesa, D&S y Copec.

4.8. Análisis Descriptivo

En esta sección se hará un sondeo, para entender como se han comportado las 7 AFP, desde un punto de vista descriptivo.
En esta primera parte se realizara un conteo para saber cuantas veces han estado primera, es decir, con la mayor rentabilidad, en el caso que una AFP tiene la mayor rentabilidad se le asignara un valor 1 y al resto de las AFP 0, entonces, finalmente se sumara todos los casos para saber cuantas veces estuvo primera en rentabilidad, si hay mas de un empate, es decir, que mas de dos AFP tienen la mayor rentabilidad, se le asignaran a cada una el valor 1.
La siguiente tabla muestra los resultados obtenidos, para las AFP en estudio.

AFP	CONTEO
Cuprum	35
Habitat	38
Magister	44
Planvital	26
Provida	37
Sta Maria	21
Summa	38

Se puede apreciar que la AFP que estuvo con mayor rentabilidad y esta primera con 44 veces dentro de este conteo es la AFP Magister, seguido de la AFP Provida que estuvo 37 veces primera y la AFP Cuprum con 35 veces.

□

Figura 4.1: Dispersión de la rentabilidad de las AFP.

Se puede observar que en gráfico de Cajón con bigotes, se puede apreciar que hay datos que escapan a estos Cajones. Además, media son similares en todas las AFP. Acontinuación, se mostrara un sumario con algunas medidas descriptivas.

Figura 4.2: 7 Series durante 1990 - 2004.

El siguiente gráfico se muestra el comportamiento de la rentabilidad en el tiempo de las 7 AFP, si se observa el gráfico se puede apreciar una linea gruesa, este fenómeno se denomina efecto manda el cual se ha mencionado anteriormente y en consecuencia no se puede notar cual de estas AFP esta por encima de otra.

En la siguiente tabla se muestra un análisis descriptivo por AFP.

Estadístico	Cuprum	Habitat	Magister	Planvital
Mínimo	-8	-7	-7,4	-6,8
1st Qu	-0,2	-0,1	-0,17	-0,1
Mediana	0,8	0,8	0,85	0,9
Media	0,82	0,83	0,83	0,85
3rd Qu	1,7	1,7	1,7	1,7
Máximo	6,6	6,9	6,8	7,2
Estadístico	Provida	Sta Maria	Suma	
Mínimo	-6,3	-6,9	-7,4	
1st Qu	-0,17	-0,17	-0,17	
Mediana	0,8	0,85	0,9	
Media	0,77	0,79	0,84	
3rd Qu	1,7	1,7	1,7	

Si consideramos, las AFP como variables fijas se puede observar que la rentabilidad media de Cuprum es de 0.8265, las rentabilidades extremas se pueden apreciar con un mínimo de -6.3 y un máximo de 6.7. Además, se puede apreciar que en 25 % la rentabilidad llega a -0.175 y en el 75 % llega a 1.7. Es claro, mencionar que para el restos de las AFP la interpretación es la misma.

Más adelante se estudiaran estas AFP, para representar, a través, de modelos paramétricos y ver cual modelo puede representar mejor esta situación.

4.8.1. Análisis Descriptivo AFP CUPRUM

En la siguiente tabla se muestra la rentabilidad de la AFP CUPRUM con las medias por año y mes.

	Mes	E	F	M	A	M	J	J	A	S	O	N	D	
Año	n	1	2	3	4	5	6	7	8	9	10	11	12	Media año
1990	1	1	-1,1	3,6	-0,4	0,8	2,6	0,9	2,4	1,3	-0,9	2	4,7	1,40
1991	2	2,2	4,7	3,1	2,2	-0,3	1,6	5,1	3	3,9	5,2	-4	0,6	2,27
1992	3	0,1	0,3	5,2	1,9	-0,2	0,9	0,8	-1,1	-1,8	-1,9	0,6	-1,1	0,30
1993	4	1,2	2,5	-0,4	-0,4	-1,7	1,5	3,5	1,6	0,9	1,8	1,2	3,5	1,26
1994	5	6,6	4,5	-1,4	-2,5	2,3	2,8	-0,8	1	1,5	3,5	1,7	-0,9	1,52
1995	6	-2,2	-1,4	-1,7	1,7	2,8	1,5	0,8	-2,3	-1,9	0,1	0,2	0,9	-0,12
1996	7	1,8	-0,4	-0,1	-0,5	1,6	0,1	2	-0,2	-0,2	1,2	-0,5	-1,5	0,27
1997	8	0,9	1,8	0,8	0,6	1,1	2,7	0,7	-0,2	-0,4	-1,6	-2,1	0,1	0,36
1998	9	-2	-1,1	3,4	1	-1,6	-1,8	-0,6	-1,1	-8	3,6	4,8	1,3	-0,17
1999	10	1,0	3,0	1,9	2,3	2,7	1,9	1,7	-0,4	1,1	-0,2	0,8	1,9	1,48
2000	11	1,7	1,3	0,0	0,4	0,3	1,3	1,3	0,6	0,8	-0,4	1,1	0,6	0,74
2001	12	1,4	1,4	0,6	0,0	1,9	1,3	0,7	-0,2	0,2	0,6	1,3	0,2	0,79
2002	13	0,7	0,4	1,2	0,7	0,3	-0,1	-0,9	0,5	1,2	-0,6	0,8	0,6	0,38
2003	14	1,7	0,4	0,7	1,1	2,7	1,9	0,3	1,5	0,3	1,7	-0,8	0,1	0,97
2004	15	1,1	1,4											1,20
	Media mes	1,14	1,17	1,20	0,58	0,90	1,30	1,10	0,36	-0,07	0,86	0,50	0,79	

4.8.2. Análisis Descriptivo AFP HABITAT

Habitat

	Mes	E	F	M	A	M	J	J	A	S	O	N	D	
Año	n	1	2	3	4	5	6	7	8	9	10	11	12	Media año
1990	1	0,4	-1,8	3,3	-0,4	0,8	2,4	0,7	2,5	1,4	-0,8	1,8	4,8	1,3
1991	2	2,2	4,7	3,3	2,7	-0,3	2	5	2,9	3,3	4,2	-3,4	0,5	2,3
1992	3	0,3	0,2	4,6	1,4	0,1	0,9	0,8	-1	-2,1	-1,9	0,6	-1,1	0,2
1993	4	1,2	2,6	-0,5	-0,6	-1,8	1,7	3,7	1,6	0,9	1,9	0,9	3,6	1,3
1994	5	6,9	4,4	-1,8	-2,7	2,3	2,7	-0,8	0,8	1,4	3,6	1,5	-1,2	1,4
1995	6	-2,2	-1,6	-2,1	1,7	2,9	1,3	0,8	-2,7	-2	-0,1	0,2	1	-0,2
1996	7	1,9	-0,5	-0,1	-0,6	1,9	0	2,3	-0,2	-0,2	1,1	-0,6	-1,4	0,3
1997	8	1,2	2	0,9	0,5	1,2	3	0,8	-0,1	-0,4	-1,5	-2,1	0,3	0,5
1998	9	-1,7	-0,8	3,3	1	-1,4	-1,6	-0,3	-0,5	-7	3	4,5	1,5	0,0
1999	10	1,1	2,9	1,7	2,3	2,5	1,8	1,9	-0,4	1,2	-0,2	1,2	1,9	1,5
2000	11	1,7	1,5	0,1	0,1	0,0	1,4	1,4	0,6	0,8	-0,5	1,0	0,7	0,7
2001	12	1,5	1,3	0,4	-0,1	2,0	1,5	0,9	0,2	0,1	0,8	1,2	0,3	0,8
2002	13	0,7	0,5	1,1	0,7	0,4	0,1	-0,1	1,1	1,8	-0,4	0,6	0,2	0,6
2003	14	1,4	0,7	0,9	0,8	2,1	1,7	0,3	1,4	0,3	1,6	-0,7	0,0	0,9
2004	15	0,9	1,2											1,0
	Media por mes	1,2	1,2	1,1	0,5	0,9	1,3	1,2	0,4	0,0	0,8	0,5	0,8	

4.8.3. Análisis Descriptivo AFP MAGISTER

magister

	Mes	E	F	M	A	M	J	J	A	S	O	N	D	
Año	n	1	2	3	4	5	6	7	8	9	10	11	12	Media año
1990	1	0,6	-1,6	3,5	-0,5	0,6	2,3	0,8	2,4	1,3	-0,9	1,9	4,5	1,2
1991	2	2	5,3	3,4	3	-0,4	2,1	5,9	3,1	4,4	6,2	-5,1	0,5	2,5
1992	3	0,2	0,5	5,5	1,5	0	1	0,9	-1,3	-2	-2,3	0,6	-1,4	0,3
1993	4	1,1	2,8	-0,4	-0,7	-1,9	1,5	3,8	1,7	1	2	1,2	3,5	1,3
1994	5	6,8	4,4	-1,7	-2,5	2,5	2,8	-0,9	0,9	1,1	3,5	1,5	-1,2	1,4
1995	6	-2,6	-1,7	-1,7	2,1	2,9	1,3	0,8	-2,9	-2,3	0	0,5	0,9	-0,2
1996	7	1,7	-0,4	0	-0,7	1,6	-0,1	2,1	-0,1	-0,1	1,3	-0,6	-1,5	0,3
1997	8	1,2	1,8	0,8	0,6	1,1	2,8	0,7	-0,3	-0,5	-1,7	-2,4	0,2	0,4
1998	9	-2,1	-1,1	3,6	0,9	-1,4	-1,7	-0,5	-0,6	-7,4	2,7	5,4	1,4	-0,1
1999	10	1,1	3,3	2,1	2,5	2,7	1,9	2,0	-0,5	1,2	-0,2	1,0	1,8	1,6
2000	11	1,6	1,4	-0,2	0,0	0,2	1,5	1,3	0,7	1,0	-0,2	1,0	0,8	0,8
2001	12	1,3	1,3	0,7	-0,2	1,6	1,5	1,0	0,4	0,3	1,0	1,1	0,0	0,8
2002	13	0,5	0,3	0,8	0,4	0,2	0,1	0,0	1,1	2,5	-0,8	0,2	0,6	0,5
2003	14	1,4	0,4	0,8	1,0	2,4	1,0	0,7	2,4	0,2	1,6	-0,7	0,0	0,9
2004	15	0,5	1,3											0,9
	Media por mes	1,0	1,2	1,2	0,5	0,9	1,3	1,3	0,5	0,0	0,9	0,4	0,7	

4.8.4. Análisis Descriptivo AFP PLANVITAL

Planvital

	Mes	E	F	M	A	M	J	J	A	S	O	N	D	
Año	n	1	2	3	4	5	6	7	8	9	10	11	12	Media año
1990	1	1,6	-1	3,8	-0,4	0,9	2,5	0,9	2,2	1,3	-1	2,1	4,4	1,4
1991	2	2	5,2	3,5	2,9	-0,7	1,6	5,3	3	4,2	5,5	-4,6	0,6	2,4
1992	3	0,1	0,4	5,7	2,6	-0,3	0,9	1	-1,3	-1,9	-2,2	0,6	-1,6	0,3
1993	4	1,1	2,7	-0,5	-0,9	-1,8	1,6	3,9	1,6	0,9	2,1	1,6	3,7	1,3
1994	5	7,2	4,5	-1,4	-2,5	2,7	2,8	-0,9	1	1,3	3,7	1,5	-1,3	1,6
1995	6	-2,5	-1,7	-1,8	1,8	2,9	1,4	0,8	-2,8	-2,3	0	0,5	1	-0,2
1996	7	1,7	-0,5	0	-0,7	1,8	0	2,3	-0,2	-0,1	1,1	-0,7	-1,6	0,3
1997	8	1,1	2	0,8	0,6	1,1	2,9	0,7	-0,1	-0,5	-1,7	-2,2	0,2	0,4
1998	9	-2	-1	3,4	0,9	-1,5	-1,6	-0,4	-0,6	-6,8	3	4,7	1,3	-0,1
1999	10	1,1	3,2	2,1	2,4	2,5	1,8	1,9	-0,5	1,1	-0,2	1,0	1,8	1,5
2000	11	1,5	1,4	-0,1	0,1	0,2	1,5	1,4	0,7	0,9	-0,3	1,1	0,8	0,8
2001	12	1,4	1,3	0,6	-0,1	1,7	1,4	0,9	0,4	0,2	0,9	1,2	0,2	0,8
2002	13	0,6	0,4	0,9	0,5	0,2	0,0	-0,1	1,2	2,2	-0,7	0,4	0,5	0,5
2003	14	1,4	0,4	0,9	1,0	2,4	1,4	0,5	1,9	0,3	1,6	-0,8	0,1	0,9
2004	15	0,8	1,3											1,0
	Media por mes	1,1	1,2	1,3	0,6	0,9	1,3	1,3	0,5	0,1	0,8	0,5	0,7	

4.8.5. Análisis Descriptivo AFP PROVIDA

Provida

	Mes	E	F	M	A	M	J	J	A	S	O	N	D	
Año	n	1	2	3	4	5	6	7	8	9	10	11	12	Media año
1990	1	-0,1	-1,9	3	-0,4	0,7	2,1	0,5	2,3	1,3	-0,8	1,6	4,4	1,1
1991	2	2,1	4	3,3	2,2	-0,3	1,5	4	2,5	2,9	3,4	-2,8	0,7	2,0
1992	3	0,3	0,1	4	1,8	0,1	0,8	0,9	-0,8	-2	-1,8	0,7	-1	0,3
1993	4	1,5	2,9	-0,6	-0,7	-1,9	1,4	3,5	1,6	1	1,8	1	3,5	1,3
1994	5	6,7	4,7	-1,6	-2,8	2,3	2,6	-0,8	0,8	1,5	3,4	1,6	-1,4	1,4
1995	6	-2,1	-1,6	-2,2	1,4	3,1	1,7	1	-2,4	-2	-0,2	0,2	0,8	-0,2
1996	7	1,9	-0,4	-0,1	-0,7	1,7	0	2,3	-0,2	-0,1	1,1	-0,6	-1,6	0,3
1997	8	1	1,9	0,8	0,6	1,1	2,9	0,7	-0,2	-0,5	-1,6	-2,3	0,2	0,4
1998	9	-1,7	-0,9	3,3	1,1	-1,4	-1,6	-0,3	-0,4	-6,3	2,9	4,4	1,4	0,0
1999	10	1,0	2,9	1,8	2,3	2,4	1,9	2,1	-0,3	1,2	-0,1	1,0	1,6	1,5
2000	11	1,3	1,2	0,0	0,4	0,3	1,4	1,3	0,7	0,8	-0,4	1,2	0,7	0,8
2001	12	1,4	1,3	0,4	-0,1	1,9	1,3	0,8	0,1	0,0	0,5	1,3	0,3	0,8
2002	13	0,6	0,6	1,1	0,5	0,3	0,0	-0,3	1,0	1,7	-0,5	0,5	0,4	0,5
2003	14	1,2	0,4	0,8	0,8	2,4	1,7	0,4	1,5	0,3	1,8	-0,8	0,4	0,9
2004	15	1,0	1,1											1,0
	Media por mes	1,1	1,1	1,0	0,5	0,9	1,3	1,2	0,4	0,0	0,7	0,5	0,7	

4.8.6. Análisis Descriptivo AFP STA MARIA

sta maria

	Mes	E	F	M	A	M	J	J	A	S	O	N	D	
Año	n	1	2	3	4	5	6	7	8	9	10	11	12	Media año
1990	1	0,4	-2	3,1	-0,5	0,7	2,2	0,5	2,4	1,3	-0,8	1,9	4,7	1,2
1991	2	2,1	4,6	3,2	2,6	-0,3	1,9	5	2,8	3,7	4,8	-4,1	0,6	2,2
1992	3	0,3	0,2	4,9	1,6	0	0,9	0,9	-1,1	-2,1	-2,2	0,8	-1,3	0,2
1993	4	1,2	2,7	-0,5	-0,6	-1,7	1,5	3,6	1,6	1	1,9	1,1	3,4	1,3
1994	5	6,9	4,3	-1,6	-2,6	2,5	2,8	-0,9	0,9	1,4	3,7	1,4	-1,6	1,4
1995	6	-2,4	-1,7	-2	1,6	2,9	1,4	0,9	-2,8	-2,1	-0,1	0,2	0,9	-0,3
1996	7	1,8	-0,5	0	-0,6	1,7	0,1	2,2	-0,2	-0,1	1,2	-0,6	-1,5	0,3
1997	8	1,1	1,9	0,8	0,5	1,1	2,8	0,8	-0,2	-0,5	-1,6	-2,4	0,2	0,4
1998	9	-1,9	-1	3,5	1,1	-1,4	-1,6	-0,3	-0,6	-6,9	2,9	4,6	1,4	0,0
1999	10	1,1	2,9	1,9	2,2	2,4	1,8	2,0	-0,4	1,1	-0,2	1,5	1,4	1,5
2000	11	1,5	1,4	0,0	0,2	0,2	1,5	1,3	0,6	0,8	-0,4	1,2	0,7	0,8
2001	12	1,4	1,2	0,4	-0,1	2,0	1,4	0,9	0,3	0,0	0,6	1,3	0,4	0,8
2002	13	0,6	0,4	1,1	0,6	0,3	0,0	-0,3	1,0	1,7	-0,5	0,6	0,4	0,5
2003	14	1,4	0,4	0,7	0,9	2,4	1,8	0,4	1,5	0,3	1,9	-0,9	0,1	0,9
2004	15	0,8	1,1											1,0
	Media por mes	1,1	1,1	1,1	0,5	0,9	1,3	1,2	0,4	0,0	0,8	0,5	0,7	

4.8.7. Análisis Descriptivo AFP SUMMA

summa

	Mes	E	F	M	A	M	J	J	A	S	O	N	D	
Año	n	1	2	3	4	5	6	7	8	9	10	11	12	Media año
1990	1	0,7	-1,7	3,7	-0,3	1	2,5	0,7	2,5	1,5	-0,8	2	5,1	1,4
1991	2	2,4	5,1	3,2	3	-0,3	2	5,3	2,7	3,8	5,5	-4,1	0,6	2,4
1992	3	0,3	0,5	4,8	1,5	-0,4	0,9	0,8	-1	-1,9	-1,7	0,3	-1	0,3
1993	4	1,1	2,4	-0,4	-0,4	-1,8	1,6	3,8	1,9	1	2,1	1,3	3,2	1,3
1994	5	6,5	4	-1,3	-2,8	2,4	2,6	-0,7	0,9	1,1	3,5	1,5	-1,5	1,4
1995	6	-2,3	-1,9	-1,8	1,6	3,1	1,3	1	-2,7	-2,2	0,3	0,7	1,1	-0,2
1996	7	1,9	-0,6	0,1	-0,7	1,7	0	2,2	-0,2	-0,1	1,1	-0,7	-1,7	0,3
1997	8	1	2	0,8	0,6	1,1	2,9	0,8	-0,1	-0,4	-1,7	-0,23	0,1	0,6
1998	9	-2	-1,1	3,5	0,9	-1,5	-1,7	-0,5	-0,8	-7,4	3	5,1	1,4	-0,1
1999	10	1,1	3,1	1,9	2,2	2,6	1,8	1,9	-0,4	1,1	-0,3	1,1	1,8	1,5
2000	11	1,6	1,4	0,1	0,1	0,1	1,4	1,3	0,6	0,8	-0,4	1,0	0,6	0,7
2001	12	1,5	1,3	0,4	-0,1	1,9	1,4	0,9	0,3	0,0	0,7	1,3	0,3	0,8
2002	13	0,7	0,5	1,1	0,7	0,3	0,0	-0,3	1,0	1,7	-0,6	0,6	0,3	0,5
2003	14	1,3	0,5	0,9	0,9	2,5	1,8	0,4	1,5	0,4	1,9	-0,7	0,2	1,0
2004	15	1,1	1,1											1,1
	Media por mes	1,1	1,1	1,2	0,5	0,9	1,3	1,3	0,4	0,0	0,9	0,7	0,8	

4.9. Aplicación Índice de Comovimiento o Codispersión

Si se aplica el Coeficiente de Codispersión al sistema de AFP, se puede apreciar los siguientes resultados. Índice de Comovimiento con $h = 1$.

$\rho(h = 1)$	Cuprum	Habitat	Provida	Planvital	Sta Maria	Summa
Habitat	0,990					
Provida	0,984	0,982				
Planvital	0,983	0,987	0,993			
Sta Maria	0,982	0,993	0,970	0,979		
Summa	0,991	0,995	0,988	0,992	0,989	
Magister	0,986	0,985	0,983	0,986	0,975	0,987

Se puede observar que el comovimiento entre las AFP es similar, es decir, estas comueven iguales en cualquier instante de tiempo.
Ahora con $h = 2$

$\rho(h = 2)$	Cuprum	Habitat	Provida	Planvital	Sta Maria	Summa
Habitat	0,990					
Provida	0,987	0,985				
Planvital	0,989	0,987	0,993			
Sta Maria	0,983	0,994	0,970	0,979		
Summa	0,992	0,995	0,988	0,992	0,989	
Magister	0,990	0,985	0,983	0,986	0,975	0,987

Ahora con h=3

$\rho(h=3)$	Cuprum	Habitat	Provida	Planvital	Sta Maria	Summa
Habitat	0,988					
Provida	0,986	0,987				
Planvital	0,988	0,987	0,993			
Sta Maria	0,981	0,994	0,970	0,979		
Summa	0,990	0,995	0,988	0,992	0,989	
Magister	0,983	0,985	0,983	0,986	0,975	0,987

Ahora con h=4

$\rho(h=4)$	Cuprum	Habitat	Provida	Planvital	Sta Maria	Summa
Habitat	0,990					
Provida	0,986	0,987				
Planvital	0,989	0,989	0,993			
Sta Maria	0,982	0,993	0,970	0,979		
Summa	0,992	0,995	0,988	0,992	0,989	
Magister	0,990	0,985	0,983	0,986	0,975	0,987

4.10. Modelamiento

Como se menciono anteriormente, se estudiara la estacionalidad y el comportamiento de la rentabilidad, para poder modelar esta situación se usara las definiciones de series de tiempo mencionadas anteriormente y se estudiara como se comportan respecto al índice o coeficiente de Codispersión para ver cual es su grado de comovimiento.

Por otra parte, se aplicará el índice de disimilaridad, para clasificar las AFP respecto a su rentabilidad y considerando la estructura de correlación temporal, esto ayudara a responder los objetivos de este Proyecto de Titulo.

Dentro de las hipótesis que contempla este Proyecto de Titulo, es que las series que se estudiaran son estacionarias.

Para analizar la estacionalidad, hay varios aspectos que se tienen que considerar una es por simple inspección, proponer transformaciones, que ayudaran a estabilizar la varianza unas propuestas pueden ser $\log x$, $\sqrt{x}$, $\frac{1}{x}$ y desplazar la serie en c por nombrar algunas. Otra alternativa es, a través, de la estructura de correlación temporal que en este caso es la función de autocorrelación (ACF) y función de autocorrelación parcial (Partial ACF), estas estructuras ayudaran a proponer que modelo se puede ajustar mejor a esta situación.

Para saber cual es el modelo que mejor se ajusta, existen valores que nos ayuda a comparar cuales de los modelos se ajusta mejor, como es el criterio Akaike (AIC), donde el AIC más pequeño evidencia un mejor ajuste. Si observamos las series por separado, como muestra la siguiente figura.

Figura 4.3: Rentabilidad AFP.

En estos Gráficos, se puede apreciar el comportamiento similar de las 7 AFP. Donde se puede observar que en todas las AFP, en el período de Noviembre 1997 y en Enero, Junio y Septiembre de 1998, las AFP muestran sus caídas mas significativas. Con Septiembre de 1998 la caída mas importante de estas, este dato atípico podría afectar la modelación en sus paramétros o quizás en validar el supuesto de normalidad.
Como las AFP muestran rentabilidades con valores negativos y además, se puede apreciar que la varianza de las AFP no es constante, proponer una tansformación que nos ayude a homogenizar la varianza y dejar las series estacionaria, se tiene que considerar estos valores. Por lo general,para homogenizar la varianza se ocupan la $\sqrt{(x)}$ y $\log(x)$, pero como existen valores negativos, habrían casos indeterminados,es por eso que se trasladara esta serie en 10 unidades, un número un poco mayor al de los valores extremo que es -7.4, para así tener valores positivos y homogenizar la serie con el $\log(x)$. Acontinuación, se muestran las series transformadas:

Figura 4.4: Transformación AFP.

En estos Gráficos se puede apreciar, que si bien la serie esta más estacionaria, aun se puede apreciar que la caída de septiembre 1998 afecta, querer homogenizar la varianza.
Acontinaución se analizaran las funciones ACF y Partial ACF.

Figura 4.5: Función de Autocorrelación AFP.

sfsffggsaggsgsagsg

Figura 4.6: Función de Autocorrelación Parcial AFP.

En base a los gráficos anteriores se pueden proponer algunos modelos a priori, para modelar la rentabilidad de las AFP.
Se puede apreciar que la parte autoregresiva contribuye en 2 lag o retardos, Por otra parte, la estructura de media movíl contribuye con 1 lag o retardo. Acontinuación algunas propuestas de modelos para esta situación.

Modelos Propuestos	AIC
ARMA(1,1)	-60.71
ARMA(2,1)	-58.73
AR(1)	-60.98
MA(1)	-61.84

Si se observa el AIC, se puede apreciar que de los modelos propuestos, el modelo que minimiza la varianza de los residuos es el modelo de media movíl MA(1).
Seguidamente, se tratara se verificar los supuestos de normalidad.

4.11. Análisis de Cluster

En esta sección se implementará el algoritmo de clasificación para series temporales con el Índice de Disimilaridad Distancias convencionales.

Figura 4.7: Dendogramas.

Figura 4.8: Alineamiento entre AFP.

Índide de Disimilaridad Adaptativo con h=1,k=1,2,3

Figura 4.9: Dendogramas.

Figura 4.10: Dendogramas.

4.12. Interpretación

Por definir

Figura 4.11: Dendogramas.

Figura 4.12: Dendogramas.