



Centro de Investigación en Matemáticas, A.C.

CIMAT

"Formación y Análisis de Conglomerados
de Series de Tiempo.
Caso Índice Nacional de Precios al
Consumidor (Enero 1989-Febrero 2008)"

T E S I S

Que para obtener el grado de
Maestra en Ciencias

con Orientación en
Estadística Oficial

P r e s e n t a

Jackeline Martínez Ponce

Co-Directora de Tesis:

**Dra. Graciela María de los
Dolores González Farías**

Co-Director de Tesis:

**M. en C. José Ramón Domínguez
Molina**

Guanajuato, Gto. Octubre de 2009

A mi esposo

Pedro Ismael

A mi hijo

Pedro Iván

Agradecimientos

Deseo expresar mi más sincero agradecimiento a todas las personas que hicieron posible, no solo la elaboración de este trabajo de tesis sino mis estudios de maestría.

En primer lugar mi agradecimiento a la **Dra. Graciela González Farías**, es imposible en unas cuantas líneas expresar mi gratitud y admiración por una persona tan preparada, me siento doblemente privilegiada por haber sido su alumna y tesista.

Al **M. en C. José Ramón Domínguez Molina**, por su dirección, por su infinita paciencia y por abrirme las puertas de su casa y de su hermosa familia.

A los **Dres. Jorge Valero Gil de la Universidad Autónoma de Nuevo León (UANL)** y **José Elías Rodríguez Muñoz de la Universidad de Guanajuato (UG)** cuyos comentarios enriquecieron este trabajo.

A todos los profesores del **Centro de Investigación en Matemáticas, A.C. (CIMAT)** que participaron en la maestría, por compartir sus conocimientos conmigo.

A las autoridades del **Instituto Nacional de Estadística y Geografía (INEGI)** por darme la oportunidad de realizar estudios de maestría.

A mi familia: Mi esposo **Pedro Ismael Reyna Gris** y mi hijo **Pedro Iván Reyna Martínez**, por su apoyo incondicional y por cubrir todas mis ausencias y omisiones durante mis estudios.

Contenido

Introducción

1 Preliminares	2
1.1 El Índice Nacional de Precios la Consumidor	2
1.2 Las series de tiempo vistas como procesos estocásticos.	14
1.3 Metodología para la conformación de conglomerados	18
2 Propuesta.....	28
2.1 Medidas de distancia entre series.	28
2.2 Métrica para la clasificación de series de tiempo basada en un periodograma. ...	29
2.3 Medida de distancia utilizada.	37
3 Caso de Estudio.....	51
3.1 Resultados obtenidos por el método de K medias.	54
3.2 Resultados obtenidos por el método PAM.....	63
3.3 Resultados obtenidos por el método AGNES.	71
4 Conclusiones	79
5 Bibliografía.....	90
Anexos	92
Étienne Laspeyres (1834-1913)	93
Ponderadores del Índice Nacional de Precios al Consumidor.	95
La Encuesta Nacional de Ingresos y Gastos en los Hogares.	104
Conglomerado de cada genérico por método de agrupación.....	106
Causalidad.....	113
Error de pronóstico.....	120
Código	126

Introducción

El objetivo de este estudio es agrupar las series que conforman el Índice Nacional de Precios al Consumidor (INPC) con el fin de descubrir si existen patrones de comportamiento en el incremento de los precios de distintos bienes y servicios.

Contamos con 315 series de tiempo que corresponden a los productos genéricos involucrados en el cálculo del INPC. Se utilizarán métodos conocidos de agrupamiento jerárquicos y no jerárquicos.

Se tomará la medida de similitud entre series de tiempo, propuesta por el M.E. José. Ramón Domínguez Molina y la Dra. Graciela González Farías (2001). Esta medida de similitud se basa en las estimaciones de los coeficientes de un modelo autorregresivo de orden grande, en este caso un $AR_{(24)}$, que permitirá captar las dependencias entre los incrementos de precios de dos años de los diferentes bienes y servicios que conforman la canasta del INPC.

Una vez conformados los conglomerados se buscarán las características que comparten las series que las llevaron a ser incluidas en un mismo conglomerado.

1 Preliminares

Una serie de tiempo es el registro metódico de observaciones numéricas de una variable a intervalos de tiempo regulares. Una gran cantidad de información acerca de las características principalmente económicas, de los individuos, las empresas o países, se recopila con fines de análisis.

El Instituto Nacional de Estadística y Geografía (INEGI) integra en el Banco de Información Económica (BIE) más de 100,000 series históricas con información económica de México y de países seleccionados. Indicadores oportunos sobre todos los sectores económicos, así como datos sobre la Ocupación y el Empleo, del Índice Nacional de Precios al Consumidor y del Comercio Exterior, entre otros.

Asimismo, con el propósito de ofrecer a los usuarios una referencia adicional para el manejo y uso de la estadística incluida en el BIE, se incorporan constantemente series estadísticas desestacionalizadas de los principales indicadores económicos de coyuntura.

La Ley del Sistema Nacional de Información Estadística y Geográfica, publicada en el Diario Oficial de la Federación el 16 de abril del 2008 en su artículo 59 otorga al INEGI como una de sus facultades exclusivas elaborar los índices nacionales de precios (al consumidor y al productor) que eran responsabilidad del Banco de México.

1.1 El Índice Nacional de Precios la Consumidor

Un índice de precios es una medida de la proporción o porcentaje de los cambios que experimentan un conjunto de precios en el tiempo. Un índice de precios al consumidor (IPC) mide los cambios en los precios de bienes y servicios que consumen los hogares. Tales cambios afectan el poder adquisitivo real del ingreso de los consumidores y su bienestar. Dado que los precios de los diferentes bienes y servicios no cambian en la misma razón, el índice de precios puede únicamente reflejar el promedio de estos movimientos. Los índices de precios también pueden ser utilizados para medir las diferencias en el nivel de precios en diferentes ciudades, regiones o países en un mismo punto en el tiempo.

En la práctica la mayoría de los índices de precios al consumidor se calculan como promedios ponderados del porcentaje de cambio de los precios para un conjunto específico “canasta” de productos. Los pesos o ponderadores reflejan la importancia relativa en el consumo de un producto en los hogares durante un período. La calidad del índice depende de que tan apropiados y oportunos son los ponderadores.

1.1.1 Orígenes y usos del índice de precios al consumidor.

El índice de precios al consumidor debe servir a un propósito. La forma como se construyen y definen depende en mucho en para qué serán utilizados y por quién. Los índices de precios al consumidor tienen una larga historia que data del siglo XVIII.

Los índices de Laspeyres y Paasche, que aún son muy utilizados, fueron propuestos en la década de 1870 y se explican más adelante. El concepto de índice del costo de la vida fue introducido en el siglo XX.

Tradicionalmente, uno de los propósitos principales de calcular el índice de precios al consumidor fue el de compensar a los asalariados por la inflación, ajustando sus salarios en proporción al porcentaje de cambio del índice de precios al consumidor, este procedimiento es conocido como indexación. Por esta razón la responsabilidad del cálculo del índice de precios al consumidor recaía en los ministerios del trabajo, actualmente en la mayoría de los países lo calculan las oficinas nacionales de estadística.

Los índices de precios al consumidor tienen las siguientes características: se publican con cierta periodicidad, comúnmente, mensual o trimestral. Se obtienen rápidamente, en unas dos semanas después de que finaliza el período de referencia, son cercanamente monitoreados y atraen mucha atención.

El índice de precios al consumidor provee información oportuna acerca de la inflación y es utilizado para diferentes propósitos, además de indexar los salarios, por ejemplo, se emplean para indexar las pensiones y otros beneficios de la seguridad social, también para indexar pagos como intereses, rentas o el valor de otras obligaciones.

Los índices de precios son comúnmente utilizados como una tasa general de inflación, aún cuando miden únicamente la inflación del consumo y son utilizados por algunos gobiernos o bancos centrales para fijar objetivos de inflación con propósitos de política monetaria.

Los precios que se colectan para el cálculo del índice de precios al consumidor también pueden ser utilizados para calcular otros índices, tales como el índice de precios empleado en contabilidad nacional para deflactar los gastos en el consumo de los hogares o la paridad del poder adquisitivo utilizado para comparar los niveles reales de consumo en diferentes países.

La variedad de uso que se da a estos índices puede crear conflictos de interés. Por ejemplo, utilizando el IPC como indicador general de la inflación creará presión para extender su cobertura, incluyendo elementos que no son bienes y servicios consumidos por los hogares, cambiando así la naturaleza y concepto del IPC.

Debido al extenso uso del IPC para indexar una gran variedad de pagos, se mueven grandes sumas de dinero alrededor de sus movimientos, suficientes para tener un impacto significativo en el estado de las finanzas de un gobierno. Así, pequeñas diferencias en los movimientos del IPC, resultado del uso de formulas y métodos ligeramente diferentes, pueden tener implicaciones financieras considerables. La metodología para el cálculo del IPC es importante en la práctica y no solo en la teoría.

El propósito de un número índice será comparar los valores del gasto en el consumo de bienes y servicios en los hogares en dos períodos de tiempo. Conocer que el gasto se ha incrementado en cinco por ciento no provee mucha información si no sabemos cuánto de este cambio es atribuible a cambios en los precios de bienes y servicios y cuánto a cambios en los hábitos de consumo. El propósito de un número índice es descomponer proporcionalmente los cambios en los valores agregados en todos sus componentes de cambio de precios y cantidades.

La intención del IPC es medir la componente de precios en los cambios en el gasto por el consumo de los hogares. Una manera de hacer esto es medir el cambio en el valor agregado, manteniendo las cantidades constantes.

1.1.2 Características del INPC

El índice mide la evolución en el tiempo del precio en el mercado de una canasta de bienes y servicios tomada como representativa de los hábitos de consumo de los hogares urbanos.

Para este estudio se tomaron las series cuya base es la segunda quincena de junio de 2002=100.

El INPC en México, es un índice del tipo propuesto por Laspeyres.

$$I_{t/o} = \frac{\sum_{i=1}^n P_{it} q_{io}}{\sum_{i=1}^n P_{io} q_{io}}$$

La fórmula anterior también puede expresarse:

$$I_{t/o} = \sum_{i=1}^n \frac{P_{io} q_{io}}{\sum_{i=1}^n P_{io} q_{io}} \left(\frac{P_{it}}{P_{io}} \right)$$

Donde

$I_{t/o}$ = Índice que mide la variación promedio ponderado de los precios del periodo (t) en relación con los precios del periodo (o).

$\left(\frac{P_{it}}{q_{io}} \right)$ = Índice de precios relativos del i-ésimo producto en el periodo (t), mide la variación del precio del periodo (t) en relación con el periodo base (o).

$\frac{P_{io}q_{io}}{\sum_{i=1}^n P_{io}q_{io}}$ = Ponderación del precio relativo del i-ésimo producto en el índice y se obtiene dividiendo el valor de las (q) unidades del i-ésimo producto en el periodo base (o) entre el valor total de los (n) productos del índice en el mismo periodo base (o).

La cobertura del índice abarca a consumidores urbanos que viven en poblaciones de más de 20,000 habitantes. La información de los precios se obtiene a partir de 15,000 puntos de venta al menudeo y 5,000 hogares, cada mes se obtienen 190,000 precios de 315 productos genéricos y servicios en 46 ciudades, estas ciudades representan a todas las ciudades pequeñas, medianas y grandes en las siete regiones en las que se ha dividido el país. Para los alimentos se obtiene información diariamente y para el resto de los componentes se obtienen observaciones semanales. Los datos no se ajustan estacionalmente.

La publicación del INPC se realiza dentro de los 10 días a partir del fin del periodo de referencia.

La cobertura de productos se refiere a todos los productos de uso común y servicios adquiridos por los hogares urbanos con propósito de consumo (no reventa) incluyendo impuestos. Se clasifican en 8 grupos, 17 subgrupos, 36 sub subgrupos, 77 conjuntos de productos genéricos y 315 productos genéricos.

Los componentes del INPC se agrupan en ocho grupos: (1) Alimentos, bebidas y tabaco, (2) Ropa, calzado y accesorios, (3) Vivienda, (4) Muebles, aparatos y accesorios domésticos, (5) Salud y cuidado personal, (6) Transporte, (7) Educación y esparcimiento y (8) Otros servicios.

El Índice Nacional de Precios al Consumidor se calcula utilizando para cada concepto de consumo familiar las ponderaciones cuya fuente es La Encuesta Nacional de Ingresos y Gastos en los Hogares, 2000 (ENIGH), realizada por el INEGI. Consiste en una muestra de 12,000 hogares y el Censo de Población y Vivienda 2000., el detalle de los ponderadores por producto se muestra en el anexo dos, para los 8 grandes grupos son los siguientes:

	Ponderaciones
Índice Nacional de Precios al Consumidor	100.00000
Alimentos, bebidas y tabaco	22.74156
Ropa, calzado y accesorios	5.59224
Vivienda	26.40994
Muebles, aparatos y enseres domésticos	4.85674
Salud y cuidado personal	8.57766
Transporte	13.40559
Educación y esparcimiento	11.53647
Otros servicios	6.87980

Las ponderaciones corresponden a la nueva base de cálculo del Índice Nacional de Precios al Consumidor que es la segunda quincena de junio de 2002= 100, y permanecerán constantes durante la vigencia de dicha base de cálculo. Por definición, la fórmula de Laspeyers no admite cambio en sus ponderaciones durante la vigencia de la base de cálculo del índice respectivo.

La información para medir la importancia relativa de los genéricos dentro de la canasta del INPC, se obtiene a partir de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). Así, el gasto que la ENIGH asocia con cada bien o servicio se compara contra el gasto total de las familias mexicanas. De esta comparación se obtiene un cociente mediante la siguiente fórmula:

$$w = \frac{\text{Gasto en el bien o servicio } i \text{ de todas las familias mexicanas}}{\text{Gasto total de las familias mexicanas}}$$

Estos cocientes indican la importancia relativa o el peso de cada satisfactor dentro de los gastos de las familias mexicanas. A su vez, cada uno de esos porcentajes tiene una correspondencia con los genéricos del INPC. Por tanto, los pesos así estimados determinan el impacto que tendrá un cambio en el precio de un genérico dentro del presupuesto familiar.

A cada uno de estos pesos relativos se le conoce como ponderación. Así, al conjunto de todos esos pesos relativos se le denomina estructura de ponderaciones.

La fórmula de Laspeyres, ciudades, estados, zonas conurbadas y ramas de actividad económica, son las que se dieron a conocer en el Diario Oficial de la Federación el 25 de julio de 2002.

El Banco de México conserva por seis meses las cotizaciones de precios utilizadas para calcular el Índice Nacional de Precios al Consumidor del mes de que se trate, las cuales estarán a disposición de los interesados que necesiten consultarlas. Dicha información, así como cualquier otra relativa al mencionado índice podrá ser solicitada a la Subgerencia de Precios de este Banco.

Frecuencia de actualización de los ponderadores: Anteriormente no se contaba con una agenda de actualización de los ponderadores, nuevos ponderadores se calcularon en 1980 utilizando la

ENIGH 1977 y en 1994 utilizando la ENIGH 1989. En el año 2002 se actualizaron tanto los ponderadores como la canasta utilizando la ENIGH 2000 y el Censo de ese mismo año. En el futuro se planea actualizar los ponderadores cada 6 años.

Selección de ítems: 315 productos genéricos cubren el universo de hábitos de consumo de las familias urbanas mexicanas en la ENIGH 2000. Productos específicos fueron seleccionados con probabilidad proporcional al tamaño de las ventas. La selección de establecimientos también se realiza por muestreo probabilístico proporcional al tamaño del volumen de ventas al público.

Tamaños de muestra: Cada mes se recolectan 190,000 precios en 15,000 establecimientos y 5,000 hogares, cubriendo 315 productos genéricos en 46 ciudades.

Método de recolección de los precios: Los precios son recolectados directamente de los establecimientos por enumeradores, excepto las rentas, el gas, la electricidad, el petróleo y los servicios telefónicos, que se recolectan directamente de los hogares. Los precios que se colectan cada día son enviados electrónicamente al Banco de México.

Prácticas de recolección: precios omitidos, no disponibles temporalmente de productos específicos, los recolectores de precios hacen una búsqueda del precio del producto en el establecimiento más cercano y similar al de la muestra. Productos discontinuados son substituidos por nuevos productos tan similares como sea posible.

Diferencias en calidad: cuando la calidad de un producto cambia significativamente, el siguiente método de ajuste se utiliza en el índice: descontar el costo de las características opcionales en el precio de automóviles, aparatos eléctricos, electrónica. Las unidades de precio se calculan de los artículos vendidos por peso o por volumen.

Para garantizar la representatividad de los precios que intervienen en el cálculo del INPC se realiza una selección de fuentes de información en cada una de las 46 ciudades de cotización. Una vez elegidos los establecimientos se lleva a cabo una investigación de marcas y presentaciones para seleccionar los productos específicos de cuyo precio se hace un seguimiento constante. La selección de los servicios se realiza con base en la opinión de los informantes que escogen los más representativos.

El sistema del INPC se integra de 46 ciudades y áreas metropolitanas agrupadas en siete regiones. A su vez, por su tamaño, las ciudades se clasifican en pequeñas, medianas y grandes. De este modo se calculan los índices de precios para cada una de las siete regiones en que se divide el territorio nacional y para las 46 ciudades que conforman el sistema, así como para cada tamaño de localidad. Cabe señalar que al menos en una ciudad por estado se recoge información para el cálculo del INPC. De esta manera se asegura la representatividad espacial del INPC.

Las ciudades que se agrupan en cada región son las siguientes:

<p>Región Frontera Norte</p> <p>Mexicali, B.C. Cd. Juárez, Chih. Tijuana, B.C. Matamoros, Tamps. La Paz, B.C.S. Cd. Acuña, Coah.</p>	<p>Región Centro Norte</p> <p>Guadalajara, Jal. (área metropolitana) León, Gto. (área metropolitana) San Luis Potosí, S.L.P. (área metropolitana) Morelia, Mich. Aguascalientes, Ags. (área metropolitana) Colima, Col. (área metropolitana) Jacona, Mich. Cortázar, Gto. Querétaro, Qro. (área metropolitana) Tepatitlán, Jal.</p>
<p>Región Noroeste</p> <p>Culiacán, Sin. Hermosillo, Son. Tepic, Nay. Huatabampo, Son.</p>	<p>Región Metropolitana</p> <p>Área metropolitana de la Ciudad de México</p>
<p>Región Noreste</p> <p>Monterrey, N.L. (área metropolitana) Torreón, Coah. (área metropolitana) Tampico, Tamps. (área metropolitana) Chihuahua, Chih. Monclova, Coah. (área metropolitana) Fresnillo, Zac. Cd. Jiménez, Chih. Durango, Dgo.</p>	<p>Región Centro Sur</p> <p>Puebla, Pue. (área metropolitana) Veracruz, Ver. (área metropolitana) Córdoba, Ver. (área metropolitana) Acapulco, Gro. Toluca, Méx. (área metropolitana) Iguala, Gro. Tulancingo, Hgo. Cuernavaca, Mor. Tlaxcala, Tlax. (área metropolitana) San Andrés Tuxtla, Ver.</p>
<p>Región Sur</p> <p>Mérida, Yuc. (área metropolitana) Tapachula, Chis. Villahermosa, Tab. Oaxaca, Oax. (área metropolitana) Campeche, Camp. Chetumal, Q.R. Tehuantepec, Oax.</p>	

El cálculo del INPC considera cuatro estratos diferentes de acuerdo con el nivel de ingreso de los hogares. Estrato 1. Hogares con ingresos menores que un salario mínimo. Estrato 2. Hogares con ingresos entre uno y tres salarios mínimos. Estrato 3. Hogares con ingresos entre tres y seis salarios mínimos. Estrato 4. Hogares con ingresos mayores a seis salarios mínimos.

Es posible consultar los índices por estrato en el ámbito nacional. Asimismo, éstos pueden conocerse también por región. En el ámbito nacional, se calculan índices por estrato de ingreso para cada uno de los ocho grandes grupos de bienes y servicios, clasificados por objeto de gasto.

El año base es el punto de referencia en el tiempo a partir del cual se efectúan las comparaciones del cambio en los precios.

En México, desde el comienzo de la medición del INPC en 1968, se han hecho cuatro cambios de base (1978, 1980, 1994 y 2002).

La teoría moderna sobre la medición de la inflación señala que a medida que la canasta y la base de ponderación de los índices de precios se alejan del periodo de comparación o punto de referencia, la inflación estimada tiende a presentar un sesgo a la alza.

Las distorsiones que puede presentar un índice de precios que se calcula con una canasta y una base de ponderadores muy antigua son: Un sistema de ponderadores antiguo no refleja la transformación de la estructura del consumo a que normalmente dan lugar los cambios relativos que experimentan los precios. Es decir, en la medida de lo posible los consumidores tienden a sustituir en su canasta los artículos de mayor precio por otros más baratos.

Un sistema de ponderadores obsoleto es incapaz de reflejar las modificaciones en la estructura de consumo de los hogares que se derivan de los cambios económicos y sociodemográficos que tienen lugar a través del tiempo.

Una canasta que deja de ser plenamente representativa no capta las alteraciones de los hábitos de compra de los consumidores causadas por la aparición de nuevos productos y servicios en el mercado.

Una canasta desactualizada es incapaz de detectar los cambios en las conductas de consumo derivadas de la mejora de los atributos físicos, de calidad y de la comercialización de los bienes y servicios que se ofrecen en el mercado.

La inflación subyacente es el aumento continuo de los precios de un subconjunto del INPC que contiene a los genéricos con cotizaciones menos volátiles. La inflación subyacente mide la tendencia inflacionaria de mediano plazo y su comportamiento se ve influido principalmente por variables macroeconómicas como el tipo de cambio y los salarios.

Su estimación consiste en separar de forma sistemática y predeterminada del cálculo aquellos componentes del INPC cuyos precios tienen un comportamiento que se sabe, por experiencia,

difieren frecuentemente de la tendencia general del resto de los genéricos que forman el sistema de precios.

Para identificar a los genéricos más volátiles, se calculan las tasas mensuales de crecimiento de sus respectivos precios y se hace una comparación de éstas con una medición de la tendencia mensual de la inflación. Esta última se estima mediante cuatro métodos distintos (X12 ARIMA, Promedio de los próximos 12 meses, Media Móvil Centrada de 12 meses, Inflación mensual del INPC).

Una vez definida la tendencia de la inflación mensual, se calculan mes con mes las desviaciones entre ésta y el crecimiento mensual del precio de cada genérico. Al elevar estas desviaciones al cuadrado, sumarlas a lo largo de toda la muestra y dividir por el número de meses, se obtiene una medición de la variabilidad relativa de cada genérico que se denomina error cuadrático medio. Gran parte del descenso del error cuadrático medio se logra vía la eliminación del 35 por ciento de los genéricos más volátiles, mismos que se muestran en el siguiente cuadro:

Genéricos Más Volátiles Dentro del INPC

Genérico	Método que lo sitúa en el 35% más volátil	Genérico	Método que lo sitúa en el 35% más volátil
Tomate verde	1,2,3,4	Electricidad	1,2,3,4
limón	1,2,3,4	Tenencia automóvil	1,2,3,4
cebolla	1,2,3,4	Metro o transporte eléctrico	1,2,3,4
Chile poblano	1,2,3,4	Larga distancia internacional	1,2,3,4
Chícharo	1,2,3,4	Autobus urbano	1,2,3,4
Chayote	1,2,3,4	Larga distancia nacional	1,2,3,4
Chile serrano	1,2,3,4	Transporte colectivo	1,2,3,4
Calabacita	1,2,3,4	Cuotas de autopista	1,2,3,4
Jitomate	1,2,3,4	Taxi	1,2,3,4
Ejotes	1,2,3,4	Servicio telefónico local	1,2,3,4
Aguacate	1,2,3,4	Estacionamiento	1,2,3,4
Nopales	1,2,3,4	Autobús foráneo	1,2,3,4
Pepino	1,2,3,4	Gas doméstico	1,4
Naranja	1,2,3,4	Jardín de niños y guardería	1,2,3,4
Guayaba	1,2,3,4	Primaria	1,2,3,4
Plátano tabaco	1,2,3,4	Secundaria	1,2,3,4
Toronja	1,2,3,4	Preparatoria	1,2,3,4
Otros chiles frescos	1,2,3,4	Universidad	1,2,3,4
Melón	1,2,3,4	Carrera corta e idiomas	1,2,3,4
Piña	1,2,3,4	Otras diversiones	1,2,3,4
Papaya	1,2,3,4	Tortilla de maíz	1,2,3,4
Ajo	1,2,3,4	Joyas y bisutería	1,2,3,4
Mango	1,2,3,4	Cerveza	1,3,4
Uva	1,2,3,4	Seguro de automóvil	1,2,3,4
Col	1,2,3,4	Cigarrillos	1,3,4
Zanahoria	1,2,3,4	Ferrocarril	1,2,3,4
Sandía	1,2,3,4	Masa de maíz	1,2,3,4
Otros plátanos	1,2,3,4	Pan de caja	1,2,3,4
Papa	1,2,3,4	Otras legumbres secas	1,2,3,4
Lechuga	1,2,3,4	Leche sin envasar	1,2,3,4
Durazno	1,2,3,4	Pantalón de niño o materiales	1,2,3,4
Frijol	1,2,3,4	Leche pasteurizada envasada	1,2,3,4
Chile seco	1,2,3,4	Instrumentos musicales y otros	1,2,3,4
Manzana	1,2,3,4	Papel higiénico	1,2,3,4
Elote	1,2,3,4	Pañuelos desechables	1,2,3,4

Genéricos Más Volátiles Dentro del INPC

(continuación)

Genérico	Método que lo sitúa en el 35% más volátil	Genérico	Método que lo sitúa en el 35% más volátil
Otras legumbres	1,2,3,4	Club deportivo	1,2,3
Pera	1,2,3,4	Leche evaporada	1,2,3
Camarón	1,2,3,4	Velas y veladoras	1,2,3,4
Mojarra	1,2,3,4	Azúcar	1,2,3
Otros pescados	1,2,3,4	Material de curación	1,3
Robalo y Mero	1,2,3,4	Pastelillos y pasteles	1,2,3,4
Otros mariscos	1,2,3,4	Servicio doméstico	1
Huachinango	1,2,3,4	Salchichas	1
Pollo en piezas	1,2,3,4	Muebles para cocina	1,2,4
Pollo entero	1,2,3,4	Medias y pantimedias	1,2,3
Pierna	1,2,3,4	Cajetas	1,3,4
Huevo	1,2,3,4		

*Los métodos utilizados para medir la tendencia inflacionaria son: (1) ARIMA, (2) Promedio de los próximos 12 meses, (3) Media móvil centrada de 12 meses, (4) INPC.

Fuente: Banco de México. Metodología para el Cálculo de la Inflación Subyacente.

De los 93 genéricos que constituyen el 35 por ciento de la canasta del INPC, 66 pertenecen a algunos de los siguientes subíndices: Precios Administrados y Concertados por el Sector Público, Productos Agropecuarios y Educación. Por tanto, estos subíndices están casi totalmente representados dentro del subconjunto más volátil.

Los errores cuadráticos medios, con respecto a la tendencia de la inflación, calculados para las variaciones mensuales, del INPC sin el subconjunto más volátil y del INPC sin los subíndices de Productos Agropecuarios, de Educación y de Precios Administrados y Concertados por el Sector Público son casi iguales.

Con base en los resultados anteriores, se define la Inflación Subyacente como la medición que elimina del INPC los subíndices Productos Agropecuarios, de Educación y de Precios Administrados y Concertados por el Sector Público.

Una clase muy amplia y popular de índices de precios se obtiene definiendo al índice como un porcentaje de cambio entre dos períodos comparados en el costo total de compra de un conjunto dado de cantidades, generalmente descrita como canasta. El significado de tal índice es fácil de comprender y de explicar a los usuarios. Algunos ejemplos de estos índices son el INPC respecto al mes anterior, el INPC respecto al mismo mes del año anterior, etc.

El Banco de México es una institución autónoma que opera bajo la Ley del Banco de México publicada en el Diario Oficial de la Federación el 23 de diciembre de 1993. La Ley autoriza al Banco

de México a publicar estadísticas económicas y financieras. El artículo 62 (1) “ El Banco de México podrá: I. En coordinación con las demás autoridades competentes, elaborar, compilar y publicar estadísticas económicas y financieras, así como operar sistemas de información basados en ellas y recabar los datos necesarios para esos efectos”.

Confidencialidad: Los datos se proveen en forma agregada, se mantiene la confidencialidad con respecto a instituciones, individuos y empresas.

Política de calidad: Un sistema de aseguramiento de la calidad (bajo la certificación ISO 9000:2000) se estableció para el índice nacional de precios al consumidor. Esto implica que, cada mes, cuatro ciudades están sujetas a un proceso de auditoría para certificar la calidad de la investigación de los precios. En corto tiempo de 3 a 5 supervisores visitan el mismo número de ciudades diferentes para verificar los precios reportados y los procedimientos. Cada 6 meses todo el sistema es verificado dos veces por auditores de procesos internos y externos.

1.2 Las series de tiempo vistas como procesos estocásticos.

Un proceso estocástico es una familia de variables aleatorias asociadas a un conjunto índice de números reales, de tal forma que a cada elemento del conjunto le corresponde una y sólo una variable aleatoria.

$Z(\tau); \tau \in T$, en donde T es el conjunto índice y $Z(\tau)$ es la variable aleatoria correspondiente al elemento τ de T .

Si T es un intervalo de números reales, ya sea abierto o cerrado, se dirá que el proceso estocástico es continuo.

Si T es un conjunto finito o infinito pero numerable, se dirá que el proceso estocástico es discreto.

El hecho de que el proceso estocástico sea continuo o discreto, no indica nada acerca de la naturaleza de las variables aleatorias involucradas, ya que estas pueden ser discretas o continuas.

Una serie de tiempo es la sucesión de observaciones generadas por un proceso estocástico, cuyo conjunto índice se toma en relación con el tiempo, así como existen procesos estocásticos discretos y continuos, también existen series de tiempo discretas y continuas.

Si las observaciones de una serie de tiempo discreta se toman en los momentos $\tau_1, \tau_2, \dots, \tau_N$, el proceso estocástico respectivo se denotará $Z_{(\tau_1)}, Z_{(\tau_2)}, \dots, Z_{(\tau_N)}$.

Es importante notar que una serie de tiempo observada no es más que una realización de un proceso estocástico.

1.2.1 Uso de operadores y polinomios de retraso.

El operador de retraso se denota por la letra B y se define como:

$$Bz_t = z_{t-1} \text{ para toda } t$$

Por la aplicación sucesiva del operador B se obtiene:

$$B^2 z_t = B(Bz_t) = z_{t-2}$$

$$B^3 z_t = B(B^2 z_t) = z_{t-3}$$

:

$$B^k z_t = B(B^{k-1} z_t) = z_{t-k},$$

De esta forma a la expresión general a que se llega es:

$$B^k z_t = z_{t-k} \text{ para } k=0,1,2,\dots \text{ y toda } t.$$

Otro operador de uso frecuente y que está íntimamente ligado a B es el operador de diferencia ∇ y se define como:

$$\nabla Z_t = Z_t - Z_{t-1} \text{ para toda } t.$$

La relación que liga a ∇ con B es la siguiente:

$$\nabla = 1 - B \text{ o sea } \nabla Z_t = (1 - B)Z_t$$

Así como se obtuvo la expresión general para B^k , puede obtenerse la siguiente forma general para ∇^k .

$$\nabla^k Z_t = (1 - B)^k Z_t.$$

El uso de polinomios de retraso es importante porque es útil para expresar de manera simple algunos de los modelos de mayor utilidad en la práctica:

Uno de estos modelos es el de *promedios móviles (MA)*, que puede representarse mediante la expresión:

$$Z_t - \mu = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) a_t$$

En donde μ denota a la media de la serie, o dicho de otra manera el "nivel" del proceso.

$Z_t - \mu$ representa la desviación de Z_t respecto de la media.

a_t es una sucesión de variables aleatorias.

$\theta_1, \theta_2, \dots, \theta_q$ son parámetros que relacionan las sucesiones a_t y Z_t

El modelo de **promedios móviles (MA)** se denota de manera compacta:

$$Z_t - \mu = \theta(B) a_t$$

Así mismo los modelos **autorregresivos (AR)** se definen como:

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) Z_t - \mu = a_t$$

En donde $\phi_1, \phi_2, \dots, \phi_p$ son parámetros, se expresan mediante la relación

$$\phi(B)(Z_t - \mu) = a_t$$

A las combinaciones de los modelos anteriores, se les conoce con el nombre de modelos **autorregresivos de promedios móviles (ARMA)** y se les representa con la expresión:

$$\phi(B)(Z_t - \mu) = \theta(B)a_t$$

Por último, utilizando polinomios de retraso y del operador diferencia se obtienen las representaciones de los modelos **autorregresivos integrados y de promedios móviles (ARIMA)**.

$$\phi(B)\nabla^d Z_t = \theta(B)a_t$$

1.2.2 Toda serie de tiempo puede modelarse como un AR_∞

El modelo autorregresivo es muy útil en la representación de series de tiempo, en este modelo el valor actual de la serie se expresa como una ponderación de los valores pasados del mismo proceso y un error aleatorio a_t

Definiendo $\tilde{z}_t = z_t - \mu$, entonces el proceso autorregresivo de orden p ($AR_{(p)}$), puede escribirse como:

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \dots + \phi_p \tilde{z}_{t-p} + a_t$$

Definiendo el operador autorregresivo de orden p por:

$$\phi(B) = 1 - \phi_1 B + \phi_2 B^2 + \dots + \phi_p B^p$$

Se puede escribir el proceso autorregresivo como:

$$\phi(B)\tilde{z}_t = a_t$$

Una propiedad importante es la invertibilidad:

En general un proceso es invertible, si puede tener la representación:

$$\pi(B)z_t = a_t$$

$$(1 - \pi_1 B + \pi_2 B^2 + \dots + \pi_p B^p)z_t = a_t$$

En donde los procesos π_j son absolutamente sumables, esto es $\sum_{j=0}^{\infty} \pi_j < \infty$, lo que implica que la serie $\pi(B)$ converge sobre o dentro del círculo unitario.

Definición de la medida de distancia entre series.

Sea a_t *i.i.d.* $\square N(0, \sigma^2)$ y

Z_t un proceso estocástico con media cero tal que

$$Z_t \square ARIMA(p, d, q)(P, D, Q)_s$$

Esta expresión puede escribirse utilizando la notación estándar de Box y Jenkins.

$$\varphi(B)Z_t = \theta(B)a_t \dots\dots\dots(1)$$

donde

$$\varphi(B) = \phi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D \text{ y}$$

$$\theta(B) = \theta_q(B)\Theta_q(B^s)$$

Despejamos a_t de (1) se tiene

$$\frac{\varphi(B)}{\theta(B)}Z_t = a_t \dots\dots\dots(2)$$

Asumiendo que este modelo es invertible, es decir que, las raíces del polinomio $\theta(B)$ estén fuera del círculo unitario, entonces $Z_t \in \ell$, donde ℓ es la clase de modelos invertibles ARIMA.

Debido a que se trata de un modelo invertible, es posible reescribir (2) como.

$$\pi(B)Z_t = a_t$$

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

$$Z_t = \bar{Z}_{t-1} + a_t$$

$$\bar{Z}_{t-1} = \sum_{j=1}^{\infty} \pi_j Z_{t-j} = 1 - \pi(B) Z_t$$

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

$$Z_t = \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \dots + a_t$$

Donde el operador $AR_{(\infty)}$ está definido por $\pi(B) = \varphi(B) / \theta(B) = 1 - \pi_1 B - \pi_2 B - \dots$.

Por lo tanto, dados valores iniciales y ordenes conocidos, cualquier proceso $Z_t \in \ell$ es completamente caracterizado por la secuencia $\pi' = (\pi_1, \pi_2, \dots)$ la cual especifica completamente la distribución del proceso W_t y Z_t cuando a_t es un proceso de ruido blanco.

La secuencia de π contiene toda la información útil acerca de la estructura del proceso estocástico ya que cualquier otra información necesaria para especificar Z_t es justamente a_t , la cual es impredecible en el tiempo $t-1$.

Entonces una medida de diversidad estructural entre $X_t \in \ell$ y $Y_t \in \ell$ puede ser obtenida comparando sus respectivas secuencias de π .

Por lo que una métrica sobre ℓ puede definirse por la distancia:

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

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

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

1.3 Metodología para la conformación de conglomerados

De manera intuitiva y casi subconsciente, todos los seres humanos aprendemos a clasificar objetos, desde niños podemos distinguir hombres de mujeres, perros de gatos, etc. (esto explica porqué el análisis de conglomerados es considerado como una rama de patrones de reconocimiento e inteligencia artificial. La clasificación juega un rol importante en la ciencia. En los últimos 30 años un gran número de algoritmos y programas de computadora se han desarrollado para el análisis de conglomerados.

Existen diferentes métodos de agrupamiento según el tipo de datos que vamos a analizar, en nuestro caso se trata de una variable continua.

Supongamos que tenemos n objetos a ser agrupados, que pueden ser, digamos, personas. Los algoritmos para hacer conglomerados, típicamente trabajan de dos maneras. La primera de ellas representa a los objetos a través de p medidas o atributos, en el ejemplo de las personas podrían ser la altura, el peso, la edad, el sexo, raza, etc. Estas medidas pueden arreglarse en una en una matriz de $n \times p$, en donde los renglones corresponden a los objetos y las columnas a los atributos. En la terminología de Tucker (1964) se conoce a estas matrices como de dos vías, dado que los objetos en los renglones y sus atributos en las columnas son dos entidades diferentes.

La segunda estructura es una colección de similitudes (proximidades) que deberá calcularse para todo par de objetos. Estas similitudes conforman una matriz de $n \times n$ que es llamada matriz de una vía dado que las entidades en los renglones y las columnas son el mismo conjunto de objetos. Es posible trabajar con similitudes (medir qué tan parecidos o cercanos son dos objetos) o disimilaridades (miden qué tan lejos están dos objetos).

En el caso de las variables continuas n objetos son caracterizados por p medidas continuas. Esta información puede ser organizada en una matriz de $n \times p$, donde los renglones corresponden a los objetos (o casos) y las columnas corresponden a las variables:

(p) variables

$$(n) \text{objetos} \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & \dots & \dots & \dots \\ \vdots & \dots & \dots & \dots \\ x_{n1} & \dots & \dots & x_{np} \end{pmatrix}$$

En algunas aplicaciones, cambiar la unidad de medida de las observaciones puede producir estructuras de agrupamiento diferentes. Por ello es recomendable que los datos que ingresen al análisis sean producto de una transformación. Las más comunes son sustituir a la variable por su inverso, o su cuadrado o estandarizar los datos.

El siguiente paso consiste en calcular las distancias entre los objetos i y j . La más conocida es la distancia Euclídeana:

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

Esta fórmula corresponde a la distancia geométrica entre los puntos cuyas coordenadas son

$$(x_{i1}, \dots, x_{ip}) \text{ y } (x_{j1}, \dots, x_{jp})$$

La distancia Euclideana satisface los siguientes requerimientos matemáticos, de toda función de distancia:

- i. $d(i, j) \geq 0$
- ii. $d(i, i) = 0$
- iii. $d(i, j) = d(j, i)$
- iv. $d(i, j) \leq d(i, h) + d(h, j)$

Para todos los objetos i, j y h .

Disimilaridades

En adelante ya no hablaremos de distancias sino de disimilaridades. Las disimilaridades son números no negativos $d(i, j)$ que son pequeñas (cerca de cero) cuando i y j están cerca uno del otro y se hacen grandes cuando i y j son muy diferentes. La diferencia con distancias es que el punto IV ya no es válido.

Las disimilaridades pueden obtenerse de muchas maneras, con frecuencia deben obtenerse a partir de distintos tipos de variables como son binarias, nominales, ordinales, de intervalo o bien de una combinación de éstas. Las disimilaridades también pueden ser medidas subjetivas acerca de que tanto difieren unos objetos de otros.

Si se desea realizar un análisis de conglomerados de un conjunto de variables que se han observado en una población existen diferentes medidas de disimilaridad por ejemplo es posible calcular la correlación paramétrica de Pearson:

$$R(i, g) = \frac{\sum_{i=1}^n (x_{if} - m_f) (x_{ig} - m_g)}{\sqrt{\sum_{i=1}^n (x_{if} - m_f)^2} \sqrt{\sum_{i=1}^n (x_{ig} - m_g)^2}}$$

O su equivalente en la correlación no paramétrica de Spearman. Ambos coeficientes se mueven entre -1 y 1 y su valor no depende de las unidades de medida de las variables. Los coeficientes de correlación son útiles en el análisis de conglomerados porque miden que tan relacionadas están dos variables.

Los coeficientes de correlación paramétricos y no paramétricos, pueden ser convertidos a disimilaridades de la siguiente forma:

$$d(f, g) = (1 - R(f, g)) / 2$$

Con esta fórmula, las variables con una estrecha relación positiva, reciben un coeficiente de disimilaridad cercano a cero, mientras que variables con una estrecha correlación negativa serán consideradas disimiles. En otras aplicaciones es preferible utilizar:

$$d(f, g) = 1 - |R(f, g)|$$

En cuyo caso a las variables con una estrecha correlación negativa les será asignada una medida de disimilaridad pequeña.

En muchas aplicaciones los datos de entrada consisten en una matriz de disimilaridades y no se cuenta con los datos originales. Por esta razón es útil contar con algoritmos que pueden operar directamente con una matriz de disimilaridades, sin tener que recurrir a otra medida. Tal es el caso de los programas PAM, FANNY, AGNES y DIANA que describiremos brevemente más adelante.

Similaridades

En lugar de utilizar un coeficiente de disimilaridad $d(i, j)$ para mostrar que tan distantes están dos objetos i y j , es posible utilizar coeficientes de similaridad $s(i, j)$. Entre más parecidos o cercanos estén dos objetos i, j , más grande se hace $s(i, j)$. Tal medida de similaridad típicamente toma valores entre 0 y 1, donde 0 significa que los objetos no son similares en absoluto y 1 refleja la máxima similaridad. Valores entre 0 y 1 indican diferentes grados de semejanza. Se asumen las siguientes condiciones:

- i. $0 \leq s(i, j) \leq 1$
- ii. $s(i, i) = 1$
- iii. $s(i, j) = s(j, i)$

Para todos los objetos i, j . Los valores $s(i,j)$ pueden colocarse en una matriz n por n que será llamada matriz de similaridad. Ambas matrices de similaridad o disimilaridad son llamadas matrices de proximidad o matrices de semejanzas.

Con el fin de definir las similaridades entre variables, podemos recurrir nuevamente a los coeficientes de correlación de Pearson y Spearman. Es importante resaltar que ninguno de ellos puede ser utilizado directamente como medida de similaridad ya que pueden tomar valores negativos. Se requiere de una transformación para traerlos a valores de cero a uno. Existen dos formas de hacer esto dependiendo del sentido de los datos y del propósito de la aplicación: Si las variables deben ser consideradas como muy diferentes porque están orientadas en dirección opuesta, es mejor utilizar:

$$s(f, g) = (1 + R(f, g)) / 2$$

Que da como resultado $s(f,g)=0$ cuando $R(f,g)=-1$. Por otro lado, existen situaciones las variables con una estrecha relación negativa deben ser agrupadas porque miden esencialmente lo mismo (por ejemplo para reducir el número de variables en el análisis de regresión), en ese caso es mejor utilizar la expresión:

$$s(f, g) = |R(f, g)|$$

Que reporta $s(f,g)=1$ cuando $R(f,g)=1$.

En el caso en que los datos consistan en una matriz de similaridades y se desea utilizar un algoritmo diseñado para disimilaridades, será necesario transformar las similaridades en disimilaridades. Entre más grande sea $s(i, j)$, más pequeña será $d(i, j)$, por lo que se requiere una transformación decreciente como la siguiente:

$$d(i, j) = 1 - s(f, g)$$

Para trabajar con similaridades y coeficientes de correlación, resulta de utilidad el programa DAISY.

La selección de un algoritmo para la construcción de conglomerados depende tanto del tipo de datos de que se dispone, como de propósito particular que se busca. En ocasiones varios algoritmos resultan aplicables y es recomendable correr más de un programa para analizar y comparar los resultados, la interpretación de estos resultados debe tener sentido con los datos originales.

1.3.1 Algoritmos no jerárquicos o de partición.

En este trabajo utilizaremos dos tipos de algoritmos para la elaboración de conglomerados conocidos como de partición y jerárquicos. Un algoritmo de partición construye k conglomerados, esto es clasifica los datos en k grupos, que juntos satisfacen dos requerimientos:

Cada grupo contiene por lo menos a un objeto.

Cada objeto pertenece solamente a un grupo.

Estas condiciones implican que como máximo se tiene tantos grupos como objetos.

$$k \leq n$$

La segunda condición implica que dos conglomerados diferentes no tienen elementos en común y que los k grupos juntos forman el conjunto de datos completo.

Una característica importante es que k es asignado por el investigador, por lo que es previsible que se ejecutara el algoritmo diferente número de veces para diferentes valores de k . El valor definitivo de k resultará del análisis gráfico que provea mayor significado a la interpretación. La selección de k también puede realizarse a través de un algoritmo computación deteniéndose bajo algún criterio numérico. El objetivo es descubrir una estructura que se encuentra presente en los datos.

Algoritmos de partición alrededor de medioides (PAM).

El programa PAM crea conglomerados de objetos que son medidos en p variables con escala de intervalos y es posible utilizarlo cuando se cuenta como datos de entrada con una matriz de disimilaridades.

Con el fin de obtener k conglomerados, el método selecciona k objetos (llamados representantes) del conjunto de datos. Los conglomerados se forman asignando cada objeto restante al conglomerado cuyo representante esté más cercano. La clave es seleccionar a los representantes que en algún sentido estén localizados en el centro de los conglomerados. Para ser exactos, la distancia promedio del objeto representante a todos los objetos del mismo conglomerado debe ser minimizada. Por esta razón a tan buen representante se le conoce como mediodide del conglomerado y al método de partición alrededor de los medioides, se le conoce como técnica de las k medias. Por construcción, el método de las k medias trata de formar conglomerados esféricos.

Clustering Large Applications (CLARA).

El programa CLARA, fue desarrollado con el propósito de analizar conjuntos de datos grandes. Su objetivo de aglomeración es el mismo que el programa PAM, pero se pierden algunas características importantes como es el hecho de que CLARA no acepta como entrada una matriz de disimilaridades (debido a que sería demasiado grande).

El programa CLARA funciona de la siguiente manera: inicia utilizando un generador de números aleatorios para seleccionar una muestra de objetos del conjunto de datos y forma conglomerados de la misma forma que el programa PAM. Después cada objeto del conjunto de datos es asignado al conglomerado cuyo medianoide está más cercano, se repite este proceso y se retiene el resultado que mejor cumpla el objetivo.

Dado que estamos hablando en el marco del método de las k medias, CLARA comparte la robustez de PAM.

Fuzzy Analysis (FANNY).

El programa FANNY puede ser aplicado a los mismos conjuntos de datos que el programa PAM, pero este algoritmo es de diferente naturaleza, FANNY evita las decisiones difíciles en lugar de decir “el objeto a pertenece al conglomerado l ”, dice “el objeto a pertenece con un 90% al conglomerado 1, con 5% al conglomerado 2 y con 5% al conglomerado 3”, lo que significa que a será asignado al conglomerado 1 pero existe una pequeña duda a favor de los conglomerados 2 y 3.

Desafortunadamente los cálculos son muy complejos y por lo mismo nada transparentes o intuitivos.

Otra desventaja es que proporciona como salida una matriz de n por k con coeficientes de membresía (probabilidad de pertenencia a un conglomerado), que pueden ser difíciles de interpretar tan solo por su tamaño.

Una de las metas de este trabajo es encontrar la técnica más adecuada, para agrupar las series de tiempo en k grupos tan homogéneos como sea posible, teniendo en cuenta que por tratarse de observaciones pertenecientes a series de tiempo, las observaciones aparecen en un orden específico que no es posible cambiar sin alterar el resultado. Adicionalmente, es muy importante tener en cuenta que en este tipo de observaciones existe dependencia, entre más cercanas sean las observaciones, la dependencia será mayor.

El objetivo del análisis de conglomerados es agrupar elementos en grupos homogéneos en función de similitudes entre ellos. También es posible utilizarlo para agrupar variables con fines de disminuir la dimensión, pero ese no es el objetivo de éste trabajo.

Al agrupar observaciones, debemos tener en cuenta que (a) Todo elemento debe pertenecer a uno y solamente a uno de los grupos; (2) todo elemento debe clasificarse y (3) cada grupo debe ser internamente homogéneo.

Uno de los métodos no jerárquicos clásicos para la partición en grupos es el conocido como algoritmo de las k-medias y se describe a continuación:

Se tiene una muestra de n elementos y p variables. Se quiere dividir a la muestra en G grupos, para ello se deben seguir cuatro pasos:

- (1) Formar G grupos iniciales. Esto puede hacerse de manera arbitraria, por ejemplo asignando los elementos de manera aleatoria.
- (2) Se calculan los G centros.
- (3) Se calculan las distancias euclidianas de cada elemento a los centros de los G grupos y se reasigna cada elemento al grupo cuyo centro está más próximo. Esta asignación es secuencial, al ingresar un nuevo elemento en un grupo se recalcula el centro de ese nuevo grupo
- (4) Se define un criterio de optimalidad y se verifica si éste mejora al reasignar un nuevo elemento a un grupo y de no ser así termina el proceso.

1.3.2 Algoritmos Jerárquicos

Los algoritmos jerárquicos no construyen una sola partición con k conglomerados, trabajan con todos los valores de k en una misma corrida. Desde $k=1$ (todos los objetos están en el mismo conglomerado) hasta $k=n$ (cada objeto forma un conglomerado con un solo elemento).

Existen dos tipos de algoritmos jerárquicos: los aglomerativos y los divisivos. Ambos construyen sus jerarquías en dirección opuesta, posiblemente con resultados muy distintos. Los algoritmos aglomerativos inician cuando todos los objetos están separados (esto es en el paso cero tenemos n conglomerados) y en cada paso surgen dos conglomerados hasta que finalmente solo queda uno. Por otro lado los algoritmos divisivos inician cuando todos los objetos están juntos (esto es en el paso cero solamente existe un conglomerado) y en cada paso un conglomerado es dividido, hasta que se tienen n conglomerados. El resultado puede coincidir pero usualmente es diferente.

Un método de partición trata de seleccionar los mejores conglomerados con k grupos, lo que no es el objetivo de los métodos jerárquicos. Un método jerárquico tiene el defecto de que no puede reparar lo hecho en pasos anteriores. La rigidez de los métodos jerárquicos constituye tanto la

clave de su éxito (requiere de pocos cálculos) como su principal desventaja (no corrige decisiones erróneas).

Agglomerative Nesting (AGNES).

El programa AGNES acepta exactamente los mismos conjuntos de datos que PAM y FANNY. De hecho los diálogos de interacción son casi idénticos, excepto claro que el número de conglomerados k no es requerido. El programa es aglomerativo, en el primer paso, los dos objetos más cercanos se unen quedando $n-1$ conglomerados, uno de los cuales tiene 2 objetos mientras que el resto tienen solamente uno. En los siguientes pasos los dos conglomerados más cercanos se unen, por lo que se requiere de una definición de disimilaridad entre conglomerados, existen varias definiciones, el algoritmo AGNES utiliza el método de promedios no ponderados para pares de grupos, UPGMA por sus siglas en inglés (unweighted pair group average method). En este método la disimilaridad entre dos conglomerados R y Q se calcula como el promedio de todas las disimilaridades $d(i, j)$ en donde i es un elemento de R y j es un elemento de Q .

Divisive Analysis (DIANA).

El programa DIANA analiza el mismo tipo de datos que AGNES, el usuario podría pensar que ambos programas son idénticos ya que pueden ser ejecutados de misma manera y proporcionan salidas muy similares, excepto que la construcción de las jerarquías se realiza en dirección opuesta.

Debido a que los conglomerados grandes se determinan primero, uno podría detener el algoritmo en un paso en el que ya no se tenga interés por más divisiones.

Por otro lado los algoritmos de división presentan algunos problemas computacionales al menos en un principio, por ello no son muy populares y no están generalmente disponibles.

DIANA es conveniente para conglomerados aproximadamente redondos al igual que el método de promedios implementado en AGNES, con frecuencia es útil comparar las salidas del programa AGNES y de DIANA aplicados a un mismo conjunto de datos.

Es posible estructurar los elementos de una muestra en forma jerárquica por similitud. En una clasificación jerárquica, se ordenan los elementos por niveles, de manera que los niveles superiores contienen a los inferiores. Un ejemplo de este tipo de construcción es el dendograma.

El dendograma también es conocido como árbol jerárquico y se construye de la siguiente manera:

En la parte inferior del gráfico se dispone de n elementos iniciales.

Se utilizan tres líneas rectas para unir los elementos, dos dirigidas a los elementos que se unen perpendiculares al eje de los elementos y una paralela que se dibuja en el nivel en que se unen.

El proceso termina cuando todos los elementos están conectados por líneas rectas.

2 Propuesta

2.1 Medidas de distancia entre series.

En la práctica con frecuencia pueden encontrarse casos en los que se requiera el análisis de un gran número de series de tiempo, por lo que resulta de utilidad contar con técnicas que permitan agrupar las series en grupos tan homogéneos como sea posible y modelar cada grupo con una familia más pequeña de modelos.

Existen diferentes propuestas para agrupar series de tiempo, una de ellas es la de Kakizawa, Shumway y Taniguchi (1998) quienes sugieren que las similitudes y diferencias entre series de tiempo pueden ser caracterizadas en términos de la estructura de covarianza, contenida en su distribución espectral. Una restricción en la metodología propuesta por ellos es que se debe trabajar con series estacionarias.

Otra trabajo realizado por Piccolo (1990) propone ajustar un modelo ARIMA a un gran número de series con el propósito de hacer predicciones y ajustes estacionales. Observando los modelos ajustados es posible encontrar similitudes entre ellos y puede ser útil clasificarlos para detectar modelos representativos del gran número de series. Esto hace relevante investigar medidas de disimilaridad entre modelos ARIMA.

La metodología propuesta por Piccolo requiere en primer lugar ajustar un modelo tipo ARIMA para cada una de las series y posteriormente calcular los valores de π_j , $j=1,2,\dots$. Una vez modelada cada serie se realiza la agrupación en función de esos modelos.

En un estudio reciente (2005) realizado por Jorge Caiado, Nuno Crato y Daniel Peña, proponen una nueva medida de distancia entre series basada en el periodograma normalizado, mostrando resultados de simulaciones en las que se compara esta medida con otras medidas paramétricas y no paramétricas. En particular se discuten la clasificación de series de tiempo como estacionarias o como no estacionarias y consideran el uso de algoritmos de conglomeración jerárquicos y no jerárquicos.

2.2 Métrica para la clasificación de series de tiempo basada en un periodograma.

En esta sección describimos una métrica útil para la clasificación de series de tiempo basada en un periodograma¹.

El problema de identificar similitudes o disimilitudes en series de tiempo ha sido estudiado en la literatura. Algunos estudios utilizan aplicaciones paramétricas y no paramétricas para separar conjuntos de series en conglomerados, observando su distancia Euclídeana. La métrica tiene la limitación importante de ser invariante a transformaciones que modifican el orden de las observaciones en el tiempo y por lo tanto, no toma en cuenta la estructura de correlación de las series de tiempo². Otros estudios introducen una métrica para modelos ARIMA basada en la representación autorregresiva³. Otros más, proponen modelos paramétricos para discriminar y agrupar series multivariadas, entre muchos otros⁴.

La métrica que nos ocupa está basada en el periodograma normalizado para la clasificación de series de tiempo, comparando los resultados de esta métrica con aquellas basadas en los coeficientes de autocorrelación, autocorrelación parcial y autocorrelación inversa. En particular, se discute la clasificación de series de tiempo como estacionarias o como no estacionarias.

Métricas de series de tiempo

El problema de clasificación de series de tiempo inicia con la selección de una métrica relevante.

Sea $X_t = (x_{1t}, \dots, x_{kt})'$ un vector de series de tiempo con componentes representadas por modelos integrados autorregresivos de promedios móviles ARIMA(p,d,q).

$$\phi_i(B)(1-B)^d x_{it} = \theta_i(B)\varepsilon_{it} \quad i = 1, \dots, k,$$

Donde $\phi_i(B)$ es el operador autorregresivo de orden p y $\theta_i(B)$ es el operador de promedios móviles de orden q ; B es el operador de retraso y $(1-B)^d$ es el operador de diferencia de orden d . Las polinomiales autorregresivas y de promedios móviles en el modelo se asume que

¹ Caiado *et al* (2005)

² Galeano y Peña (2002)

³ Piccolo (1990)

⁴ Kakisawa *et al* (1998)

tienen todas sus raíces fuera del círculo unitario, tal que cada proceso $X_{it} = (1-B)^d X_{it}$ es causal e invertible.

Una métrica para la clase de modelos invertibles ARIMA como la distancia Euclideana entre sus expansiones autorregresivas⁵. Sea x_i un proceso estocástico con media cero que sigue un modelo invertible $ARIMA(p,0,q)$: $\phi(B)(1-B)^d x_t = \theta(B)\varepsilon_t$. Entonces puede ser representado por el operador $AR(\infty)$, $\pi(B) = \theta^{-1}(B)\phi(B) = 1 - \pi_1 B - \pi_2 B^2 - \dots$, y los coeficientes π contienen toda la información acerca de la estructura de dependencia estocástica de una serie de tiempo. Piccolo (1990) introduce una métrica comparando las respectivas secuencias π , definidas por la distancia:

$$d_{PIC}(x, y) = \sqrt{\sum_{j=1}^{\infty} (\pi_{jx} - \pi_{jy})^2}$$

Una métrica alternativa basada en la función de autocorrelación estimada (ACF)⁶. Supongamos que tenemos un conjunto de series de tiempo $X = (x_{1t}, \dots, x_{kt})'$ y $\hat{\rho}_i = (\hat{\rho}_{i1}, \dots, \hat{\rho}_{im})$ es el vector de los coeficientes de correlación estimados de las series de tiempo i para alguna m tal que $\hat{\rho}_j \cong 0$ para $j > m$. Esta distancia entre las series de tiempo x y y está definida por:

$$d_{ACF}(x, y) = \sqrt{(\hat{\rho}_x - \hat{\rho}_y)' \Omega (\hat{\rho}_x - \hat{\rho}_y)}$$

Donde Ω es una matriz de ponderaciones. Cuando $\Omega = I$ (matriz de identidad), se obtiene la distancia Euclideana entre los coeficientes de correlación de las series de tiempo x y y . Cuando $\Omega = \text{cov}(\hat{\rho})^{-1}$ es la matriz de covarianza inversa de las autocorrelaciones, se obtiene la distancia de Mahalanobis entre las autocorrelaciones. Es común también utilizar ponderaciones que decrecen con el retraso (lag) de la autocorrelación.

⁵ Piccolo (1990)

⁶ Galeano y Peña (2000)

Otras posibles distancias, que no parecen haber sido aplicadas aún en la literatura de conglomerados, son las basadas en la función de autocorrelación parcial (PACF) y en la función de autocorrelación inversa (IACF) introducida por Cleveland (1972) y desarrollada por Chatfield(1979). Por ahora, una distancia entre las autocorrelaciones inversas puede ser dada por:

$$d_{IACF}(x, y) = \sqrt{(\hat{\rho}_x^{(I)} - \hat{\rho}_y^{(I)})' \Omega (\hat{\rho}_x^{(I)} - \hat{\rho}_y^{(I)})}$$

Donde las funciones de autocorrelación inversa muestrales $\hat{\rho}_x^{(I)}$ y $\hat{\rho}_y^{(I)}$ pueden ser estimadas por métodos presentados por distintos métodos⁷.

Una nueva distancia basada en el periodograma normalizado⁸.

Sean $P_x(w_j) = 1/n \left| \sum_{t=1}^n x_t e^{-itw_j} \right|^2$ y $P_y(w_j) = 1/n \left| \sum_{t=1}^n y_t e^{-itw_j} \right|^2$ los periodogramas de las series de tiempo x y y , respectivamente, en las frecuencias $w_j = 2\pi j/n$, $j = 1, \dots, n/2$ en el rango de 0 a π (donde $n/2$ es el mayor entero menor o igual a $n/2$). Una distancia entre x y y puede ser definida por:

$$d_p(x, y) = \sqrt{\sum_{j=1}^{n/2} [P_x(w_j) - P_y(w_j)]^2}$$

De no existir interés en la escala del proceso, sino únicamente en su estructura de correlación, es mejor utilizar el periodograma normalizado, reemplazando $P(w_j)$ por $NP(w_j) = P(w_j)/\hat{\gamma}_0$, donde $\hat{\gamma}_0$ es la varianza muestral de las series de tiempo, esto es:

$$d_{NP}(x, y) = \sqrt{\sum_{j=1}^{n/2} [NP_x(w_j) - NP_y(w_j)]^2}$$

⁷ Bhansali(1980,1983), Battaglia (1983,1986, 1988), Kanto (1987) y Subba Rao y Gabr (1989)

⁸ Caiado et al (2005)

Dado que la varianza de las ordenadas del periodograma es proporcional al valor del espectro de las frecuencias correspondientes, tiene sentido usar el logaritmo del periodograma normalizado,

$$d_{LNP}(x, y) = \sqrt{\sum_{j=1}^{n/2} [\log NP_x(w_j) - \log NP_y(w_j)]^2}$$

Las distancias NP y LNP satisfacen las propiedades usuales de una métrica:

Simetría $d(x, y) = d(y, x)$

No negatividad $d(x, y) > 0$ con $x \neq y$

Desigualdad del triángulo $d(x, y) \leq d(x, z) + d(z, y)$

Como es de esperarse, medidas basadas en autocorrelaciones y medidas basadas en periodogramas están relacionadas. Es bien sabido que, el periodograma tiene la representación equivalente:

$$P(w_j) = 2 \left[\hat{\gamma}_0 + 2 \sum_{k=1}^{n-1} \hat{\gamma}_k \cos(w_j k) \right]$$

Donde $\hat{\gamma}_k$ es la función de autocovarianza muestral y dividiendo $P(w_j)$ por $\hat{\gamma}_0$ obtenemos el periodograma normalizado dado por

$$NP(w_j) = 2 \left[1 + 2 \sum_{k=1}^{n-1} \hat{\rho}_k \cos(w_j k) \right]$$

Que es la transformación de la función de autocorrelación muestral y viceversa. La relación entre la métrica del periodograma normalizado y la métrica ACF puede ser expresada por:

$$d_{NP}(x, y) = \sqrt{\sum_{j=1}^{n/2} [NP_x(w_j) - NP_y(w_j)]^2}$$

$$= \sqrt{\sum_{j=1}^{n/2} \left[\left(2 + 4 \sum_{k=1}^{n-1} \hat{\rho}_{kx} \cos(w_{jk}) \right) - \left(2 + 4 \sum_{k=1}^{n-1} \hat{\rho}_{ky} \cos(w_{jk}) \right) \right]^2}$$

$$= \sqrt{\sum_{j=1}^{n/2} \left[4 \sum_{k=1}^{n-1} \cos(w_{jk}) (\hat{\rho}_{kx} - \hat{\rho}_{ky}) \right]^2}$$

Por propiedades de ortogonalidad de la función coseno $\sum_{j=1}^{n/2} \cos^2(w_{jk}) = n/4$ y

$\sum_{j=1}^{n/2} \cos(w_j k) \cos(w_j s) = 0$ para $k \neq s$, obtenemos

$$d_{NP}(x, y) = \sqrt{16 \left[\frac{n}{4} \hat{\rho}_{1x} - \hat{\rho}_{1y}^2 + \dots + \frac{n}{4} \hat{\rho}_{n-1,x} - \hat{\rho}_{n-1,y}^2 \right]}$$

$$= 2\sqrt{n} \sqrt{\sum_{k=1}^{n-1} \hat{\rho}_{k,x} - \hat{\rho}_{k,y}^2}$$

O, utilizando notación matricial.

$$d_{NP}(x, y) = 2\sqrt{n} \sqrt{\hat{\rho}_x - \hat{\rho}_y' I \hat{\rho}_x - \hat{\rho}_y}$$

$$= 2\sqrt{n} d_{ACF}(x, y)$$

Estas dos medidas son de este modo equivalentes. Sin embargo su aplicación con un número trunco de autocorrelaciones o un número trunco de ordenadas del periodograma pueden traer diferentes resultados.

Metodología para la clasificación de series de tiempo.

Como primer paso de deben encontrar las similitudes o disimilaridades entre cada par de series de tiempo en el conjunto de datos. Para cada dato calculamos una matriz de distancias con $k(k-1)/2$ pares diferentes utilizando las siguientes métricas:

La distancia Euclideana, $d_{EUCL}(x, y) = \sqrt{\sum_{t=1}^n (x_t - y_t)^2}$

La distancia de Piccolo, $d_{PIC}(x, y) = \sqrt{\sum_{j=1}^{\infty} (\pi_{jx} - \pi_{jy})^2}$

La aplicación de esta distancia requiere ajustar un modelo ARIMA a las series de tiempo.

La distancia ACF, Se implementaron tres posibles formas de calcular una distancia utilizando los coeficientes de autocorrelación. La primera utiliza una ponderación uniforme y es equivalente a la distancia Euclideana entre vectores de coeficientes de autocorrelaciones,

$$d_{ACF}(x, y) = \sqrt{\sum_{i=1}^L (\hat{\rho}_{ix} - \hat{\rho}_{iy})^2}.$$

La segunda utiliza un decaimiento geométrico, $d_{ACFG}(x, y) = \sqrt{\sum_{i=1}^L m_i (\hat{\rho}_{ix} - \hat{\rho}_{iy})^2}$, donde L es el número de autocorrelaciones, $m_i = pq^i$ para $i = 1, \dots, L$, $p = 1 - q$ y $0 < p < 1$.

El tercer uso es la distancia de Mahalanobis, $d_{ACFM}(x, y) = \sqrt{(\hat{\rho}_x - \hat{\rho}_y)' \Omega^{-1} (\hat{\rho}_x - \hat{\rho}_y)}$, donde Ω es la matriz de covarianzas muestral de los coeficientes de autocorrelación por elementos dados por la formula truncada de Barlett. Esto es asumimos que las series de tiempo x y y son,

$$\text{independientes y denotamos } w_{ij} = \sum_{k=1}^L \left[\hat{\rho}_{(k+i)} \hat{\rho}_{(k-i)} - 2\hat{\rho}_{(i)} \hat{\rho}_{(k)} \quad \hat{\rho}_{(k+j)} \hat{\rho}_{(k-j)} - 2\hat{\rho}_{(j)} \hat{\rho}_{(k)} \right],$$

entonces las varianzas y covarianzas en Ω están dadas por

$$\text{var}(\hat{\rho}_x - \hat{\rho}_y) = \text{var}(\hat{\rho}_x) + \text{var}(\hat{\rho}_y) = w_{ix} + w_{iy}$$

$$\text{cov}(\hat{\rho}_{ix} - \hat{\rho}_{iy}, \hat{\rho}_{jx} - \hat{\rho}_{jy}) = \text{cov}(\hat{\rho}_{ix} \hat{\rho}_{jx}) + \text{cov}(\hat{\rho}_{iy} \hat{\rho}_{jy}) = w_{ijx} + w_{ijy}$$

La distancia Euclideana PACF, $d_{PACF}(x, y) = \sqrt{\sum_{i=1}^L \hat{\phi}_{ix} - \hat{\phi}_{iy}^2}$, donde $\hat{\phi}_{ii}$ son las

autocorrelaciones parciales muestrales. También se exploró ponderar estos coeficientes, pero dado que los resultados fueron similares, no se incluyeron.

Como segundo paso agrupamos las series de tiempo en dos conglomerados (estacionarias y no estacionarias) utilizando un algoritmo apropiado aglomerativo jerárquico con un solo acoplamiento (el cuál maximiza la distancia mínima entre objetos del mismo grupo), el acoplamiento completo (el cual minimiza la distancia máxima entre objetos en el mismo grupo) o el acoplamiento promedio (el cual promedia la distancia entre objetos en diferentes grupos). Estos algoritmos de acoplamiento están relacionados con la partición de un conjunto de objetos en grupos o conglomerados, de modo tal que objetos en el mismo grupo son similares uno al otro y objetos en diferentes conglomerados son tan distintos como sea posible. Alternativamente usan

un procedimiento de conglomeración no jerárquico como el de las k medias, en donde se determina un conjunto preliminar de k conglomerados, mueve cada serie de tiempo al conglomerado cuyo centroide está más cercano en distancia euclidiana, recalcula el centroide del cluster y repite el procedimiento de asignación hasta que ninguna serie es reasignada. La implementación del algoritmo de las k medias en esta aplicación se basa en las distancias euclidianas entre observaciones estandarizadas, ponderadores autorregresivos, coeficientes de autocorrelación, coeficientes de autocorrelación parcial, coeficientes de autocorrelación inversa y ordenadas normalizadas del periodograma en escala logarítmica.

Los coeficientes de autocorrelación y el espectro son usualmente definidos por procesos estacionarios pero su definición puede ser extendida para procesos integrados. Con la definición usual de series de tiempo estacionarias (S), $\hat{\rho}_k^S \rightarrow 0$ cuando $k \rightarrow \infty$ y $n \rightarrow \infty$, y para las series de tiempo no estacionarias (NS), para toda k fija, $\hat{\rho}_k^{NS} \rightarrow 1$ cuando $n \rightarrow \infty$. De ahí

$\sum_{k=1}^m \hat{\rho}_k^{NS} - \hat{\rho}_k^S \rightarrow \infty$ cuando $m \rightarrow \infty$ y $n \rightarrow \infty$ con $m < n$. Consecuentemente, se sigue que

$\sum NP^{NS}(w_j) - NP^S(w_j)^2$ también diverge. Así, la ACF estimada y el periodograma normalizado debieran ser capaces de discriminar entre series de tiempo estacionarias y no estacionarias.

2.3 Medida de distancia utilizada.

En el trabajo del M.E. José Ramón Domínguez Molina y la Dra. Graciela González Farías (2001) proponen una medida que no requiere encontrar el mejor modelo ARIMA para modelar las series antes de construir una medida de disimilitud para formar conglomerados de modelos semejantes. Es decir, se pretende primero formar grupos a través de modelar todas las series con un AR de orden grande, ya que como se explico cualquier modelo ARIMA puede ser aproximado por un AR de orden infinito, posteriormente se modela cada grupo con una familia mucho más pequeña de modelos tipo ARIMA. La medida propuesta es la siguiente:

Se tiene

$$Z_t = \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \dots + a_t$$

El proceso puede aproximarse por

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

Para algún h , ya que $\pi_j \rightarrow 0$ cuando $j \rightarrow \infty$.

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

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

Donde σ^2 es la varianza del proceso de ruido blanco, a_t , $X_t = (Z_{t-1}, \dots, Z_{t-h})$, $\pi = (\pi_1, \dots, \pi_h)$.

El estimador de máxima verosimilitud condicional de π , es:

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

Donde

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

El cual coincide con el estimador de mínimos cuadrados.

Haciendo esto se obtendrá una aproximación de los coeficientes de π . Por lo que una métrica sobre ℓ puede definirse ahora por la distancia

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

Otro método de estimación de los coeficientes del autorregresivo es a través de las ecuaciones de Yule-Walker que se describen a continuación:

Como caso general de un modelo autorregresivo, se considera el proceso $AR_{(p)}$ que se describe mediante la ecuación:

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \phi_2 \tilde{Z}_{t-2} + \dots + \phi_p \tilde{Z}_{t-p} + a_t$$

En donde $\tilde{Z}_t = Z_t - \mu$.

Un proceso $AR_{(p)}$ será estacionario si y sólo si las raíces de la ecuación característica

$$1 - \phi_1 x - \phi_2 x^2 - \dots - \phi_p x^p = 0$$

se encuentran fuera del círculo unitario.

Una manera de verificar el supuesto de estacionariedad es mediante las autocorrelaciones $\rho_0, \rho_1, \rho_2, \dots$, definidas por

$$\rho_k = \frac{E[(Z_t - \mu)(Z_{t+k} - \mu)]}{E(Z_t - \mu)^2}$$

para eso se usan las ecuaciones de Yule-Walker, correspondientes al proceso $AR_{(p)}$, esto es:

$$\rho_1 = \phi_1 + \phi_2 \rho_1 + \dots + \phi_p \rho_{p-1}$$

$$\rho_2 = \phi_1 \rho_1 + \phi_2 + \dots + \phi_p \rho_{p-2}$$

:

$$\rho_p = \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + \dots + \phi_p$$

De donde se obtienen los valores de las primeras p autocorrelaciones en función de los parámetros autorregresivos $\phi_1, \phi_2, \dots, \phi_p$; las demás autocorrelaciones se obtienen de la relación:

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \dots + \phi_p \rho_{k-p} \quad , \quad k \geq p+1$$

Lo anterior muestra que los coeficientes del modelo autorregresivo contienen información de la estructura de correlación de las series.

Ejemplo de formación de grupos con simulación de series de tiempo.

En el problema multivariado tradicional se tiene:

$$\begin{pmatrix} peso & altura & \dots & var\ m \\ indiv1 & \dots & \dots & \dots \\ \vdots & \dots & \dots & \dots \\ indivn & \dots & \dots & \dots \end{pmatrix}$$

Existe independencia entre individuos (independencia entre los renglones), pero cuando se trabaja con series de tiempo:

$$\begin{pmatrix} ene73 & ropa & \dots & muebles \\ feb73 & \dots & \dots & \dots \\ \vdots & \dots & \dots & \dots \\ dic07 & \dots & \dots & \dots \end{pmatrix}$$

Sí se utilizara la técnica tradicional se ignoraría la dependencia entre las observaciones (en este caso mensuales) y se obliga a comparar series de igual tamaño y en los mismos periodos.

La medida de distancia que utilizaremos toma en cuenta la estructura de correlaciones entre las series, por lo tanto es común que los grupos formados no tengan semejanza a simple vista.

Será necesario estacionarizar las series, ya que se trata de precios, es decir, de series con una clara tendencia positiva, además si se trabajara con series no estacionarias se requeriría un orden p muy grande para recoger toda la estructura de correlaciones, mientras que en series estacionarias, se puede obtener modelos que cumplan con el principio de parsimonia. No se estima ϕ_0 para la serie de datos relativos, porque ésta ya está estandarizada.

Para aplicar esta técnica no es necesario que las series tengan igual longitud, ya que se estimará un modelo $AR_{(p)}$, lo importante es que contemos con suficientes datos para tener una buena

estimación del modelo. Tampoco será necesario que las series inicien o terminen en el mismo tiempo.

Se trabajó con un orden $p=24$ (para tomar en cuenta las dependencias de hasta dos años) y el algoritmo empleado para estimar los coeficientes del modelo $AR_{(p)}$ fue el de Yule-Walker, con estos coeficientes se obtiene la matriz de distancias y posteriormente se aplican tres métodos de agrupamiento (KMEANS, PAM, AGNES).

Si formáramos conglomerados con el índice, la agrupación sería deficiente, ya que se trata de series de tiempo y no resulta adecuado comparar valor con valor. Lo que debemos comparar es la estructura de dependencias en el tiempo, en nuestro caso los coeficientes de un modelo autorregresivo, calculados a partir del índice relativo al mes anterior.

La medida de similitud que utilizaremos, será la distancia euclídeana entre los coeficientes del modelo autorregresivo.

Ilustramos lo anterior, con un ejemplo:

Simulamos 1000 observaciones de quince series de tiempo, cinco autorregresivos, cinco de media móvil y cinco autorregresivos de media móvil, posteriormente se conforman grupos utilizando AGNES y se muestra el dendograma.

Simulación para validar la técnica de agrupación

```
nr <- 1000
```

```
nc <- 15
```

```
MAT <- matrix(0,nr,nc)
```

```
set.seed(2000)
```

```
for(i in 1:5)
```

```
  MAT[,i] <- arima.sim(n=nr,model=list(ar=0.75))
```

```
for(i in 6:10)
```

```
  MAT[,i] <- arima.sim(n=nr,model=list(ma=0.75))
```

```
for(i in 11:15)
```

```
  MAT[,i] <- arima.sim(n=nr,model=list(ar=0.5,ma=0.75))
```

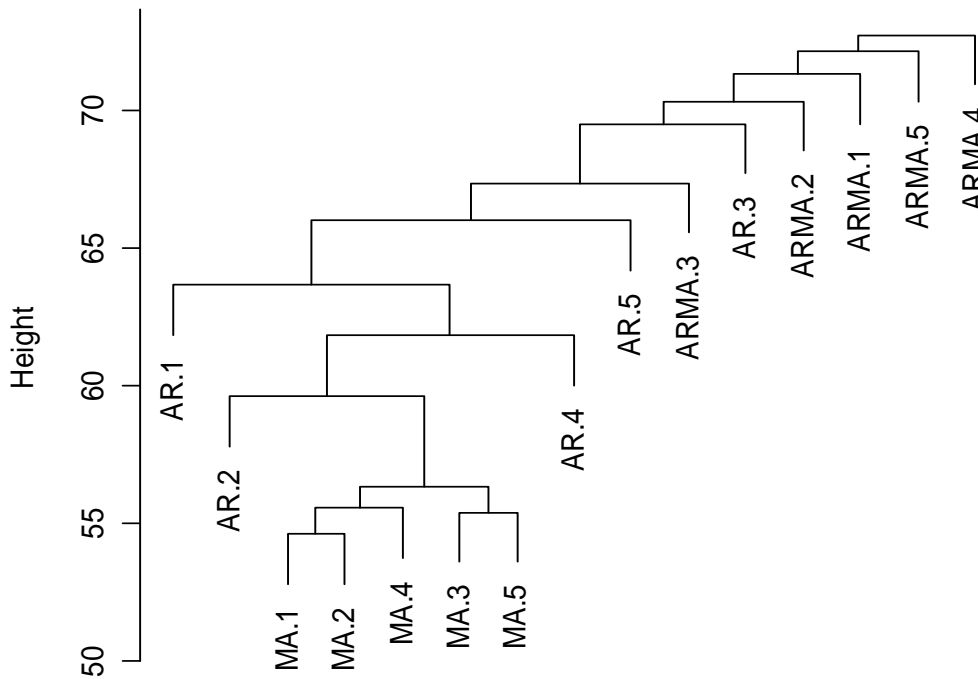
```
dimnames(MAT)[[2]] <- paste(rep(c("AR","MA","ARMA"),each=5),1:5,sep=".")
```

```
library(cluster)
```

```
hc <- agnes(t(MAT))
```

```
plot(hc)
```

Dendrogram of agnes(x = t(MAT))



t(MAT)
Agglomerative Coefficient = 0.13

Puede observarse que las series obtenidas a partir de diferentes modelos no fueron agrupadas de manera adecuada.

Vamos ahora a aplicar la metodología propuesta por Caiado *et al* (2005) referente al uso de una métrica para la clasificación de series de tiempo basada en el periodograma. Para esta simulación, utilizaremos únicamente la técnica de formación de grupos a través de los ρ 's .

La distancia ACF,
$$d_{ACF}(x, y) = \sqrt{\sum_{i=1}^L (\hat{\rho}_{ix} - \hat{\rho}_{iy})^2} .$$

La distancia Euclideana PACF,
$$d_{PACF}(x, y) = \sqrt{\sum_{i=1}^L \hat{\phi}_{ix} - \hat{\phi}_{iy}^2} ,$$
 donde $\hat{\phi}_i$ son las autocorrelaciones parciales muestrales.

Nótese que la distancia PACF es similar a la distancia Domínguez-González (2001)

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

Simulación para validar la técnica de agrupación

```
nr <- 1000
```

```
nc <- 15
```

```
MAT <- matrix(0,nr,nc)
```

```
set.seed(2000)
```

```
for(i in 1:5)
```

```
  MAT[,i] <- arima.sim(n=nr,model=list(ar=0.75))
```

```
for(i in 6:10)
```

```
  MAT[,i] <- arima.sim(n=nr,model=list(ma=0.75))
```

```
for(i in 11:15)
```

```
  MAT[,i] <- arima.sim(n=nr,model=list(ar=0.5,ma=0.75))
```

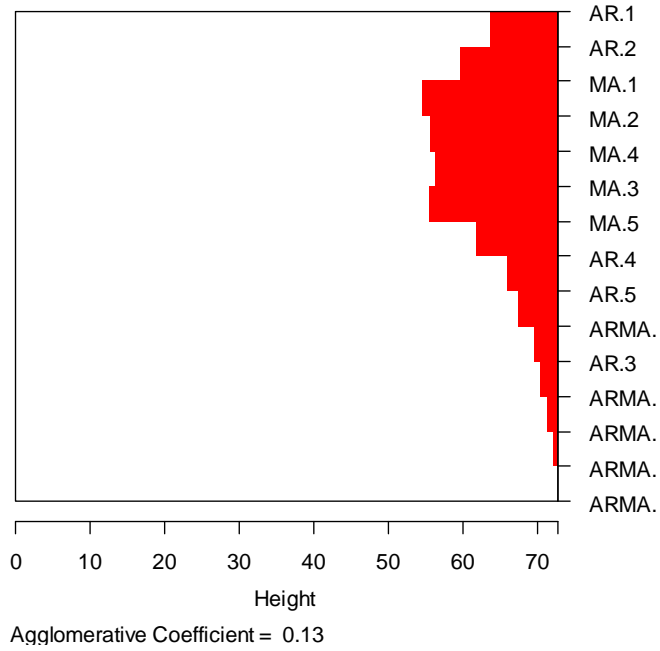
```
dimnames(MAT)[[2]] <- paste(rep(c("AR", "MA", "ARMA"),each=5),1:5,sep=".")
```

```
library(cluster)
```

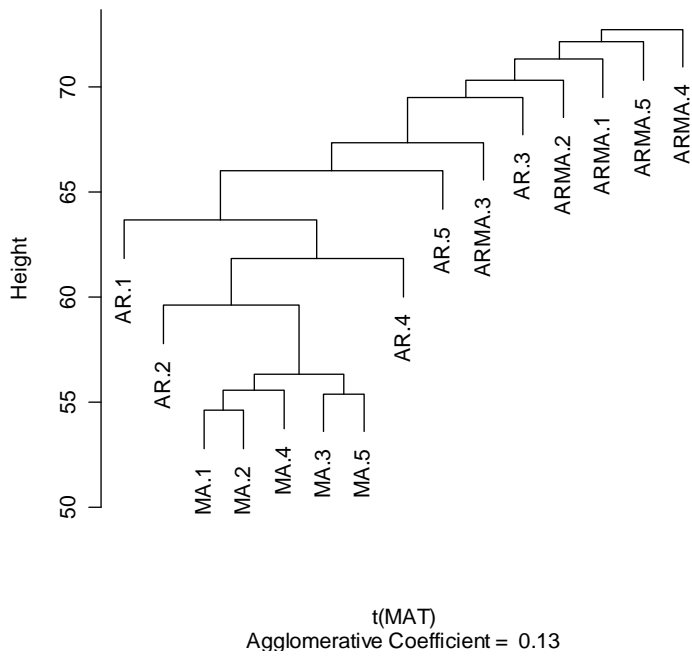
```
hc <- agnes(t(MAT))
```

```
plot(hc,main="Grupos sobre Variables Directas")
```

Grupos sobre Variables Directas



Grupos sobre Variables Directas



nar <- 24

```
MATCOEF <- MATPENACF <- MATPENAPACF <- matrix(0,nar,nc)
for(i in 1:nc)
```

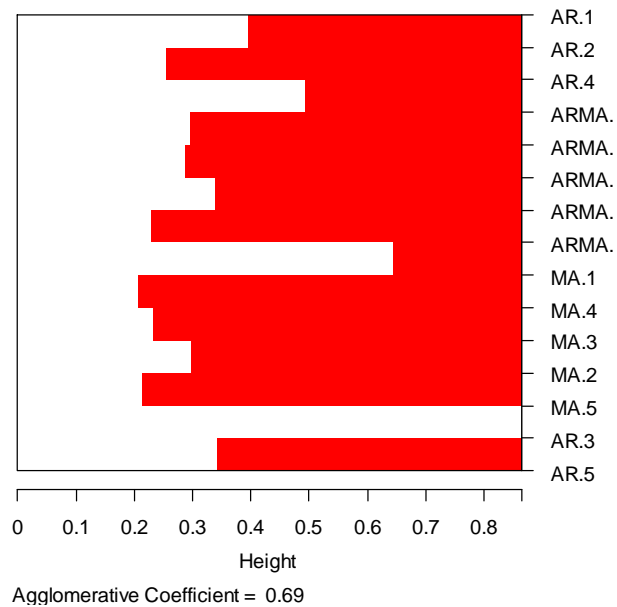
```

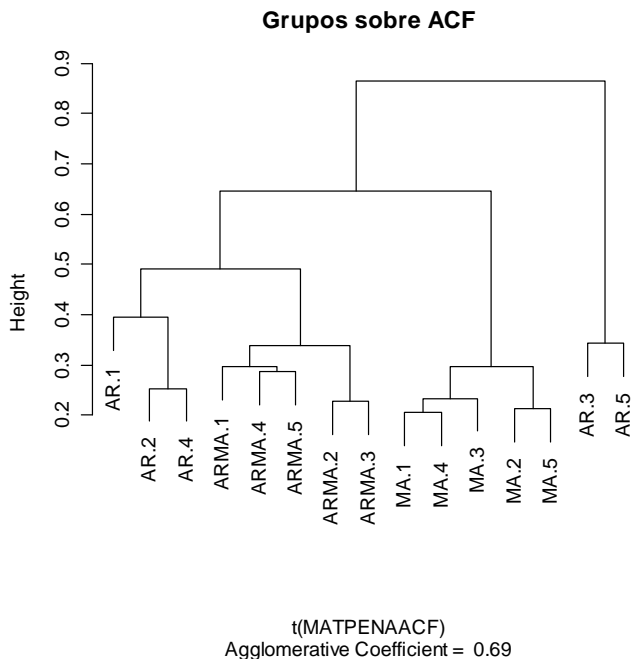
{
  MATPENAACF[,i] <- as.numeric(acf(MAT[,i],lag.max=nar,type="correlation",plot=FALSE)[[1]][-1])
  MATPENAPACF[,i] <- as.numeric(acf(MAT[,i],lag.max=nar,type="partial",plot=FALSE)[[1]])
  MATCOEF[,i] <- ar(MAT[,i],aic=FALSE,order.max=nar,method="yule-walker")$ar
}
dimnames(MATPENAACF)[[2]] <- dimnames(MAT)[[2]]
dimnames(MATPENAPACF)[[2]] <- dimnames(MAT)[[2]]

hc <- agnes(t(MATPENAACF))
plot(hc,main="Grupos sobre ACF")

```

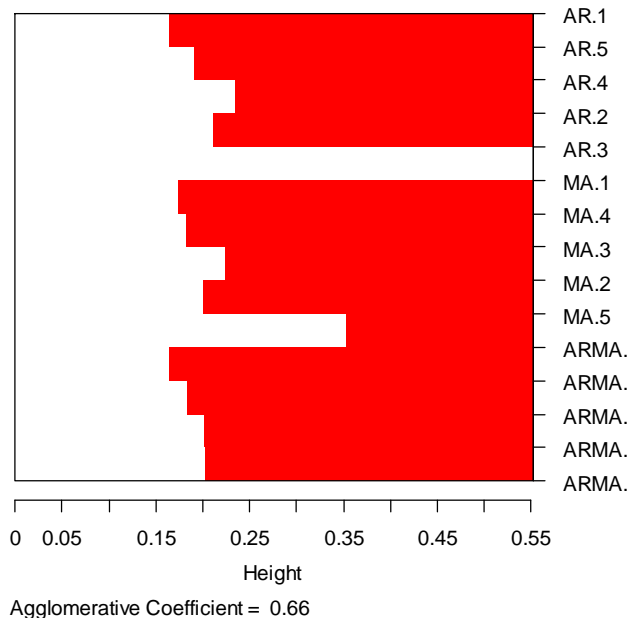
Grupos sobre ACF

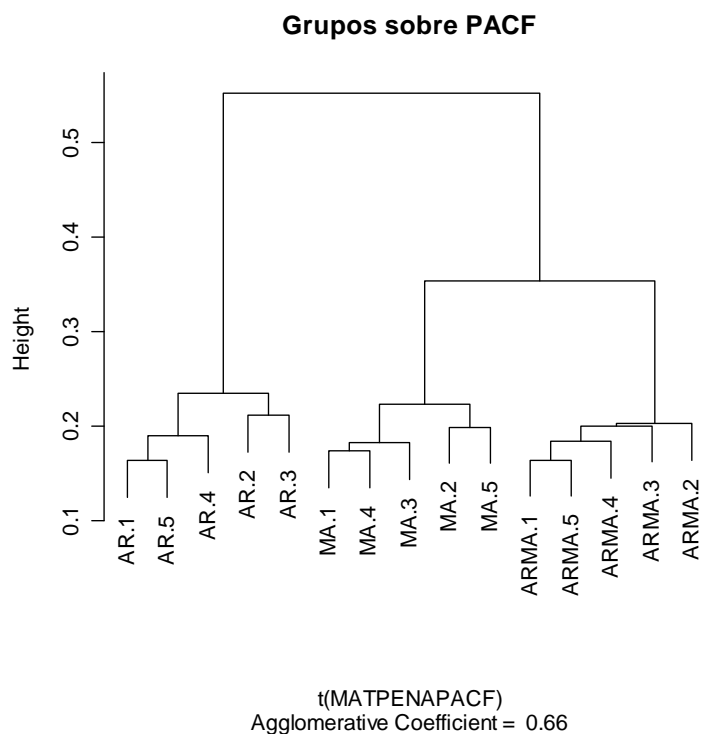




```
hc <- agnes(t(MATPENAPACF))
plot(hc, main="Grupos sobre PACF")
```

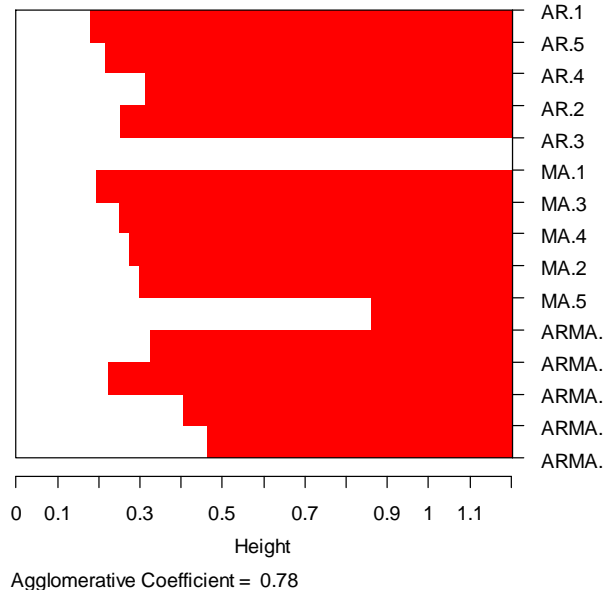
Grupos sobre PACF



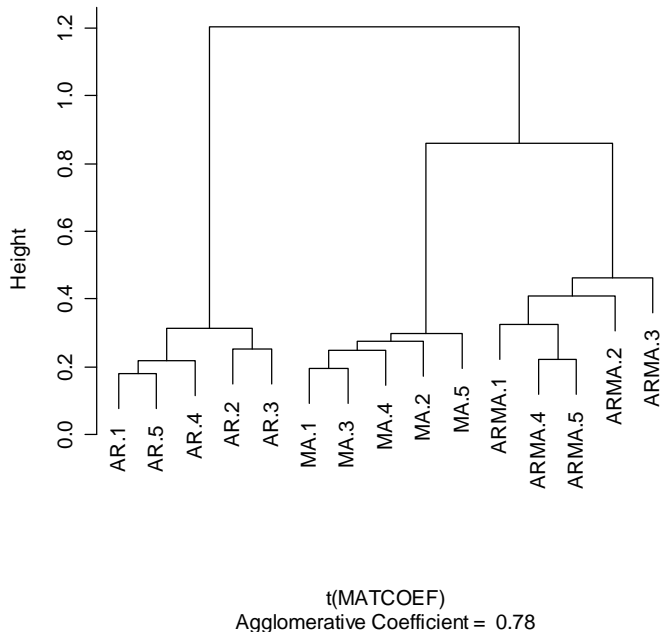


```
dimnames(MATCOEF)[[2]] <- dimnames(MAT)[[2]]
hc <- agnes(t(MATCOEF))
plot(hc,main="Grupos sobre Phi's")
```

Grupos sobre Phi's



Grupos sobre Phi's



Ahora se forman los conglomerados a partir de los coeficientes de un AR_{24} estimados con el método Yule-Walker y se muestra el nuevo dendrograma.

```

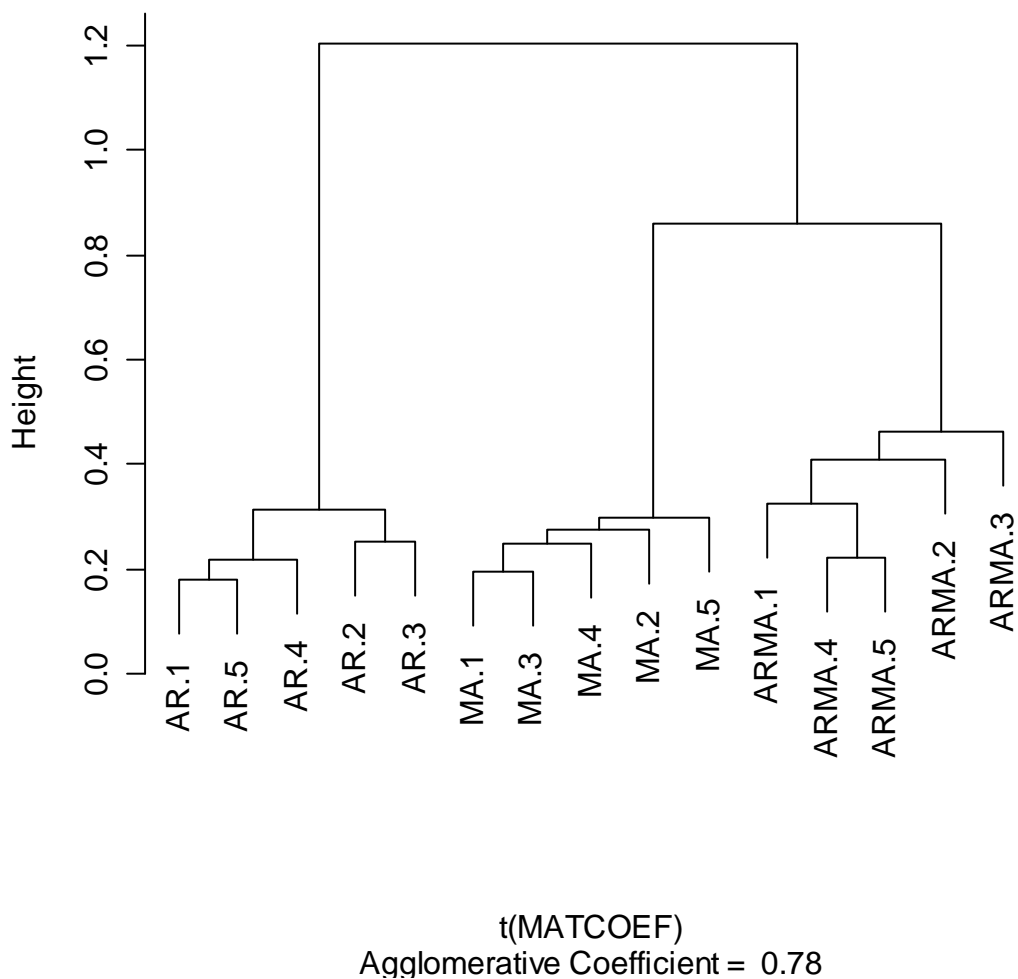
nar <- 24
MATCOEF <- matrix(0,nar,nc)
for(i in 1:nc)
  MATCOEF[,i] <- ar(MAT[,i],aic=FALSE,order.max=nar,method="yule-walker")$ar

dimnames(MATCOEF)[[2]] <- dimnames(MAT)[[2]]

hc <- agnes(t(MATCOEF))
plot(hc)

```

Dendrogram of agnes(x = t(MATCOEF))



El gráfico muestra lo eficiente que resulta cuando se trabaja con series de tiempo, elaborar conglomerados a partir de la estructura de dependencias.

Causalidad y Estacionaridad

Existe confusión entre las nociones de estacionaridad y causalidad para modelos AR, acerca de la certeza de que siempre exista un $AR_{(p)}$.

La condición de causalidad se refiere a que un proceso útil AR deberá depender únicamente de su historia y no de valores futuros.

Para el caso $AR_{(1)}$ la expresión causal es $Y_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$.

Un proceso se dice que es causal si puede ser expresado como valores presentes y pasados de procesos de ruido $Z_t, Z_{t-1}, Z_{t-2}, \dots$.

Un proceso $AR_{(p)}$ será causal si las raíces del polinomio característico $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$ caen todas fuera del círculo unitario

Sea Y_t un modelo causal estacionario $AR_{(p)}$. Multiplicamos Y_t por Y_{t-k} :

$$Y_t Y_{t-k} = \phi_1 Y_{t-1} Y_{t-k} + \dots + \phi_p Y_{t-p} Y_{t-k} + Z_t Y_{t-k}$$

Tomamos la esperanza y obtenemos.

$$\phi_1 \gamma(k-1)$$

Dividiendo entre $\gamma(0)$, obtenemos

$$\rho(k) = \phi_1 \rho(k-1) + \dots + \phi_p \rho(k-p) \text{ para toda } k.$$

Hemos llegado a un conjunto de ecuaciones de diferencia, las ecuaciones de Yule-Walker, cuyas soluciones generales están dadas por:

$$\rho(k) = A_1 \pi_1^{-|k|} + \dots + A_p \pi_p^{-|k|},$$

Donde π_i son las soluciones de la correspondiente ecuación característica de un proceso $AR_{(p)}$,

$$1 - \phi_1 z^{p-1} - \dots - \phi_p z^p = 0.$$

Al final se incluyen en la sección de anexos unas notas sobre el tema de causalidad extraídas del libro de series de tiempo de Chan.

Error de pronóstico

Uno de los objetivos más importantes en el análisis de series de tiempo es el pronóstico de valores futuros.

El pronóstico es esencial para la planeación y control de operación. Se ha desarrollado el error cuadrático medio mínimo para el pronóstico de series estacionarias y no estacionarias. Estos modelos pueden ser utilizados para actualizar pronósticos cuando se dispone de nueva información.

Existen diferentes implicaciones del modelo de series de tiempo construido en términos de su eventual función de pronóstico.

El mínimo error cuadrático medio de pronóstico de Z_{n+l} está dado por la esperanza condicional. Esto es:

$$\hat{Z}_n(l) = E(Z_{n+l} | Z_n, Z_{n-1}, \dots)$$

El error de pronóstico $e_n(l)$ es una combinación lineal de valores futuros aleatorios entrando al sistema después del tiempo n . Especialmente, el error de pronóstico en el tiempo 1 es:

$$e_n(1) = Z_{n+1} - \hat{Z}_n(1) = a_{n+1}$$

Entonces los errores de pronóstico de un tiempo son independientes. Esto implica que $\hat{Z}_n(1)$ es en verdad el mejor pronóstico de Z_{n+1} .

Al final se incluye en la sección de anexos unas notas sobre el tema de error de pronóstico extraídas del libro de series de tiempo de Wei.

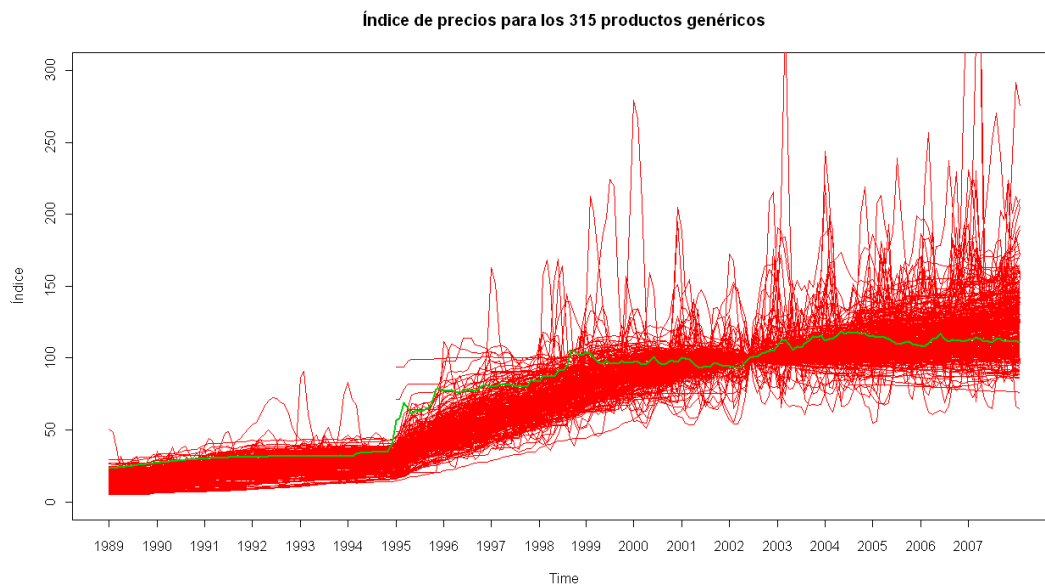
3 Caso de Estudio

Una serie de tiempo es el resultado de observar valores de una variable a lo largo del tiempo en intervalos regulares (cada día, cada mes, cada año, etc.). En nuestro estudio vamos a trabajar con los valores mensuales del Índice Nacional de Precios al Consumidor, originalmente obtuvimos información de las series históricas de precios desde enero de 1973 hasta febrero del 2008, esto hacía un total de 455 series, cada una con 421 observaciones, se pudo haber trabajado también con las observaciones quincenales con lo que se tendrían series con el doble de datos, sin embargo, por la naturaleza de este estudio decidimos trabajar con los datos mensuales. Adicionalmente se decidió trabajar únicamente con 20 años de información y con las series correspondientes a los 315 productos genéricos ya que las series correspondientes a grupos y subgrupos son agregaciones de información.

Se incorporó la serie con información del tipo de cambio, porque es conocida la estrecha relación que tiene esta variable con los precios y de esta forma podremos observar las relaciones entre ciertos clusters y el tipo de cambio, particularmente, podremos observar que grupos tienen un comportamiento más parecido al tipo de cambio.

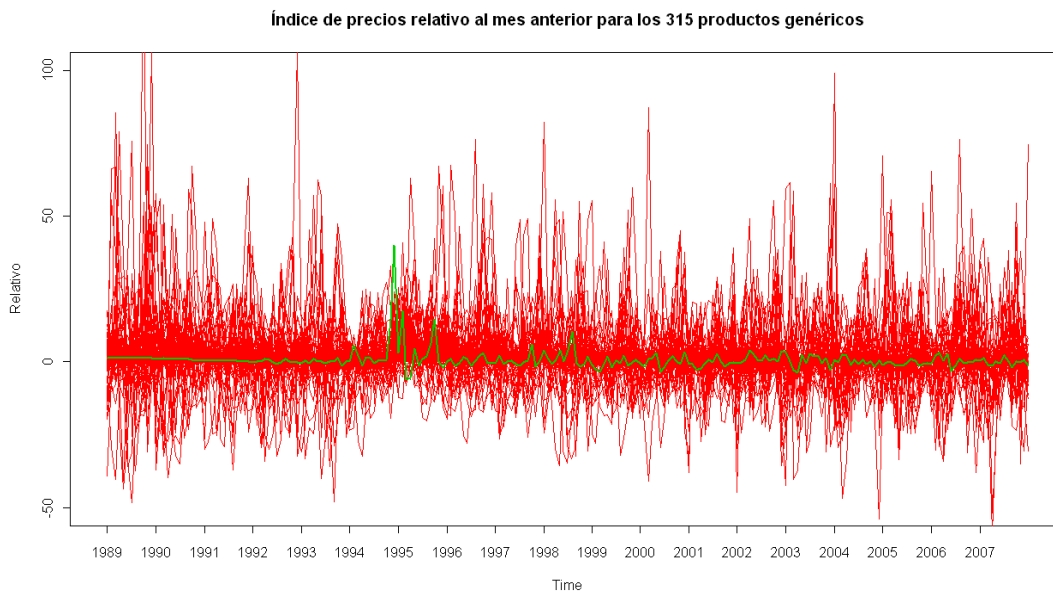
Las series de tiempo que se mantienen alrededor de un valor constante, se dice que son estacionarias, también es posible encontrar series de tiempo estacionarias alrededor de una tendencia.

Las series de tiempo del INPC no se mantienen alrededor de un valor constante, en éste caso se trata de series con tendencia creciente, como puede apreciarse en el gráfico siguiente:



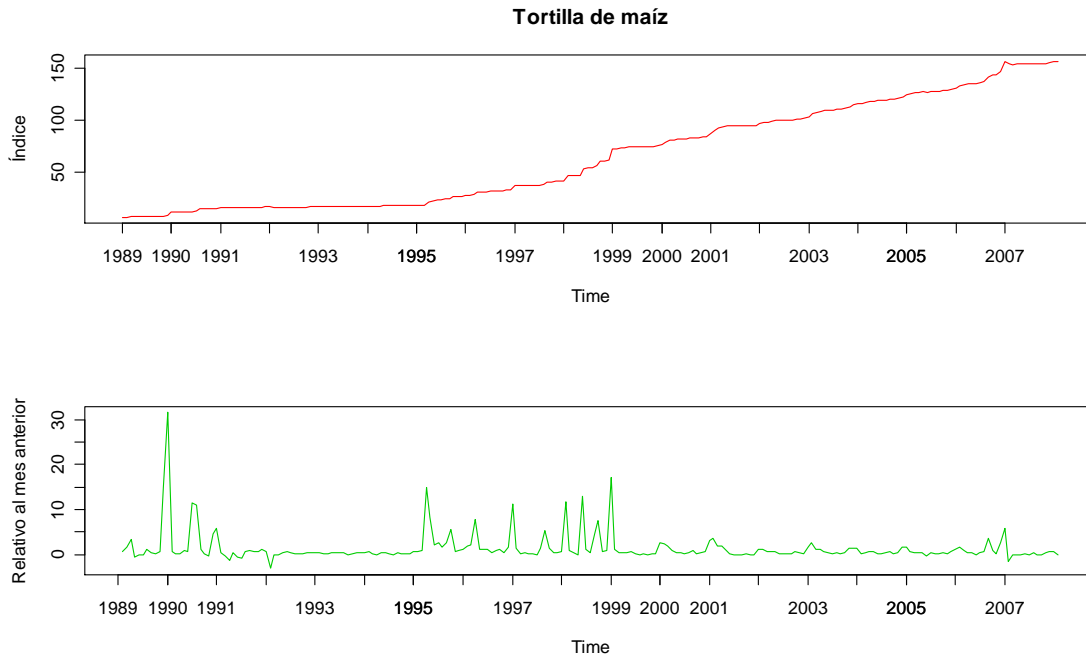
La línea verde corresponde a la serie del tipo de cambio.

La manera natural de eliminar la tendencia es trabajar con el índice relativo al mes anterior, lo cual tiene a medir naturalmente el efecto inflacionario mensual, aunado a que se alivian posibles no estacionalidades al tenerlas a ese nivel.



Una extensión natural de este análisis sería trabajar con series relativas al mismo mes del año anterior. Este tipo de series son de uso común, cuando se habla de precios.

El gráfico de una serie de tiempo es una herramienta muy valiosa para entender su comportamiento, sin embargo en los gráficos previos mostramos 315 series, en las cuales es posible visualizar que tienen un comportamiento general, pero resulta imposible visualizar productos específicos, por ello, a manera de ejemplo mostramos el gráfico de un solo producto, en este caso la tortilla de maíz, para el caso del índice de precios y del índice de precios relativo al mes anterior.



En el gráfico correspondiente al Índice se observa que la serie no es estable, ya que su nivel aumenta con el tiempo, en este caso se observa una clara tendencia positiva. La mayoría de las series económicas no son estacionarias y presentan tendencias constantes en el tiempo. En este caso se observa una tendencia aproximadamente lineal, aunque se antoja ajustar tres rectas, la primera entre 1989 y 1995, la segunda de 1995 a 1999 y la última a partir de 1999 por sus diferencias en pendiente.

Se trata de un producto con precios controlados, por lo menos en ciertos periodos, sin embargo, en el gráfico de los precios relativos al mes anterior, es posible observar los picos que representan las alzas en el precio de este producto.

Encontramos que no todas las series están completas, es decir no tienen datos para el período completo y esto es debido a que como ya se explicó, van saliendo productos y se van incorporando nuevos productos del cálculo del índice de precios debido principalmente a cambios en los hábitos de consumo de los hogares.

Modelaremos un AR de orden 24 para capturar dependencias de hasta dos años en las series de precios. Esto no afecta para las series que están incompletas. Con el fin de capturar el comportamiento actual de los incrementos en los precios trabajamos con datos a partir de 1989.

Finalmente, calcularemos la distancia Euclídeana del centro de cada grupo a la serie del tipo de cambio del peso frente al dólar, para observar cuál de los grupos es más sensible a esta variable.

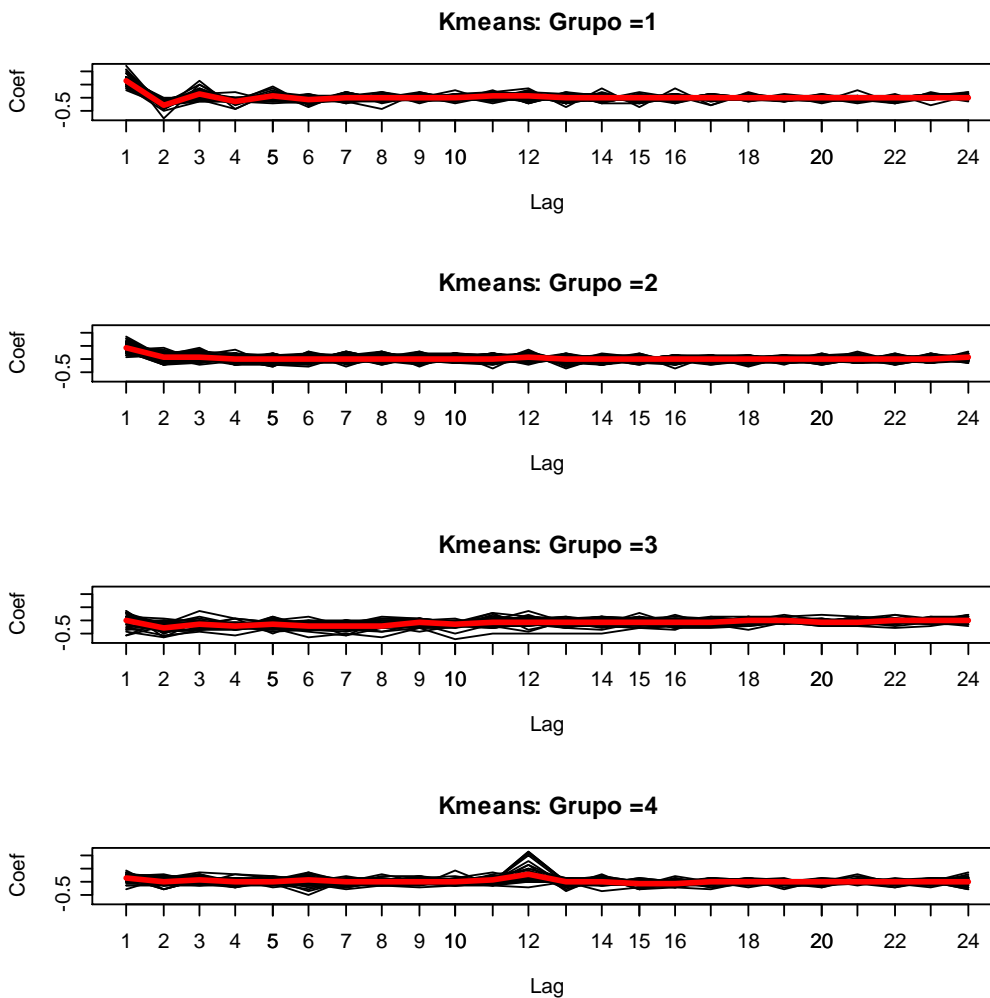
Es importante recordar que se está trabajando con series de incrementos en los precios respecto al mes inmediato anterior.

3.1 Resultados obtenidos por el método de K medias.

Gráfica de los perfiles de los coeficientes para cada uno de los grupos. Método para la elaboración de conglomerados: K medias

Con éste tipo de gráfico observaremos las características generales de correlaciones de las series en cada grupo.

Los conglomerados se forman a partir de los coeficientes, aquí graficamos los coeficientes del $AR_{(24)}$, de cada serie de incrementos de precios para cada grupo. Los conglomerados se formaron con el método de la K medias.



El grupo 1 muestra una dependencia mayor del mes inmediato anterior, mostrando también algunas dependencias de los meses 1 a 5.

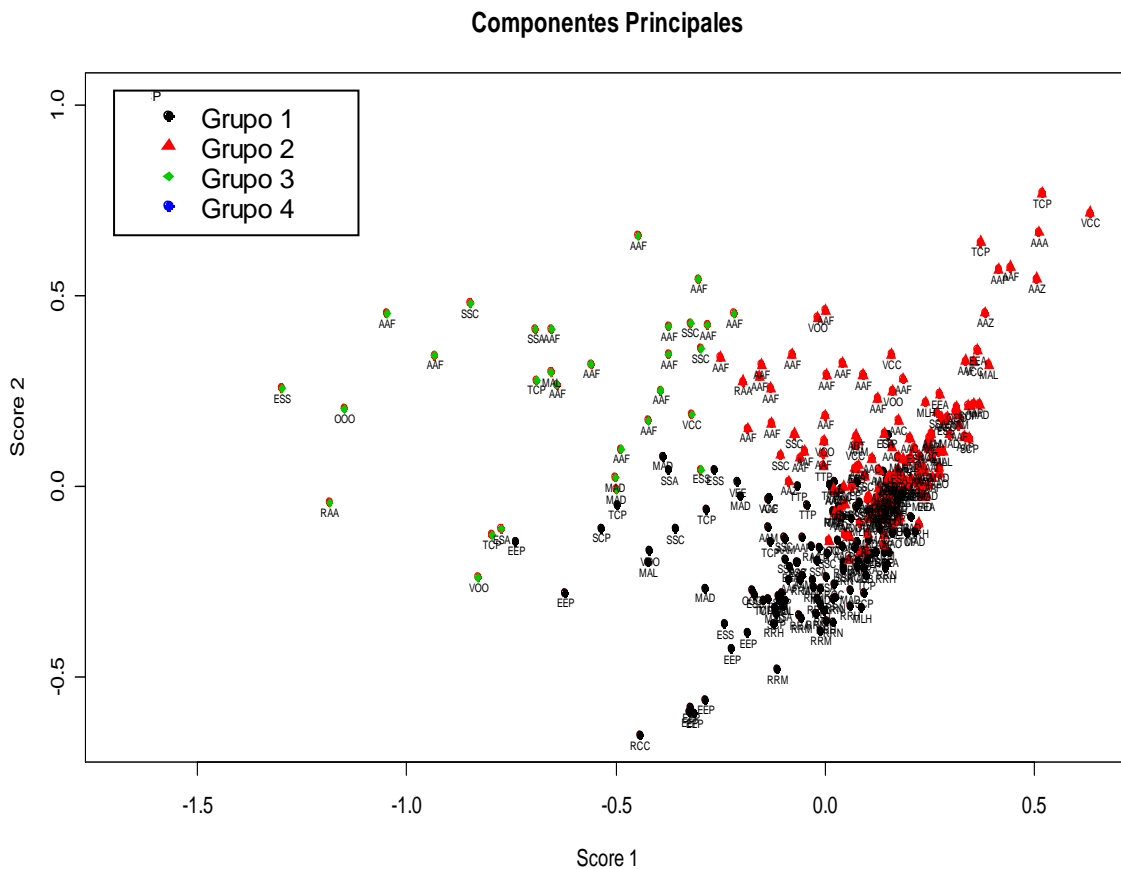
El grupo 2 muestra dependencias casi exclusivamente con el mes inmediato anterior.

El grupo 3 no depende tanto del mes inmediato anterior sino del segundo y tercero.

En el grupo 4 se observa que se tienen series de incrementos de precios que dependen más del mismo mes del año anterior que de su historia más reciente.

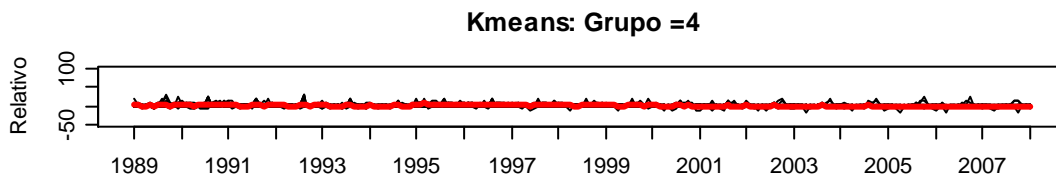
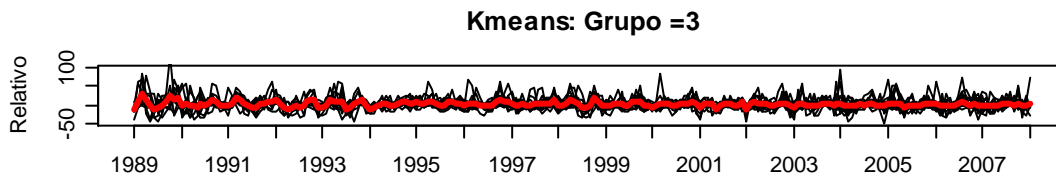
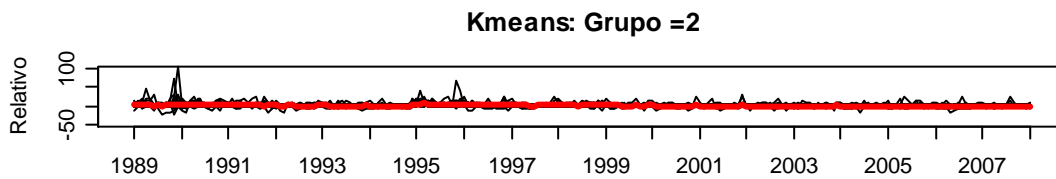
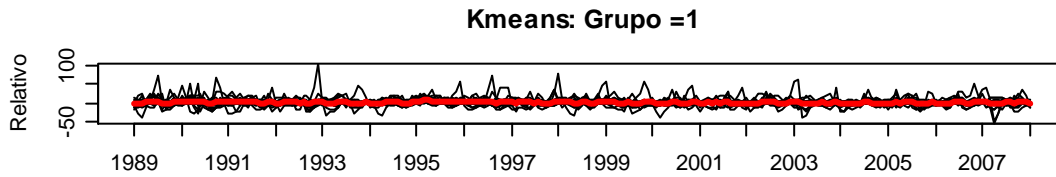
Gráfica de componentes principales Método para la elaboración de conglomerados: K Medias

Se utiliza el método de componentes principales con fines de visualización, sólo como herramienta de reducción de dimensiones de los coeficientes de 24 a 2 dimensiones. Si los grupos se iluminan de manera separada en la nube de puntos esto es una indicación de grupos agradable. El método utilizado en este caso para la formación de los conglomerados es el de las k medias.



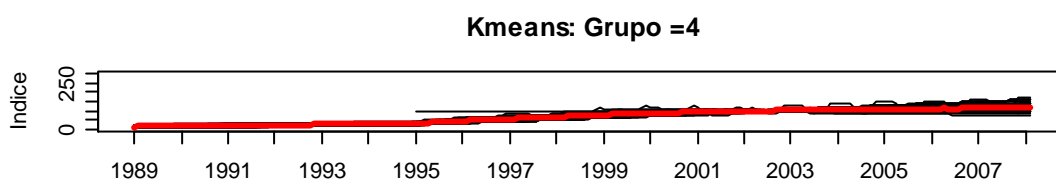
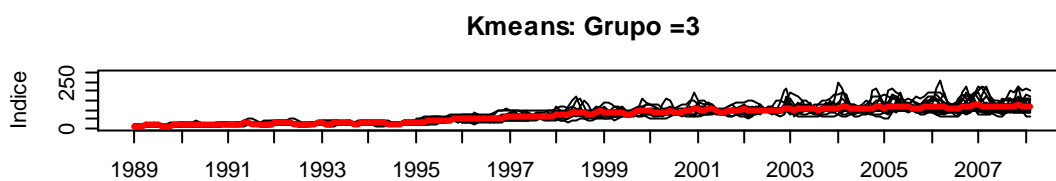
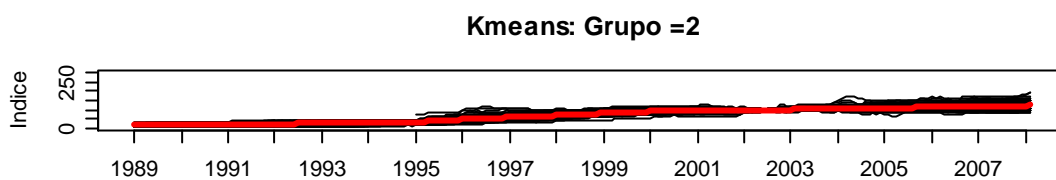
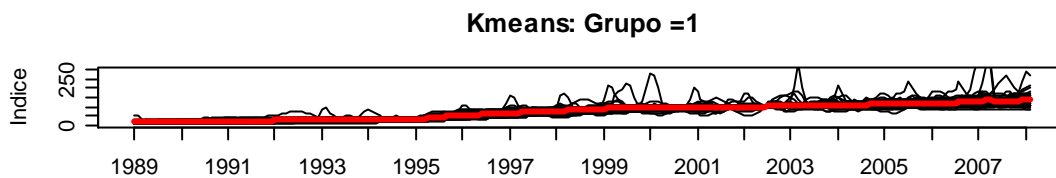
Gráfica de los perfiles de las series relativas para cada uno de los grupos Método para la elaboración de conglomerados: K Medias.

Quando se trabaja con el índice respecto al mes anterior, se elimina la tendencia, en este caso se puede observar la volatilidad de las series.



Gráfica de los perfiles de las series en índice. Método para la elaboración de conglomerados: K Medias.

A pesar de que esta es la información base de todo el análisis se tiene que aquí se observa sólo la tendencia de las series.



Composición de los Conglomerados
Método: k-medias

Conglomerado	Descripción del Grupo	Elementos del Grupo en el Conglomerado	Relativo al Conglomerado	Relativo al Grupo	Distancia de Centro del grupo a la serie del tipo de cambio
1	Alimentos	32	66.7	26.2	.55
1	Educación	2	4.2	6.7	
1	Muebles	2	4.2	4.8	
1	Otros	1	2.1	14.3	
1	Ropa	1	2.1	2.6	
1	Salud	1	2.1	2.5	
1	Transporte	3	6.2	14.3	
1	Vivienda	6	12.5	42.9	
2	Alimentos	69	39.9	56.6	.18
2	Educación	12	6.9	40.0	
2	Muebles	29	16.8	69.0	
2	Otros	3	1.7	42.9	
2	Ropa	18	10.4	46.2	
2	Salud	25	14.5	62.5	
2	Transporte	13	7.5	61.9	
2	Vivienda	4	2.3	28.6	
3	Alimentos	13	37.1	10.7	.60
3	Educación	4	11.4	13.3	
3	Muebles	4	11.4	9.5	
3	Otros	1	2.9	14.3	
3	Ropa	1	2.9	2.6	
3	Salud	7	20.0	17.5	
3	Transporte	3	8.6	14.3	
3	Vivienda	2	5.7	14.3	
4	Alimentos	8	13.6	6.6	.22
4	Educación	12	20.3	40.0	
4	Muebles	7	11.9	16.7	
4	Otros	2	3.4	28.6	
4	Ropa	19	32.2	48.7	
4	Salud	7	11.9	17.5	
4	Transporte	2	3.4	9.5	
4	Vivienda	2	3.4	14.3	

El conglomerado uno muestra una mayor dependencia respecto a los meses previos y está conformado principalmente por alimentos, básicamente por alimentos frescos como son la carne y las frutas, algunos ejemplos son el pollo, el lomo, pierna, mojarra, manzana, naranja aguacate, mango, limón, uva, pera, toronja, piña, guayaba, cebolla, zanahoria. Del grupo ropa únicamente aparecen los zapatos de hombre. Algunos elementos del grupo vivienda no parece que debieran haber quedado en este conglomerado como son el costo de la renta o el impuesto predial, sí lo relacionado con el servicio doméstico y los servicios de mantenimiento de la vivienda. Del grupo

transporte aparecen los automóviles, la tenencia de automóvil y el transporte aéreo, finalmente del grupo educación aparecen los libros de texto y los cuadernos y carpetas.

En el grupo 2 quedan el 56.6% de los alimentos, sin embargo se observa que tienen en su mayoría algún grado de procesamiento como es el caso de la tortilla, las galletas, quesos, yogurth, helados, jamón, salchichas, chorizo, azúcar, mayonesa, mostaza, chocolates, dulces, gelatina, los alimentos preparados como las pizzas, carnitas, pollo rostizado, barbacoa o birria, otros licores, ron, brandy y vino de mesa. Aparece casi la mitad de los elementos del grupo ropa, destacando aquellos elaborados a base de algodón. Este grupo contiene también casi al 70% de los elementos del grupo muebles y a poco más del 60% de los elementos de los grupos de transporte y salud, así como al 40% de los elementos del grupo educación. En este grupo se observa dependencia con los meses previos.

El grupo 3 es el más heterogéneo ya que contiene entre el 10 y el 15 por ciento de cada grupo de genéricos, es decir agrupa a los elementos con comportamientos atípicos, que no muestran dependencias claras con los meses anteriores.

Finalmente el conglomerado 4 está compuesto principalmente por ropa y muestra una mayor dependencia con respecto a los precios del mismo mes pero del año anterior. De manera intuitiva, puede ser claro que la ropa respete una temporada, es decir que tenga más relación el precio de la ropa de invierno, con el precio que tuvo en el invierno anterior más que con el precio de la ropa en verano. Además de la ropa, en este grupo quedan algunos bienes clasificados en el grupo de muebles como son los colchones, recámaras, colchas, otros blancos del hogar, algunos aparatos eléctricos. Del grupo salud, se incluyen material de curación, lentes, análisis clínicos, que muestran un comportamiento similar. De manera natural queda comprendidos en este conglomerado elementos del grupo educación que van desde jardín de niños, primaria, secundaria, preparatoria, carreras cortas y universidad, además se incluye servicios funerarios y licencias del grupo otros.

Para ver la información a detalle, será necesario revisar el anexo 4.

En relación al comportamiento respecto al tipo de cambio vemos que en este caso el centro del conglomerado 2 tiene menor distancia respecto a la serie del tipo de cambio, es decir los productos incluidos en este conglomerado muestran mayor sensibilidad a variaciones cambiarias y son los siguientes:

Del grupo alimentos: tortilla de maíz, masa y harina de maíz, maíz, pan de caja, pastelillos y pasteles, pasta para sopa, galletas populares, tortillas de harina y trigo, otras galletas, arroz, pollo en piezas, pulpa de cerdo, bistec de res, carne molida de res, chuletas y costillas de res, retazo, cortes especiales de res, hígado de res, otras vísceras de res, otros cortes de carne, jamón, salchichas, chorizo, otros embutidos, carnes secas, tocino, camarón, otros pescados y mariscos, leche pasteurizada y fresca, leche evaporada, condensada y maternizada, queso fresco, yogurt, crema de leche, queso chihuahua o manchego, otros quesos, queso amarillo, mantequilla, huevo, plátanos, papaya, papa, otras legumbres, legumbres secas, jugos o néctares envasados, chiles procesados, verduras envasadas, puré de tomate y sopas enlatadas, azúcar, agua embotellada, mayonesa y mostaza, concentrados de pollo y sal, otros condimentos, papas fritas y similares, concentrados para refresco, chocolate, dulces, cajetas y miel, gelatina en polvo, otros alimentos cocinados, carnitas, pollos rostizados, barbacoa o birria, pizzas, tequila, otros licores, ron, brandy, vino de mesa.

Del grupo ropa, calzado y accesorios: camisas, ropa interior para hombre, calcetines, pantalón hombre base algodón, otras prendas para hombre, ropa interior para mujer, pantalón mujer base algodón, vestido para niña, camiseta para bebé, sombreros, suéter para niño y niña, zapatos tenis, zapatos para mujer, zapatos para niños, otros gastos del calzado, servicios de tintorería, relojes, joyas y bisutería.

Del grupo vivienda: suministro de agua, gas doméstico, servicio telefónico local, larga distancia nacional.

Del grupo muebles aparatos y accesorios domésticos: muebles para cocina, ante comedores, estufas, calentadores para agua, salas comedores refrigeradores, lavadoras de ropa, ventiladores, planchas eléctricas, licuadoras, computadoras, equipos modulares, radios y grabadoras, focos, cerillos, pilas, velas y veladoras, escobas, loza y cristalería, utensilios de plástico para el hogar, otros utensilios de cocina, sábanas, cobijas, toallas, cortinas, detergentes, desodorantes ambientales, plaguicidas.

Del grupo salud y cuidados personales: antibióticos, analgésicos, cardiovasculares, nutricionales, anticonceptivos y hormonales, gastrointestinales, otros medicamentos, antigripales, consulta médica, operación quirúrgica, cuidado dental, atención médica durante el parto, corte de cabello, sala de belleza, productos para el cabello, lociones y perfumes, jabón de tocador, pasta dental, desodorantes personales, crema para la piel, navajas y maquinas de afeitarse, papel higiénico, pañales, toallas sanitarias, servilletas de papel.

Del grupo transporte: colectivo, autobús urbano, taxi, metro, autobús foráneo, bicicletas, aceites y lubricantes, neumáticos, otras refacciones, acumuladores, seguro de automóvil, mantenimiento de automóvil, estacionamiento.

Del grupo educación y esparcimiento: otros libros, plumas, lápices y otros, cine, centro nocturno, espectáculos deportivos, periódicos, revistas, juguetes, discos y casetes, alimentos para mascotas, material y aparatos fotográficos, artículos deportivos.

Del grupo otros servicios: loncherías, restaurantes y cafeterías.

Vemos que en este conglomerado aparecen productos como el maíz, del cual en la historia contemporánea del país ya no se tiene autosuficiencia de ahí la necesidad de importar de acuerdo a las cuotas establecidas por la Secretaría de Economía para no afectar a productores nacionales. Gracias a la apertura comercial, cada día tenemos acceso a nuevos productos, por ejemplo, existen cadenas de comida rápida en México que importan gran cantidad de insumos como papas fritas y similares, mayonesas, mostazas, así como concentrados para refrescos. Por ejemplo en el año 1995 que se tuvo una alta depreciación de nuestra moneda frente al dólar americano, cadenas comerciales donde más de la mitad de los productos que venden son importados decidieron suspender sus planes de expansión hasta contar con escenarios más favorables.

Aparecen también gran número de productos de consumo final como la ropa, el calzado, medicamentos y artículos de cuidado personal, así como algunos servicios. Es sabido que nada daña más a la economía que tener un déficit comercial propiciado por la importación de productos

de consumo final respecto a la importación de maquinaria, equipo e insumos que son empleados para la producción de productos cuyo destino puede ser el mercado internacional.

Por otro lado el conglomerado cuyo centro muestra mayor distancia con la serie del tipo de cambio es el conglomerado número tres, cuyos productos son los siguientes:

Del grupo alimentos: melón, sandía, jitomate, tomate verde, calabacita, chile serrano, otros chiles frescos, chile poblano, chícharo, nopales, chayote, pepino, ejotes.

Del grupo ropa, calzado y accesorios: solamente aparece el elemento servicios de lavandería.

Del grupo vivienda: materiales para el mantenimiento de la vivienda, la larga distancia internacional.

Del grupo muebles, aparatos y accesorios domésticos: muebles diversos para el hogar, aparatos de aire acondicionado, blanqueadores y televisores.

Del grupo salud: dermatológicos, prótesis dentales, hospitalización general, estudios médicos de gabinete, hospitalización por parto, artículos de maquillaje facial.

Del grupo transporte: gasolina de bajo octanaje, servicios de reparación de automóvil y lavado y engrasado de automóvil.

De grupo educación y esparcimiento: enseñanza adicional, servicios turísticos en paquete, servicios de internet y renta de películas.

Del grupo otros servicios: servicios profesionales.

En el conglomerado cuyo comportamiento se vio menos afectado por variaciones en el tipo de cambio aparecen genéricos cuya producción nacional cubre la demanda del propio país tales como variedades de chiles, frutas y verduras, melón, sandía, jitomate, y calabaza. Existe mayor producción de estos alimentos ya que tienen una demanda garantizada además de que su precio en el mercado es rentable para los productores.

Servicios como lavandería de ropa y el lavado y engrasado de automóviles son muy locales y es natural que no se vean afectados fuertemente por variaciones cambiarias.

Finalmente, resulta interesante observar la suma de los ponderadores de cada grupo:

KMEANS

Conglomerado	Elementos	Suma de Ponderadores
1	48	28.03368
2	173	46.17594
3	35	9.19290
4	59	16.59748
Total	315	100.00000

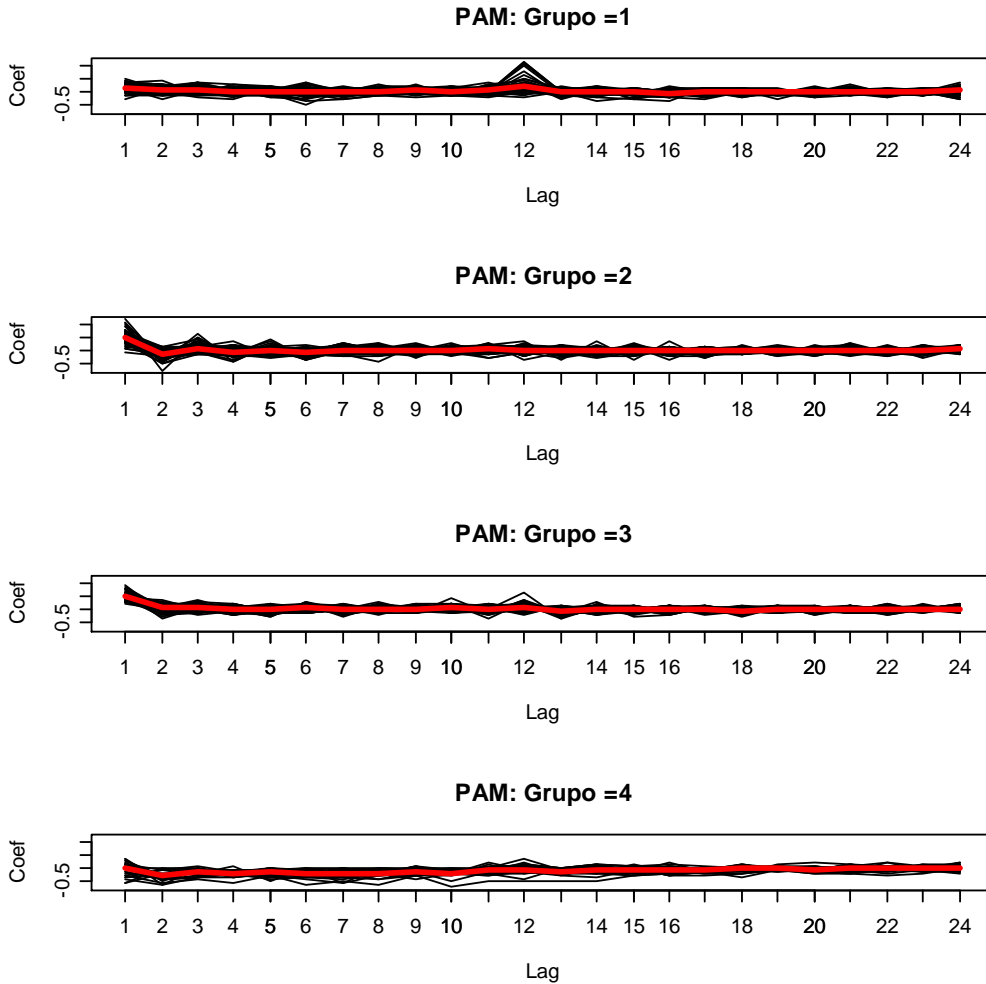
Vemos que el grupo que contiene el mayor número de elementos, de hecho 173 de los 315 productos genéricos representan el 54.92% de los productos y aportan el 46.17% del valor del índice.

3.2 Resultados obtenidos por el método PAM.

Gráfica de los perfiles de los coeficientes para cada uno de los grupos. Método para la elaboración de conglomerados: PAM

Con éste tipo de gráfico observaremos las características generales de correlaciones de las series en cada grupo.

Los conglomerados se forman a partir de los coeficientes, aquí graficamos los coeficientes del $AR_{(24)}$, de cada serie de incrementos de precios para cada grupo. Los conglomerados se formaron con el método PAM.



El grupo 1 se observa que se tienen series de incrementos de precios que dependen más del mismo mes del año anterior que de su historia más reciente.

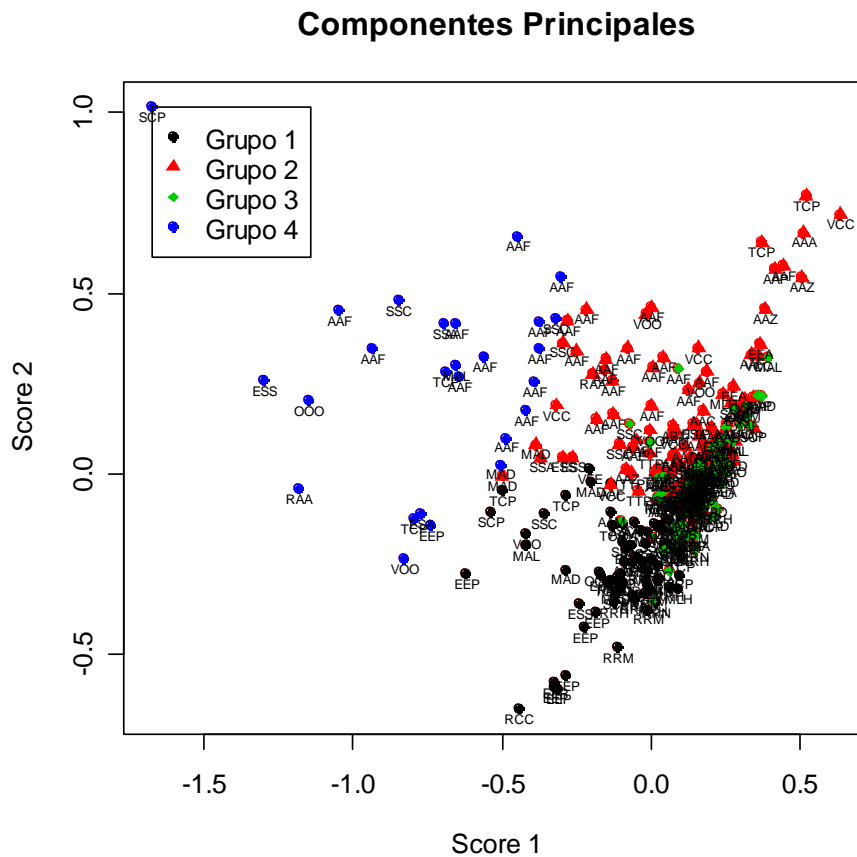
El grupo 2 muestra una dependencia mayor del mes inmediato anterior, mostrando también algunas dependencias de los meses 1 a 4.

El grupo 3 muestra dependencias casi exclusivamente con el mes inmediato anterior.

En el grupo 4 no depende tanto del mes inmediato anterior sino del segundo y tercero. En general no muestra dependencias claras.

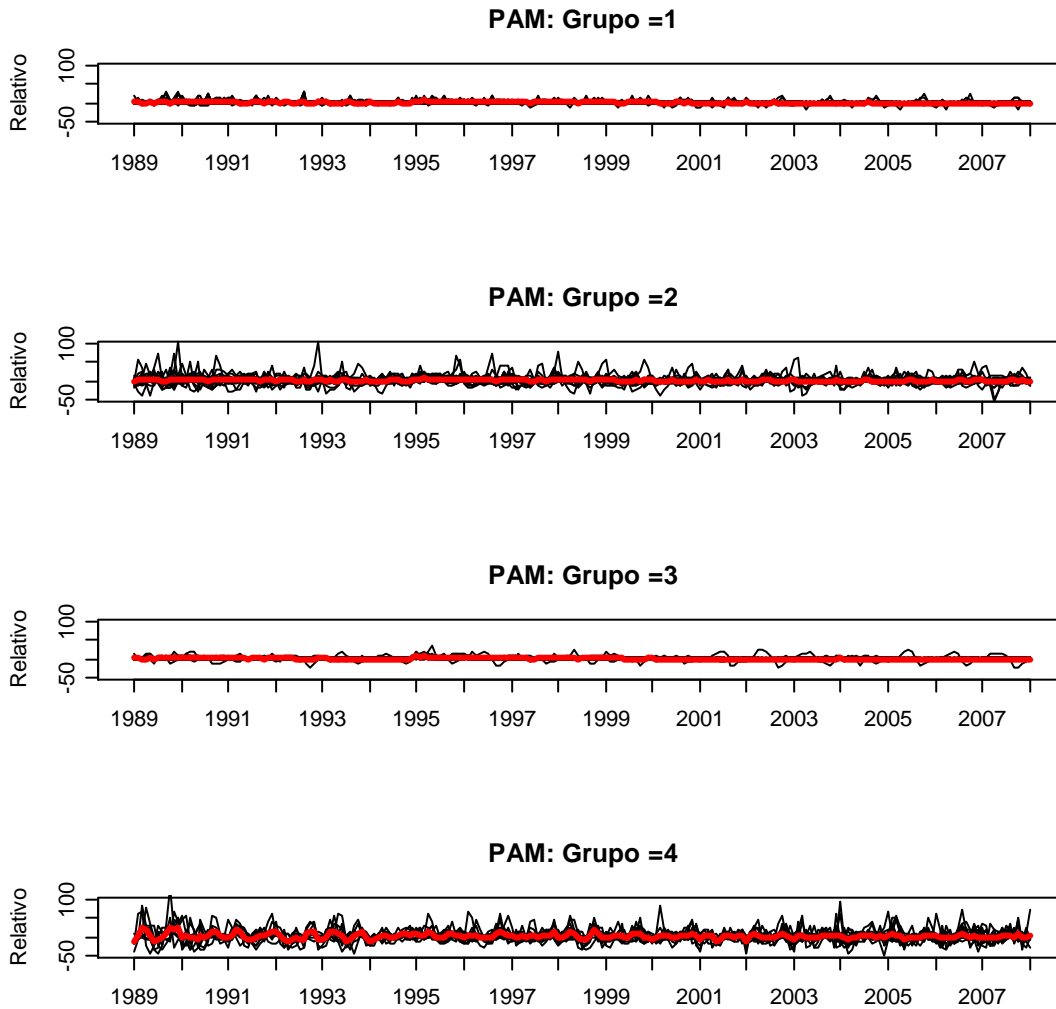
Gráfica de componentes principales Método para la elaboración de conglomerados: PAM

Se utiliza el método de componentes principales con fines de visualización, sólo como herramienta de reducción de dimensiones de los coeficientes de 24 a 2 dimensiones. Si los grupos se iluminan de manera separada en la nube de puntos esto es una indicación de grupos agradable. El método utilizado en este caso para la formación de los conglomerados es PAM.



Gráfica de los perfiles de las series relativas para cada uno de los grupos Método para la elaboración de conglomerados: PAM.

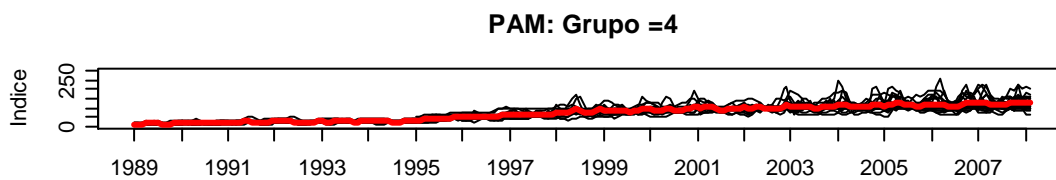
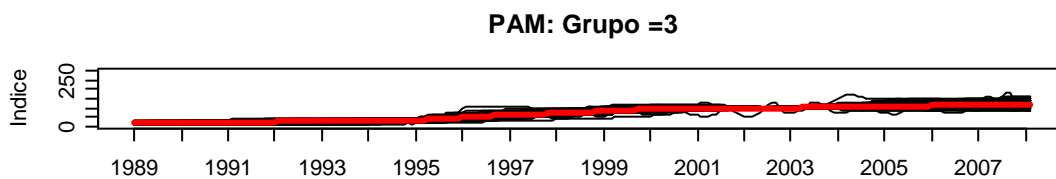
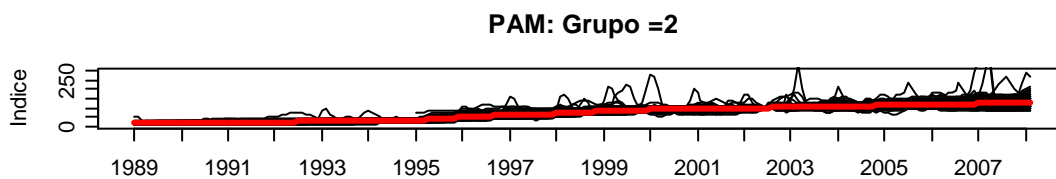
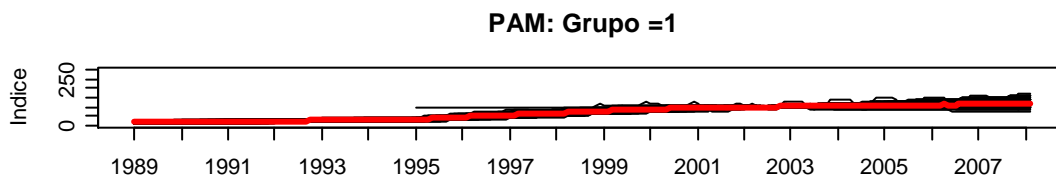
Cuando se trabaja con el índice respecto al mes anterior, se elimina la tendencia, en este caso se puede observar la volatilidad de las series



Notamos que los grupos 1 y 3 son menos volátiles que los grupos 2 y 4.

Gráfica de los perfiles de las series en índice. Método para la elaboración de conglomerados: PAM.

A pesar de que ésta es la información base de todo el análisis, aquí se observa sólo la tendencia de las series.



Composición de los Conglomerados Método: PAM

Grupo	Descripción del Grupo	Elementos del Grupo en el Conglomerado	Relativo al Conglomerado	Relativo al Grupo	Distancia de Centro del grupo a la serie del tipo de cambio
1	Alimentos	15	17.9	12.3	.18
1	Educación	12	14.3	40.0	
1	Muebles	13	15.5	31.0	
1	Otros	2	2.4	28.6	
1	Ropa	22	26.2	56.4	
1	Salud	10	11.9	25.0	
1	Transporte	7	8.3	33.3	
1	Vivienda	3	3.6	21.4	
2	Alimentos	52	47.7	42.6	.30
2	Educación	11	10.1	36.7	
2	Muebles	10	9.2	23.8	
2	Otros	2	1.8	28.6	
2	Ropa	5	4.6	12.8	
2	Salud	10	9.2	25.0	
2	Transporte	9	8.3	42.9	
2	Vivienda	10	9.2	71.4	
3	Alimentos	43	44.8	35.2	.20
3	Educación	4	4.2	13.3	
3	Muebles	17	17.7	40.5	
3	Otros	2	2.1	28.6	
3	Ropa	11	11.5	28.2	
3	Salud	16	16.7	40.0	
3	Transporte	3	3.1	14.3	
4	Alimentos	12	46.2	9.8	.8
4	Educación	3	11.5	10.0	
4	Muebles	2	7.7	4.8	
4	Otros	1	3.8	14.3	
4	Ropa	1	3.8	2.6	
4	Salud	4	15.4	10.0	
4	Transporte	2	7.7	9.5	
4	Vivienda	1	3.8	7.1	

El conglomerado 1 muestra una mayor dependencia respecto al mismo mes del año anterior que de su historia más reciente y está compuesto en un 26% por ropa, que además corresponde al 56.4% del total de elementos del grupo ropa, en este grupo caen también el 40% del grupo de educación y arriba del 30% de los grupos de muebles y transporte, mientras que solamente el 12.3% del grupo de alimentos, que no corresponde a alimentos frescos salvo el caso del durazno que es una fruta de temporada y de algunos pescados también de temporada. No es de sorprender que en este grupo caen también los precios de los servicios educativos desde primaria, secundaria, preparatoria, carreras cortas y hasta jardín de niños o guarderías cuyas colegiaturas suelen revisarse anualmente. Con el mismo comportamiento encontramos a los precios de los hoteles, clubes deportivos, servicios funerarios y licencias.

En el grupo 2 se observa una mayor dependencia respecto a los meses previos y está compuesto en un 47.7 % por alimentos, además contiene al 42.6% del grupo de alimentos, este grupo contiene a casi todos los alimentos frescos, además de los elaborados con harinas como el pan, galletas y tortillas, otros gastos como son el mantenimiento de la vivienda y el suministro de gas y servicio telefónico, del grupo de cuidados a la salud encontramos algunos medicamentos, consulta médica y hospitalización en general. Comprende también lo relacionado con el transporte público y del grupo educación pertenecen a este conglomerado lo relacionado con los precios de los libros de texto, cuadernos, periódicos y revistas.

El grupo 3 muestra un poco de dependencia respecto al mes anterior, en este grupo se encuentran algunas carnes y vísceras, además de alimentos procesados y envasados como son el jamón, chorizo y otros embutidos, la leche en polvo y evaporada o condensada, los quesos, el puré de tomate, la mayonesa, la mostaza, los concentrados de pollo y otros condimentos, la gelatina, carnitas, refrescos, vinos y cervezas. Del grupo de ropa aparecen algunos elementos principalmente a base de algodón, del grupo de muebles aparecen los aparatos eléctricos, del grupo de salud, algunos medicamentos y lo relacionado con el cuidado personal y otros servicios como son las loncherías y los restaurantes.

El grupo 4 muestra también una ligera dependencia respecto al mes anterior, pero menor a la del grupo 3, en el aparecen algunos alimentos frescos y otros pocos elementos del resto de los grupos, por lo que da la impresión de contener a aquellos elementos de los grupos con comportamientos atípicos.

Para ver la información a detalle, será necesario revisar el anexo 4.

En relación al comportamiento respecto al tipo de cambio vemos que en este caso el centro del conglomerado 1 tiene menor distancia respecto a la serie del tipo de cambio, es decir los productos incluidos en este conglomerado muestran mayor sensibilidad a variaciones cambiarias y son los siguientes:

Del grupo alimentos, bebidas y tabaco: tortilla de maíz, galletas populares, tocino, otros pescados, robaló y mero, huachinango, leche pasteurizada y fresca, queso Oaxaca o asadero, durazno, frutas y legumbres preparadas para bebida, refrescos envasados, otros alimentos cocinados, barbacoa o birria, pizzas, brandy.

Del grupo ropa, calzado y accesorios: camisas, trajes, blusas para mujer, medias y pantimedias, pantalón para mujer otros materiales, conjuntos y otras prendas para mujer, vestido para mujer, falda para mujer, pantalón niño otros materiales, camisas y playeras para niño, ropa interior para niño, ropa interior para niña, calcetines y calcetas, traje para bebé, camiseta para bebé, chamarras y abrigos, uniforme para niño, uniforme para niña, zapatos para mujer, zapatos para niños, zapatos de material sintético, bolsas, maletas y cinturones.

Del grupo vivienda: electricidad, servicio telefónico local, y línea telefónica.

Del grupo muebles, aparatos y accesorios domésticos: comedores, colchones, recámaras, otros aparatos eléctricos, computadoras, focos, otros utensilios de cocina, reproductores de video, colchas, otros blancos para el hogar, sábanas, cortinas, suavizantes y limpiadores.

Del grupo salud: otros medicamentos, material de curación, lentes y otros aparatos, consulta médica durante el embarazo, loción y perfumes, otros artículos de tocador, otros artículos de maquillaje, papel higiénico, servilletas de papel y pañuelos desechables.

Del grupo transporte: autobús foráneo, gasolina de bajo octanaje, gasolinas de alto octanaje, acumuladores, mantenimiento de automóvil, estacionamiento y cuotas de autopista.

Del grupo educación y esparcimiento: universidad, primaria, preparatoria, secundaria, preprimaria, carrera corta, jardín de niños y guardería, hoteles, otras diversiones, club deportivo, alimentos para mascotas, discos y cassettes e instrumentos musicales y otros.

Del grupo otros servicios: servicios funerarios, licencias y otros documentos.

Se observa consistencia en el tipo de productos cuyo comportamiento es más sensible a la paridad cambiaria, independientemente del método estadístico de agrupación utilizado, vemos por ejemplo del grupo alimentos, bebidas y tabaco que sobresalen especies de pescados cuya producción nacional es limitada como el robalo, mero y huachinango. De éstas y otras especies se ha incrementado la importación cuya procedencia es principalmente de países del sureste asiático.

Nuevamente aparecen productos de consumo final como la ropa, el calzado y accesorios como bolsas y cinturones.

Del grupo vivienda aparecen la electricidad y la línea telefónica cuyos insumos son en gran medida de importación como la fibra óptica.

La apertura comercial que caracteriza al período de estudio ha propiciado la aparición de importantes cadenas comerciales transnacionales que ofrecen gran cantidad de productos de importación y consumo final como son colchones, recámaras, reproductores de video, entre otros. Así como variedad de marcas de suavizantes y limpiadores.

Del grupo salud aparecen materiales de curación, lentes y otros aparatos cuya importación es cada vez más común.

El país importa también gasolina de la cual algunas cifras señalan que se importa alrededor del 50% del consumo nacional.

Del grupo educación y esparcimiento no sorprende encontrar los hoteles pero si el costo de las colegiaturas en distintos niveles educativos.

Tampoco tenemos mayor explicación para incluir en este grupo la expedición de licencias y otros documentos. Salvo con fines recaudatorios.

Por otro lado el conglomerado cuyo centro muestra mayor distancia con la serie del tipo de cambio es el conglomerado número 4.

Del grupo alimentos, bebidas y tabaco: melón, sandía, jitomate, tomate verde, calabacita, chile serrano, otros chiles frescos, chile poblano, nopales, chayote, pepino, ejotes.

Del grupo ropa, calzado y accesorios: solamente aparece el elemento servicios de lavandería.

Del grupo vivienda: solamente aparece la larga distancia internacional.

Del grupo muebles, aparatos y accesorios domésticos: aparatos de aire acondicionado y blanqueadores.

Del grupo salud: prótesis dental, estudios médicos de gabinete, hospitalización por parto, artículos de maquillaje facial.

Del grupo transporte: servicios de reparación de automóvil y lavado y engrasado de automóvil.

De grupo educación y esparcimiento: enseñanza adicional, servicios turísticos en paquete y renta de películas.

Del grupo otros servicios: servicios profesionales.

Este grupo tiene gran cantidad de elementos en común con el conglomerado de mayor distancia a la serie de precios obtenido por el método de K-MEANS, por lo que no es necesario añadir más comentarios, salvo la consistencia que se obtiene de manera independiente al método de agrupamiento.

Finalmente, resulta interesante observar la suma de los ponderadores de cada grupo:

PAM

Conglomerado	Elementos	Suma de Ponderadores
1	84	27.03156
2	109	46.54221
3	96	22.55560
4	26	3.87063
Total	315	100.00000

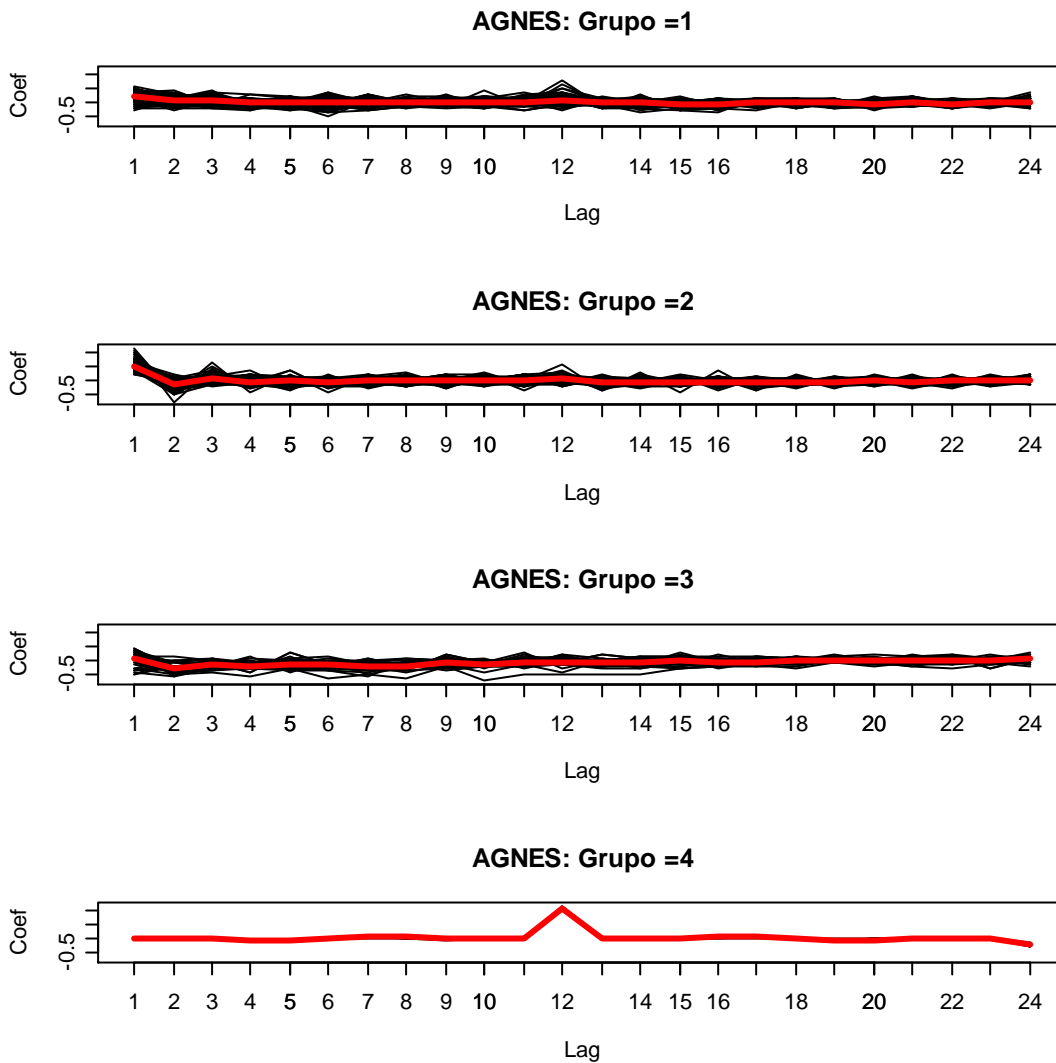
Vemos que el grupo que contiene el mayor número de elementos, de hecho 109 de los 315 productos genéricos representan el 34.6% de los productos y aportan el 46.54% del valor del índice.

3.3 Resultados obtenidos por el método AGNES.

Gráfica de los perfiles de los coeficientes para cada uno de los grupos. Método para la elaboración de conglomerados: AGNES.

Con éste tipo de gráfico observaremos las características generales de correlaciones de las series en cada grupo.

Los conglomerados se forman a partir de los coeficientes, aquí graficamos los coeficientes del $AR_{(24)}$, de cada serie de incrementos de precios para cada grupo. Los conglomerados se formaron con el método AGNES.



El grupo 1, no muestra dependencias claras casi con ningún mes, excepto tal vez con el mes inmediato anterior.

El grupo 2 muestra una dependencia mayor del mes inmediato anterior.

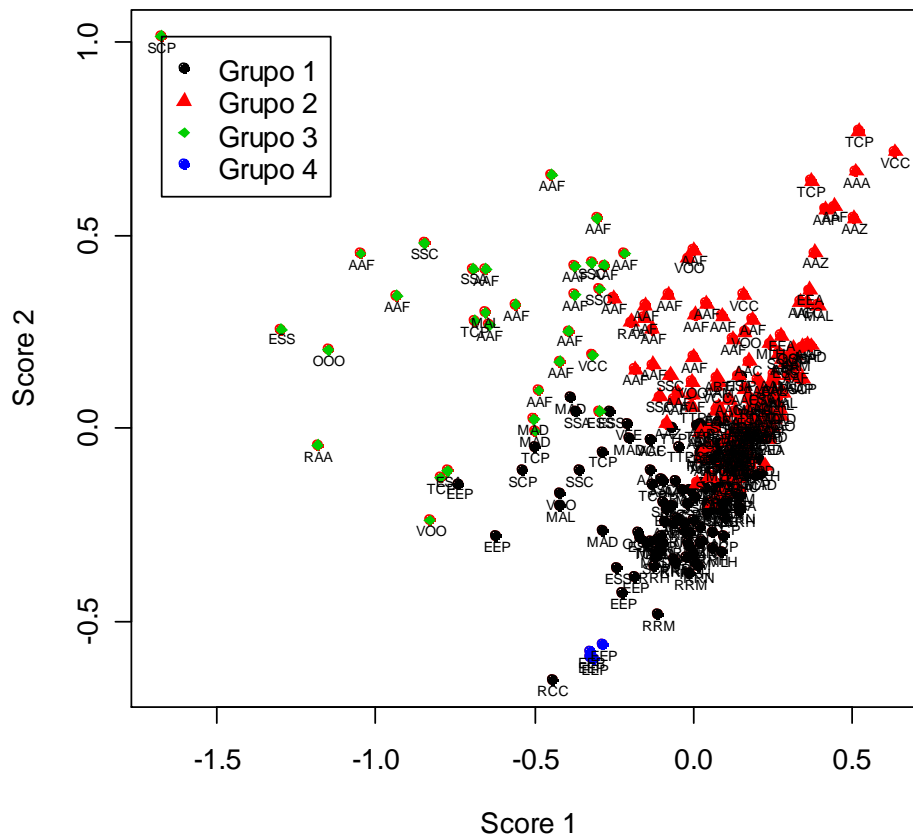
El grupo 3 muestra dependencias casi exclusivamente con el mes inmediato anterior.

En el grupo 4 muestra dependencia clara con el mismo mes del año anterior.

Gráfica de componentes principales. Método para la elaboración de conglomerados: AGNES.

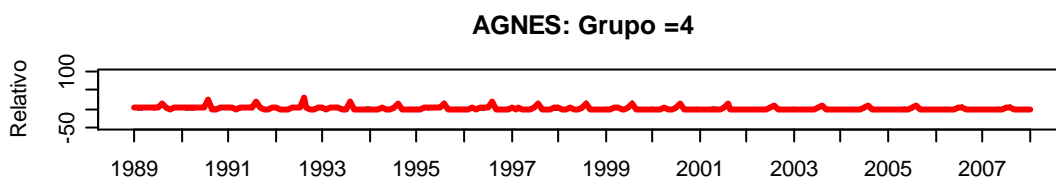
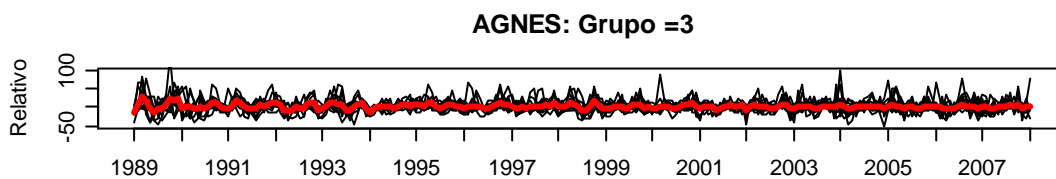
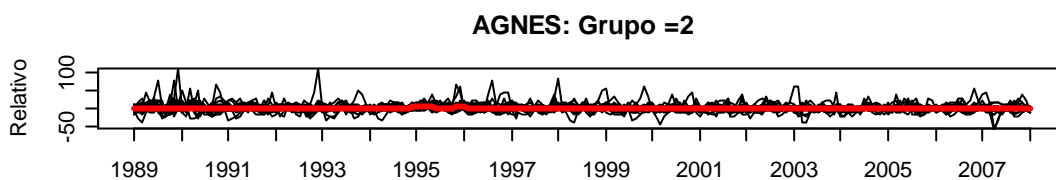
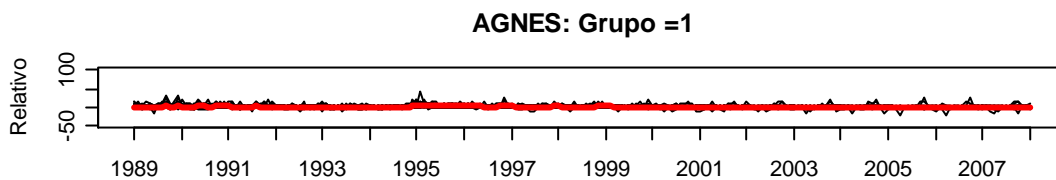
Se utiliza el método de componentes principales con fines de visualización, sólo como herramienta de reducción de dimensiones de los coeficientes de 24 a 2 dimensiones. Si los grupos se iluminan de manera separada en la nube de puntos esto es una indicación de grupos agradable. El método utilizado en este caso para la formación de los conglomerados es AGNES.

Componentes Principales



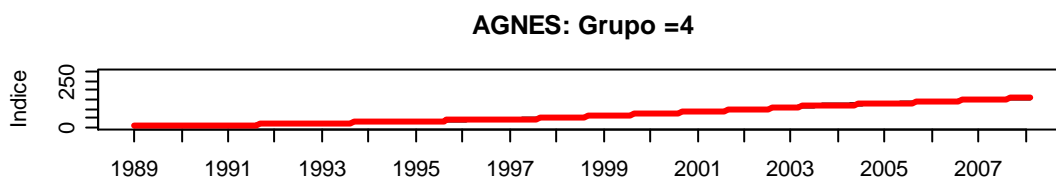
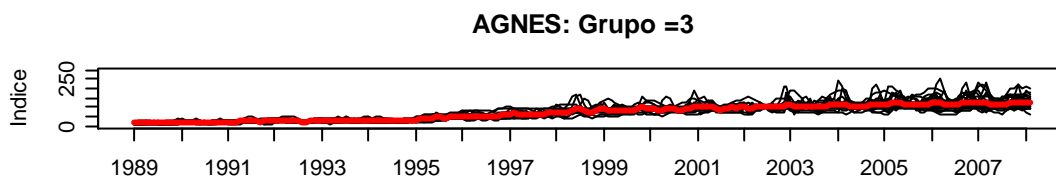
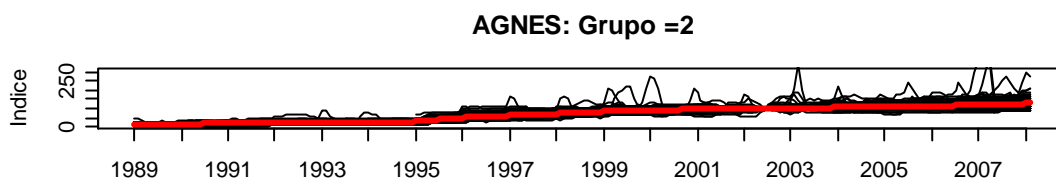
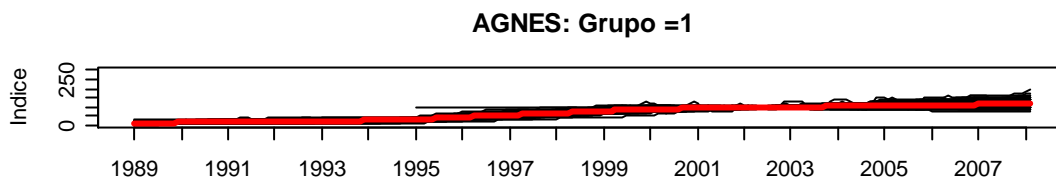
Gráfica de los perfiles de las series relativas para cada uno de los grupos. Método para la elaboración de conglomerados: AGNES.

Cuando se trabaja con el índice respecto al mes anterior, se elimina la tendencia, en este caso se puede observar la volatilidad de las series

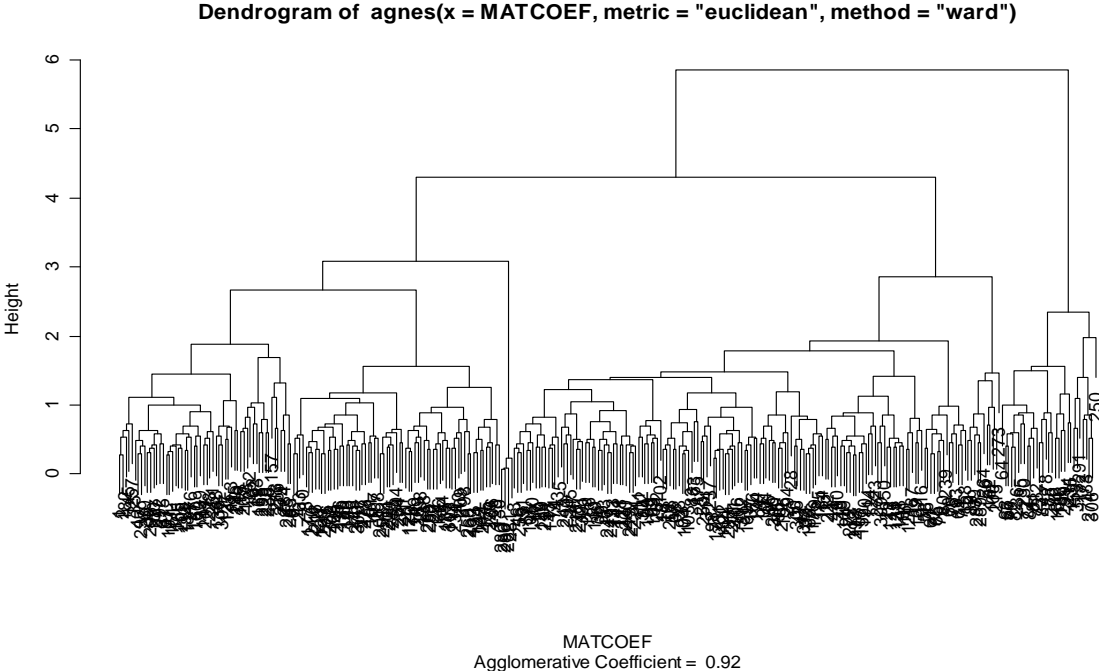


Gráfica de los perfiles de las series en índice. Método para la elaboración de conglomerados: AGNES.

A pesar de que esta es la información base de todo el análisis se tiene que aquí se observa sólo la tendencia de las series.



Para este mismo método para la conformación de conglomerados, mostramos ahora el dendograma.



De este algoritmo tomamos cuatro conglomerados con el siguiente número de series:

Grupos				Total de Series
1	2	3	4	
123	157	31	4	315

Composición de los Conglomerados Método: AGNES

Grupo	Descripción del Grupo	Elementos del Grupo en el Conglomerado	Relativo al Conglomerado	Relativo al Grupo	Distancia de Centro del grupo a la serie del tipo de cambio
1	Alimentos	23	18.7	18.9	.13
1	Educación	14	11.4	46.7	
1	Muebles	24	19.5	57.1	
1	Otros	1	0.8	14.3	
1	Ropa	27	22.0	69.2	
1	Salud	16	13.0	40.0	
1	Transporte	14	11.4	66.7	
1	Vivienda	4	3.3	28.6	
2	Alimentos	85	54.1	69.7	
2	Educación	9	5.7	30.0	
2	Muebles	15	9.6	35.7	
2	Otros	5	3.2	71.4	
2	Ropa	11	7.0	28.2	
2	Salud	19	12.1	47.5	
2	Transporte	5	3.2	23.8	
2	Vivienda	8	5.1	57.1	
3	Alimentos	14	45.2	11.5	.68
3	Educación	3	9.7	10.0	
3	Muebles	3	9.7	7.1	
3	Otros	1	3.2	14.3	
3	Ropa	1	3.2	2.6	
3	Salud	5	16.1	12.5	
3	Transporte	2	6.5	9.5	
3	Vivienda	2	6.5	14.3	
4	Educación	4	100.0	13.3	1.57

Con el método para la elaboración de conglomerados AGNES, se formaron 4 grupos. El primero de ellos muestra poca relación con los meses previos, algunas de las series que quedaron contenidas en este grupo muestran mayor relación con los precios del mismo mes pero del año anterior, por lo que no es de sorprender que este grupo contenga al 46.7% de los elementos del grupo de educación, al 57.1% de los elementos del grupo muebles, al 69.2% del grupo ropa, al 40% del grupo salud y al 66.7% del grupo transporte, en el aparecen los alimentos elaborados con harina como son la tortilla, el pan, pastelitos y la pasta para sopa, también aparecen algunas carnes y vísceras y en general elementos de otros grupos de los que ya se había hablado típicamente sufren revisiones en sus precios de manera anual.

El segundo grupo muestra relación respecto al mes anterior, en el aparecen el 69.7% de los elementos del grupo alimentos, casi la mitad de los elementos del grupo salud, lo referente al transporte público, del grupo educación aparecen los impresos como son libros de texto, cuadernos, periódicos y revistas, finalmente casi todos los elementos de grupo de otros productos y servicios como son los restaurantes, las loncherías, cafeterías y cantinas, además de los servicios funerarios.

El tercer grupo muestra también relación respecto al mes anterior, pero menor a la del grupo 2, en él aparecen algunos elementos del grupo de alimentos que corresponden a alimentos frescos que están disponibles todo el año en nuestro país, entre ellos la calabacita, el chile serrano, los nopales, el chayote, el pepino, entre otros. También aparecen algunos elementos del grupo salud relacionados con las prótesis, estudios de gabinete y la hospitalización, del grupo correspondiente a otros productos y servicios, aparecen únicamente los servicios profesionales.

Finalmente el grupo 4 muestra relación con los precios respecto al mismo mes pero del año anterior. Este conglomerado quedó compuesto únicamente por elementos del grupo educación que van desde el jardín de niños y guardería, primaria, secundaria y preparatoria.

Para ver la información a detalle será necesario revisar el anexo 4.

En relación al comportamiento respecto al tipo de cambio vemos que en este caso el centro del conglomerado 1 tiene menor distancia respecto a la serie del tipo de cambio, es decir los productos incluidos en este conglomerado muestran mayor sensibilidad a variaciones cambiarias y corresponde al conglomerado número 1 obtenido por el método de agrupamiento PAM y son los siguientes:

Del grupo alimentos, bebidas y tabaco: tortillas de maíz, masa y harina de maíz, pan de caja, pastelillos y pasteles, pasta para sopa, tortillas de harina de trigo, bistec de res, carne molida de res, retazo, hígado de res, tocino, otros pescados, robalo y mero, huachinango, durazno, queso Oaxaca o asadero, frutas y legumbres preparados para bebida, refrescos envasados, otros condimentos, otros alimentos cocinados, pizzas, tequila, brandy,

Del grupo ropa, calzado y accesorios: camisas, pantalón hombre base algodón, trajes, pantalón hombre otros materiales, otras prendas para hombre, blusas para mujer, medias y pantimedias, pantalón mujer otros materiales, conjuntos y otras prendas para mujer, vestido para mujer, falda para mujer, pantalón niño base algodón, pantalón niño otros materiales, camisas y playeras para niño, ropa interior para niño, ropa interior para niña, calcetines y calcetas, traje para bebé, camiseta para bebé, sueter para niño y niña, uniforme para niño, uniforme para niña, zapatos para mujer, zapatos para niño, zapatos de material sintético, bolsas maletas y cinturones y relojes, joyas y bisutería.

Del grupo vivienda: vivienda propia, derechos por el suministro de agua, electricidad, servicio telefónico local, línea telefónica.

Del grupo muebles: antecomedores, estufas, salas, comedores, colchones, recámaras, refrigeradores, televisores, computadoras, reproductores de video, focos, cerillos, escobas, loza y cristalería, utensilios de plástico para el hogar, otros utensilios de cocina, colchas, otros blancos para el hogar, sábanas, cobijas, toallas, suavizantes y limpiadores, desodorantes ambientales, plaguicidas.

Del grupo salud: cardiovasculares, otros medicamentos, dermatológicos, antigripales, material de curación, lentes y otros aparatos, cuidado dental, consulta médica durante el embarazo, análisis clínicos durante el embarazo, lociones y perfumes, cremas para la piel, otros artículos de tocador, otros artículos de maquillaje, pañales, toallas sanitarias, pañuelos desechables.

Del grupo transporte: transporte público urbano colectivo, autobús urbano, taxi, transporte aéreo, bicicletas, gasolina de bajo octanaje, gasolina de alto octanaje, aceites lubricantes, neumáticos, otras refacciones, acumuladores, mantenimiento de automóvil, cuotas de autopista y estacionamiento.

Del grupo educación y esparcimiento: universidad, preprimaria, carrera corta, enseñanza adicional, otros libros, hoteles, cine, servicio de televisión por cable o satélite, otras diversiones, club deportivo, discos casetes, alimento para mascotas, instrumentos musicales y otros, artículos deportivos.

Del grupo otros servicios: licencias y otros documentos.

Además de productos cuyas variaciones aparecieron cercanas a las de la serie de tipo de cambio que ya fueron comentadas, se incorporan productos cárnicos. En tiempos recientes llega al país carne procedente de la Unión Americana.

Por otro lado el conglomerado cuyo centro muestra mayor distancia con la serie del tipo de cambio es el conglomerado número 4, que únicamente contiene cuatro elementos del grupo educación y son los siguientes:

De grupo educación y esparcimiento: jardín de niños y guardería, primaria, preparatoria y secundaria.

Este grupo comprende lo correspondiente a colegiaturas de niveles educativos básico y medio, lo cual parece natural que no debieran incrementarse por una depreciación cambiaria de nuestra moneda.

Finalmente, resulta interesante observar la suma de los ponderadores de cada grupo:

AGNES		
Conglomerado	Elementos	Suma de Ponderadores
1	123	33.62011
2	157	58.12207
3	31	5.43714
4	4	2.82068
Total	315	100.00000

Vemos que el grupo que contiene el mayor número de elementos, de hecho 157 de los 315 productos genéricos representan el 49.84% de los productos y aportan el 58.12% del valor del índice.

4 Conclusiones

La formación de conglomerados de series de tiempo se ha convertido en un área relevante para la investigación por sus aplicaciones en numerosos campos como son la economía, la biología y las finanzas. Existen diferentes formas de abordar esta tarea, una de ellas es bajo el enfoque del análisis funcional, es decir se busca una expresión analítica de una función que capture la estructura y dinámica de las series de tiempo. Sin embargo, en la práctica, pocos casos ameritan llevar a los usuarios a complicaciones propias del caso continuo.

La característica principal del análisis funcional es que consiste en funciones, frecuentemente curvas suavizadas. Los objetivos del análisis funcional son los mismos que en otras ramas de la estadística: formular el problema de forma tal que sea útil para el análisis, desarrollar formas de presentación de los datos que permitan visualizar características importantes, investigar la variabilidad, así como las características de la media, construir modelos para los datos observados, incluidos aquellos que muestran dependencia entre observaciones o variables, entre otras.

Los pasos a seguir en el análisis funcional son: la representación de los datos utilizando técnicas de suavizamiento e interpolación. Registro de datos o alineación a rasgos, desplegar los datos y finalmente obtener el resumen estadístico como son las funciones de la media y la varianza, la covarianza y funciones de correlación para posteriormente explorar la variación con mayor detalle.

Es posible encontrar en la literatura, métricas que permiten medir similitudes en el análisis funcional.

Otro enfoque se presenta en la medida de distancia entre series de tiempo con base en el periodograma normalizado. En su estudio se realizan simulaciones para comparar la propuesta con otras métricas paramétricas y no paramétricas. Particularmente se analiza la clasificación de series estacionarias y no estacionarias, así como el uso de algoritmos de agrupamiento jerárquicos y no jerárquicos⁹.

En su trabajo estudian métricas basadas en diferentes medidas de dependencia para clasificar series como estacionarias o no estacionarias. Con simulaciones muestran que las métricas basadas en el logaritmo del periodograma normalizado y en los coeficientes de autocorrelación pueden ambas distinguir empíricamente con éxito entre modelos ARMA y ARIMA, lo que no sucede con la distancia Euclideana ni con la métrica basada en los ponderadores autorregresivos¹⁰.

⁹ Caiado, *et al*(2005)

¹⁰ Piccolo(1990)

En este trabajo, utilizamos tres algoritmos para la construcción de los conglomerados, formando en cada caso cuatro grupos que podríamos etiquetar como el grupo de los “alimentos frescos”, el grupo de los “alimentos procesados”, el grupo de la “ropa” y el grupo “heterogéneo”.

De manera intuitiva, resulta claro que el comportamiento de los alimentos frescos sea distinto al de los alimentos procesados ya que los primeros tienen una vida más corta. Por otro lado se antoja que los precios de la ropa tengan un comportamiento estacional, es decir el precio de la ropa de invierno, tenga una mayor relación con el precio de la ropa de esa misma temporada pero del año anterior que con la ropa de verano por ejemplo.

Revisemos brevemente los elementos comunes que aparecen en los conglomerados generados por los tres diferentes métodos de agrupamiento empleados en este trabajo.

Comparativo de los resultados obtenidos por los tres métodos.

Elementos comunes en cada conglomerado según el método de agrupación

GrupoPAM	GrupoAGNES				Total
	1	2	3	4	
1	70	10		4	84
2	19	84	6		109
3	33	63			96
4	1		25		26
Total	123	157	31	4	315

GrupoKMEANS	GrupoPAM				Total
	1	2	3	4	
1		41	7		48
2	28	59	86		173
3	2	7		26	35
4	54	2	3		59
Total	84	109	96	26	315

GrupoKMEANS	GrupoAGNES				Total
	1	2	3	4	
1	1	46	1		48
2	67	106			173
3	5		30		35
4	50	5		4	59
Total	123	157	31	4	315

Del cuadro anterior se desprenden los conglomerados elaborados por un método de agrupamiento que “corresponden” o tienen mayor número de coincidencias con los conglomerados generados por los otros métodos de agrupamiento.

Conglomerados con mayor número de elementos comunes

KMEANS	PAM	AGNES
1	2	2
2	3	2
3	4	3
4	1	1

Revisamos a continuación los elementos comunes a los conglomerados KMEANS(1), PAM(2) y AGNES(2).

Elementos comunes entre los conglomerados con mayores coincidencias formados a partir de los tres métodos.

Grupo	Grupo KMEANS	Grupo PAM	Grupo AGNES	Descripción
Alimentos	1	2	2	Pan Blanco
Alimentos	1	2	2	Harinas de trigo
Alimentos	1	2	2	Pollo entero
Alimentos	1	2	2	Chuletas y manteca de cerdo
Alimentos	1	2	2	Lomo
Alimentos	1	2	2	Pierna
Alimentos	1	2	2	Mojarra
Alimentos	1	2	2	Aceites y grasas vegetales comestibles
Alimentos	1	2	2	Manzana
Alimentos	1	2	2	Aguacate
Alimentos	1	2	2	Mango
Alimentos	1	2	2	Limón
Alimentos	1	2	2	Uva
Alimentos	1	2	2	Pera
Alimentos	1	2	2	Toronja
Alimentos	1	2	2	Piña
Alimentos	1	2	2	Guayaba
Alimentos	1	2	2	Cebolla
Alimentos	1	2	2	Zanahoria

Grupo	Grupo KMEANS	Grupo PAM	Grupo AGNES	Descripción
Alimentos	1	2	2	Lechuga y col
Alimentos	1	2	2	Frijol
Alimentos	1	2	2	Chile seco
Alimentos	1	2	2	Café soluble
Alimentos	1	2	2	Café tostado
Alimentos	1	2	2	Cigarrillos
Ropa	1	2	2	Zapatos para hombre
Vivienda	1	2	2	Vivienda alquilada Renta de vivienda
Vivienda	1	2	2	Vivienda propia Vivienda propia
Vivienda	1	2	2	Vivienda propia Mantenimiento de vivienda servicios
Vivienda	1	2	2	Vivienda propia Impuesto predial
Vivienda	1	2	2	Servicio doméstico Servicio doméstico
Vivienda	1	2	2	Servicios de la vivienda Otros servicios para el hogar
Muebles	1	2	2	Baterías de cocina
Salud	1	2	2	Expectorantes y descongestivos
Transporte	1	2	2	Vehículos automotores y de pedal Automóviles
Transporte	1	2	2	Servicios para automóvil Tenencia de automóvil
Educación	1	2	2	Libros de texto
Educación	1	2	2	Cuadernos y carpetas
Otros	1	2	2	Cantinas

Estos conglomerados tienen en común, algunas frutas y vegetales frescos, algunas carnes, del grupo ropa únicamente los zapatos para hombre, del grupo vivienda , se tiene la renta de vivienda, la vivienda propia, los servicios de mantenimiento de la vivienda; los grupos muebles, salud y otros, solamente aportan un elemento, recordemos que el conglomerado 1 de KMEANS, el conglomerado 2 de PAM y el conglomerado 2 de AGNES, tienen en común que muestran dependencias más claras con el mes inmediato anterior, además de algo de dependencia con otros meses cercanos.

Ahora revisamos los elementos comunes a los conglomerados KMEANS(2), PAM(3) y AGNES(2).

Elementos comunes entre los conglomerados con mayores coincidencias formados a partir de los tres métodos.

Grupo	Grupo KMEANS	Grupo PAM	Grupo AGNES	Descripción
Alimentos	2	3	2	Maíz
Alimentos	2	3	2	Arroz

Grupo	Grupo KMEANS	Grupo PAM	Grupo AGNES	Descripción
Alimentos	2	3	2	Pollo en piezas
Alimentos	2	3	2	Pulpa de cerdo
Alimentos	2	3	2	Otras vísceras de res
Alimentos	2	3	2	Jamón
Alimentos	2	3	2	Chorizo
Alimentos	2	3	2	Otros embutidos
Alimentos	2	3	2	Carnes secas
Alimentos	2	3	2	Camarón
Alimentos	2	3	2	Leche evaporada, condensada y maternizada
Alimentos	2	3	2	Queso fresco
Alimentos	2	3	2	Crema de leche
Alimentos	2	3	2	Queso chihuahua o manchego
Alimentos	2	3	2	Otros quesos
Alimentos	2	3	2	Otras legumbres
Alimentos	2	3	2	legumbres secas
Alimentos	2	3	2	Chiles procesados
Alimentos	2	3	2	Verduras envasadas
Alimentos	2	3	2	Puré de tomate y sopas enlatadas
Alimentos	2	3	2	Mayonesa y mostaza
Alimentos	2	3	2	Concentrados de pollo y sal
Alimentos	2	3	2	Papas fritas y similares
Alimentos	2	3	2	Concentrados para refrescos
Alimentos	2	3	2	Gelatina en polvo
Alimentos	2	3	2	Carnitas
Alimentos	2	3	2	Pollos rostizados
Alimentos	2	3	2	Otros licores
Alimentos	2	3	2	Ron
Alimentos	2	3	2	Vino de mesa
Ropa	2	3	2	Ropa interior para hombre
Ropa	2	3	2	Ropa interior para mujer
Ropa	2	3	2	Pantalón mujer base algodón
Ropa	2	3	2	Vestido para niña
Ropa	2	3	2	Sombreros

Grupo	Grupo KMEANS	Grupo PAM	Grupo AGNES	Descripción
Ropa	2	3	2	Zapatos tenis
Muebles	2	3	2	Calentadores para agua
Muebles	2	3	2	Ventiladores
Muebles	2	3	2	Planchas eléctricas
Muebles	2	3	2	Licadoras
Muebles	2	3	2	Equipos modulares
Muebles	2	3	2	Radios y grabadoras
Salud	2	3	2	Analgésicos
Salud	2	3	2	Nutricionales
Salud	2	3	2	Anticonceptivos y hormonales
Salud	2	3	2	Gastrointestinales
Salud	2	3	2	Atención médica durante el parto
Salud	2	3	2	Sala de belleza
Salud	2	3	2	Productos para el cabello
Salud	2	3	2	Desodorantes personales
Salud	2	3	2	Navajas y máquinas de afeitar
Transporte	2	3	2	Servicios para automóvil Seguro de automóvil
Educación	2	3	2	Plumas, lápices y otros
Otros	2	3	2	Loncherías
Otros	2	3	2	Restaurantes

Estos conglomerados comparten elementos comunes que corresponden a alimentos procesados o preparados como son jamón, chorizo, otros embutidos, carnes secas, leche evaporada, condensada y maternizada, algunos quesos, chiles procesados, verduras envasadas, puré de tomate y sopas enlatadas, papas fritas y similares, concentrados para refrescos, gelatina, carnitas, pollos rostizados, ron, vino de mesa, etc. Parece natural que estos alimentos quedaran comprendidos en un mismo conglomerado y que sucediera sin importar cuál de los tres métodos de agrupamiento fue utilizado. Aparecen también seis elementos del grupo ropa, del grupo muebles aparecen únicamente aparatos eléctricos con excepción de los calentadores de agua. Del grupo salud aparecen algunos medicamentos, además de productos para el cabello, desodorantes personales, máquinas de afeitar y salas de belleza, lo cual indica que el comportamiento en estos conglomerados es más representativo de los productos del cuidado personal que los propios de la salud como medicamentos, consultas médicas u hospitalización. Finalmente se incluyen también del grupo educación, únicamente la plumas y lápices y del grupo otros las loncherías y los restaurantes.

Recordemos que estos conglomerados mostraban dependencias casi exclusivamente con el mes inmediato anterior.

Ahora revisamos los elementos comunes a los conglomerados KMEANS(3), PAM(4) y AGNES(3).

Elementos comunes entre los conglomerados con mayores coincidencias formados a partir de los tres métodos.

Grupo	Grupo KMEANS	Grupo PAM	Grupo AGNES	Descripción
Alimentos	3	4	3	Melón
Alimentos	3	4	3	Sandía
Alimentos	3	4	3	Jitomate
Alimentos	3	4	3	Tomate verde
Alimentos	3	4	3	Calabacita
Alimentos	3	4	3	Chile serrano
Alimentos	3	4	3	Otros chiles frescos
Alimentos	3	4	3	Chile poblano
Alimentos	3	4	3	Nopales
Alimentos	3	4	3	Chayote
Alimentos	3	4	3	Pepino
Alimentos	3	4	3	Ejotes
Ropa	3	4	3	Servicio de lavandería
Vivienda	3	4	3	Servicio telefónico Larga distancia internacional
Muebles	3	4	3	Aparatos de aire acondicionado
Muebles	3	4	3	Blanqueadores
Salud	3	4	3	Prótesis dentales
Salud	3	4	3	Estudios médicos de gabinete
Salud	3	4	3	Hospitalización parto
Salud	3	4	3	Artículos de maquillaje facial
Transporte	3	4	3	Servicios para el automóvil Reparación de automóvil
Transporte	3	4	3	Servicios para el automóvil Lavado y engrasado de automóvil
Educación	3	4	3	Servicios turísticos en paquete
Educación	3	4	3	Renta de películas
Otros	3	4	3	Servicios profesionales

Estos conglomerados comparten elementos comunes que corresponden a algunas frutas y vegetales frescos como melón, sandía, jitomate, tomate verde, calabacita, chile serrano, chile

poblano, nopales, chayote, pepino y ejotes, los cuales además están disponibles en nuestro país prácticamente durante todo el año. De los grupos de ropa y vivienda, en realidad aparecen un par de servicios que son el de lavandería y la telefonía de larga distancia internacional, respectivamente, al igual que del grupo transporte en donde aparece el servicio de reparación de automóvil y el de lavado y engrasado. Finalmente de los grupos educación y otros también aparecen servicios como son los turísticos en paquete, la renta de películas y otros servicios profesionales.

Recordemos que estos conglomerados mostraban la característica de no depender tanto del mes inmediato anterior, sino del segundo y tercer mes.

Finalmente revisamos los elementos comunes a los conglomerados KMEANS(4), PAM(1) y AGNES(1).

Elementos comunes entre los conglomerados con mayores coincidencias formados a partir de los tres métodos.

Grupo	Grupo KMEANS	Grupo PAM	Grupo AGNES	Descripción
Alimentos	4	1	1	Otros pescados
Alimentos	4	1	1	Robalo y mero
Alimentos	4	1	1	Huachinango
Alimentos	4	1	1	Queso Oaxaca o asadero
Alimentos	4	1	1	Durazno
Alimentos	4	1	1	Frutas y legumbres preparadas para bebida
Alimentos	4	1	1	Refrescos envasados
Ropa	4	1	1	Trajes
Ropa	4	1	1	Blusas para mujer
Ropa	4	1	1	Medias y pantimedias
Ropa	4	1	1	Pantalón mujer otros materiales
Ropa	4	1	1	Conjuntos y otras prendas para mujer
Ropa	4	1	1	Vestido para mujer
Ropa	4	1	1	Falda para mujer
Ropa	4	1	1	Pantalón niño otros materiales
Ropa	4	1	1	Camisas y playeras para niño
Ropa	4	1	1	Ropa interior para niño
Ropa	4	1	1	Ropa interior para niña
Ropa	4	1	1	Calcetines y calcetas
Ropa	4	1	1	Traje para bebé

Grupo	Grupo KMEANS	Grupo PAM	Grupo AGNES	Descripción
Ropa	4	1	1	Uniforme para niño
Ropa	4	1	1	Uniforme para niña
Ropa	4	1	1	Zapatos de material sintético
Ropa	4	1	1	Bolsas, maletas y cinturones
Vivienda	4	1	1	Electricidad
Vivienda	4	1	1	Servicio telefónico Línea telefónica
Muebles	4	1	1	Colchones
Muebles	4	1	1	Recámaras
Muebles	4	1	1	Reproductores de video
Muebles	4	1	1	Colchas
Muebles	4	1	1	Otros blancos para el hogar
Muebles	4	1	1	Suavizantes y limpiadores
Salud	4	1	1	Material de curación
Salud	4	1	1	Lentes y otros aparatos
Salud	4	1	1	Consulta médica durante el embarazo
Salud	4	1	1	Otros artículos de tocador
Salud	4	1	1	Pañuelos desechables
Transporte	4	1	1	Gasolinas y aceites lubricantes Gasolina de alto octanaje
Transporte	4	1	1	Servicios para automóvil Cuotas de autopista
Educación	4	1	1	Universidad
Educación	4	1	1	Preprimaria
Educación	4	1	1	Carrera corta
Educación	4	1	1	Hoteles
Educación	4	1	1	Otras diversiones
Educación	4	1	1	Club deportivo
Educación	4	1	1	Instrumentos musicales y otros
Otros	4	1	1	Licencias y otros documentos

Estos conglomerados comparten pocos elementos comunes que corresponden al grupo de alimentos, aparecen el robalo y el huachinango que son especies de temporada y de las frutas el durazno que también es de temporada. Muchos elementos del grupo ropa, del grupo vivienda aparecen la electricidad y el servicio de línea telefónica, algunos elementos del grupo muebles como son colchones, recámaras, reproductores de video, colchas, otros blancos para el hogar y suavizantes y limpiadores. Se incluyen también cinco elementos del grupo salud como son el material de curación, los lentes, la consulta médica, otros artículos de tocador y los pañuelos

desechables. Del grupo transporte aparece la gasolina de alto octanaje y las cuotas de autopista, además de siete elementos del grupo educación como son los precios de la universidad, la preprimaria, la carrera corta, recordemos que el grupo educación incluye también el esparcimiento y en estos conglomerados quedaron incluidos los hoteles, otras diversiones, club deportivo y los instrumentos musicales y finalmente del grupo otros tenemos las licencias y otros documentos.

Recordemos que estos conglomerados mostraban la característica de dependencia con respecto al mismo mes del año anterior.

Al inicio de este estudio se habló de la importancia del Índice Nacional de Precios al Consumidor, con él se miden los cambios en los precios de los bienes y servicios que consumen los hogares, los cambios en los precios afectan el poder adquisitivo de los hogares y por tanto su nivel de bienestar.

Hemos visto que al cambiar la técnica de agrupamiento se forman conglomerados distintos pero que mantienen una cierta estructura, es decir se trata de grupos de series con comportamientos similares en lo que al incremento de precios se refiere.

Se podría profundizar en el estudio del comportamiento del grupo con mayor influencia en el índice, es decir en el grupo cuyos ponderadores sean más grandes. Estudiar el comportamiento de los bienes y servicios con mayor afectación en la inflación y por lo tanto en el nivel de bienestar de los hogares.

Así como los hogares conociendo estos patrones de comportamiento podrían tener sus previsiones en pro de la mejora de la economía del hogar, los gobiernos también podrían crear políticas orientadas a disminuir el impacto negativo que los incrementos en los precios de ciertos bienes y servicios tienen sobre la economía del país.

Una posible aplicación de este trabajo sería plantear la posibilidad de reconfigurar los grupos de bienes y servicios, en función del comportamiento histórico mostrado en el incremento de sus respectivos precios.

Punto aparte merece comentar la relación que el INPC tiene con el tipo de cambio de nuestra moneda frente al dólar americano. El periodo que comprenden nuestras series de precios caen en tres periodos de gobierno encabezados por Carlos Salinas de Gortari (1988-1994), Ernesto Zedillo Ponce de León (1994-2000) y Vicente Fox Quezada (2000-2006), periodos caracterizados por un amplio proceso de apertura económica, que a la fecha cuenta con 12 acuerdos comerciales que involucran a casi medio centenar de países.

México cuenta con una enorme oferta de productos nuevos, que cambian de manera acelerada los hábitos de consumo de la población y el aumento en la demanda de los mismos en los últimos años.

Los diferentes métodos de agrupamiento mostraron consistencia en cuanto a los elementos característicos de cada grupo.

Este trabajo plantea una estructura fija de análisis y formación de grupos del INPC, pero existen una gran variedad de posibilidades de realizar este análisis de una manera diferente, para un propósito específico, como por ejemplo:

Diferentes transformaciones del INPC (relativo al mes anterior, relativo al mismo mes año anterior, etc.).

Diferentes órdenes del modelo autorregresivo (24, 12, etc.).

Diferentes métodos de estimación de los coeficientes (Yule-Walker, Máxima Verosimilitud, etc.).

Diferentes medidas de similitud (Euclideana, Manhattan, etc.).

Diferentes métodos de agrupamiento (KMEANS, PAM, AGNES, etc.). Algunos de estos métodos de agrupamiento tienen variantes.

Diferente número de grupos a formar.

Por lo anterior es claro que no es posible trabajar con todas las variantes y nos vemos obligados a seleccionar algunas en función del propósito de este estudio.

Es amplia la gama de opciones para continuar con este estudio uno de ellos podría ser conformar clusters por regímenes y ver si existe algún cambio en los productos que los conforman, como por ejemplo si un producto cambia de un grupo a otro.

Es importante destacar que el método ayuda a agrupar características comunes en las series a través de representar cualquier serie mediante su generalidad autorregresiva, lo que facilita el análisis.

No sorprende que no se agrupen todos los alimentos, o la ropa, o los gastos de vivienda juntos en un solo grupo, puesto que pueden seguir estructuras de tiempo que son movidas por otros hechos tales como la política de apertura económica que caracteriza al periodo de estudio u otras políticas como la del sistema de tipo de cambio, recordemos que en tiempos del presidente Salinas se tenía un sistema de bandas de flotación, el cuál fue sustituido por un sistema de libre flotación a raíz de la devaluación de finales de 1994, cambios en la estructura de consumo, precios contralados, sectores protegidos, entre otros.

La importancia de los clusters radica en que es posible reducir el tiempo de análisis para fines prácticos. Se podría pronosticar los índices de un conglomerado particular, realizar pruebas a algunos índices de ese grupo y atribuirse al resto de los índices sin necesidad de probar para todos.

El estudio también podría continuar incorporando la medida de variabilidad de la estimación de los coeficientes a la distancia entre series propuesta. Como resultado del análisis realizado se ve también la necesidad de seguir con la incorporación de otros métodos de agrupamiento que consideren anidamiento como los de lógica difusa (fuzzy analysis).

5 Bibliografía

Brockwell, Peter J., Davis Rihard A., "Time Series: Theory and Methods", Second Edition., Springer-Verlag, 1987.

"Consumer Price Index Manual: Theory and Practice". International Labour Office. Bureau of Statistics, 2004.

Chan Ngai Hang, "Time Series. Applications to Finance", Ed. Wiley-Interscience, 2002.

Domínguez Molina José Ramón (2001). "Formación de Conglomerados y Detección de Observaciones Aberrantes en Series de Tiempo". Universidad de Guanajuato. Tesis para obtener el grado de Maestro en Estadística.

Guerrero Víctor, "Análisis Estadístico de Series de Tiempo Económicas", Ed. Thomson (2003)

"El Índice Nacional de Precios al Consumidor: Características y Actualización de su Base al Año 2002". Banco de México, Junio 2002.

Kaufman Leonard y Rousseeuw Peter (1990). "Finding Groups in Data". An Introduction to Cluster Analysis, Edit. Wiley Interscience.

"Metodología para el Cambio de Base del INPC". Banco de México, Junio 2002.

"Metodología para el Cálculo de la Inflación Subyacente". Banco de México.

Shumway Robert H., Stoffer David S., "Time Series Analysis and Its Applications", Ed. Springer, 2000.

INEGI. "Síntesis Metodológica de la Encuesta Nacional de Ingresos y Gastos en los Hogares". Julio 2003.

Peña Daniel, "Análisis de Series Temporales", Ed. Alianza,(2005).

Ramsey J.O., Silverman B.W., "Functional Data Analysis", Ed. Springer, 1997.

Ramsey J.O., Silverman B.W., "Applied Functional Data Analysis. Methods and Case Studies", Ed. Springer, 2002.

Wei, William W.S., "Time Series Analysis. Univariate and Multivariate Methods". Ed. Addison-Wesley, 1990.

Anexos

Étienne Laspeyres (1834-1913)

La teoría de índices de precios es conocida, sin embargo la contribución de Laspeyres se conoce por su nombre y poco más. La información sobre Laspeyres se limita a escritos de sus contemporáneos como Marshall que leyeron sus artículos o escritores ingleses como Correa Mylan Walsh cuyo libro de 1901 resumía la discusión en torno a los números índices. Las opiniones sobre Laspeyres como economista no eran del todo favorables; Suchumpeter le describió como “un hombre sin la calificación necesaria para llamarse economista” en contraposición a economistas como Jevons, Edgeworth o Fisher que también estaban interesados en los números índice.

Ernst Louis Étienne Laspeyres nació el 28 de noviembre de 1834 en Halle (Saale). Entre 1853 y 1859 estudió leyes y finanzas públicas, consiguiendo su doctorado en Ciencias Políticas y Finanzas Públicas en Heidelberg. Ocupó Plazas en institutos y universidades de Heidelberg, Basel, Riga (en la Escuela Politécnica, un instituto de habla alemana), Dorpat (una universidad alemana en Tartú, la segunda mayor ciudad de Estonia), y Karlsruhe, hasta que finalmente se estableció en Huesen en 1887, donde ocupó una cátedra en Ciencias Políticas hasta su jubilación en 1900.

Durante su estancia en Huesen, Laspeyres fundó el seminario en Estadística de las Ciencias Políticas y fue miembro del Instituto Internacional de Estadísticas (ISI), participando regularmente en sus conferencias. Murió el 4 de agosto de 1913 y fue enterrado en el Alter Friedhof de Huesen junto con otros muchos maestros incluido Röntgen.

Describió su famosa fórmula en un artículo de 1871 “Die Berechnung einer mittleren Waarenpreissteigerung” (*Cálculo del incremento de una media de precios de materias primas*). Ya entonces, tenía ciertas reservas sobre su aplicación general. Por ejemplo, decía que las compras de Hamburgo (una sola ciudad) podría no coincidir con el consumo de todo el país. Sus razones en 1871 para no usar ponderaciones en sus trabajos empíricos eran fundamentalmente prácticas. Laspeyres detalló que sus medidas aritméticas podrían no ser tan correctas como las medias ponderadas usando su fórmula, pero que los pesos eran tan difíciles de obtener para la mayoría de las compras, exceptuando los bienes de consumo, que los errores debidos a su omisión eran menores que los problemas de falta de especificación al no incorporar todos los productores. Así escribió “*Un cálculo totalmente correcto de la inflación es todavía imposible dado el estado actual de las estadísticas de consumo y precios. Por ello, no se debe despreciar un método casi correcto*”.

Laspeyres usó algunos datos de consumo para calcular índices ponderados en su trabajo en 1871 y de nuevo en 1883, pero las cantidades para las ponderaciones estaban disponibles sólo para un conjunto muy reducido de aquellos productos para los cuales el tenía ya los precios (82 frente a 313 en el artículo de 1871), no obstante, permitieron a Laspeyres demostrar su fórmula. Los cálculos elaborados con sus medias simples ofrecieron resultados muy similares a sus medias ponderadas; calculó un índice de incremento medio de los precios comparando la media de precios de la primera mitad de la década de los 50 y los precios de la segunda mitad, su fórmula dio como resultado un 8,8% de incremento para los precios ponderados frente al 9,6% de incremento sin ponderaciones. Sin ponderaciones podía usar 313 precios de productos en sus medias, entendiendo que la información adicional de incluir más productos era más valiosa que el error causado por utilizar una fórmula ligeramente peor.

Laspeyres discutió además los cambios en las características (calidad) comparando los precios de bienes importados con los correspondientes precios de productos nacionales. Además comprobó las diferencias entre los precios oficiales y los de mercado y se preocupó de los problemas derivados de agregar precios de productos heterogéneos (cambios en el tiempo y diferencias territoriales) y sobre ponderaciones de productos más caros. Mitigó los problemas de agregación y unidades indexando los precios a un año base en el que tomaban el valor 100. Con esta transformación, el incremento de los precios con respecto al año base era fácilmente observable en cada producto, el nivel de precios medio para cada periodo era la media de estos precios indexados y la inflación media era el porcentaje de cambio en el nivel medio.

Los trabajos de Laspeyres no son muy conocidos mundialmente ya que se escribieron en alemán. Su fórmula de números índice $\frac{\sum p_1 q_0}{\sum p_0 q_0}$ está muy generalizada y se usa en muchos de los índices de precios más conocidos. Sin embargo, en su época, Laspeyres era más conocido por sus trabajos en finanzas públicas e historia económica. Por todo ello, se podría afirmar, a pesar de la afirmación de Schumpeter, que Laspeyres fue un gran economista que desarrolló las herramientas estadísticas de los índices de precios.

Ponderadores del Índice Nacional de Precios al Consumidor.

Los ponderadores del Índice Nacional de Precios al Consumidor con base en la segunda quincena de junio de 2002, fueron publicados en el Diario Oficial de la Federación del 24 de julio del 2002 y son los siguientes:

ÍNDICE NACIONAL DE PRECIOS AL CONSUMIDOR BASE SEGUNDA QUINCENA DE JUNIO DE 2002

PONDERACIONES

CONCEPTO	NACIONAL
Índice general	100.00000
1. Alimentos, bebidas y tabaco	22.74156
1.1 Alimentos	19.94255
1.1.1 Pan, tortillas y cereales	3.04467
01 Tortillas y derivados del maíz	1.28627
Tortilla de maíz	1.22652
Masa y harinas de maíz	0.3616
Maíz	0.02359
02 Pan	0.98410
Pan dulce	0.52608
Pan blanco	0.21603
Pan de caja	0.18499
Pastelillos y pasteles	0.05700
03 Galletas, pastas y harinas de trigo	0.36316
Pasta para sopa	0.13527
Galletas populares	0.08369
Tortillas de harina de trigo	0.06701
Otras galletas	0.04683
Harinas de trigo	0.03036
04 Arroz y cereales preparados	0.41114
Cereales en hojuelas	0.26113
Arroz	0.15001
1.1.2 Carnes	4.79512
05 Carne de ave	1.20375
Pollo en piezas	1.05177
Pollo entero	0.15197
06 Carne y vísceras de cerdo	0.41332
Pulpa de cerdo	0.19002
Chuletas y manteca de cerdo	0.14242
Lomo	0.05338
Pierna	0.02750
07 Carne y vísceras de res	2.20464
Bistec de res	1.14286
Carne molida de res	0.57050
Chuletas y costillas de res	0.17856
Retazo	0.16479
Cortes especiales de res	0.06623
Hígado de res	0.04902
Otras vísceras de res	0.03268
08 Otros cortes de carne	0.05480
Otros cortes de carne	0.05480
09 Carnes frías, secas y embutidos	0.91861
Jamón	0.44869

CONCEPTO	NACIONAL
Salchichas	0.18002
Chorizo	0.15592
Otros embutidos	0.06229
Carnes secas	0.04324
Tocino	0.02846
1.1.3 Pescados y mariscos	0.56114
10 Pescados y mariscos	0.38673
Otros pescados	0.09341
Camarón	0.08894
Mojarra	0.08689
Otros mariscos	0.04254
Robalo y mero	0.04029
Huachinango	0.03465
11 Pescados y mariscos en conserva	0.17441
Atún y sardina en lata	0.14472
Otros pescados y mariscos en conserva	0.02969
1.1.4 Leche, derivados de leche y huevo	3.53096
12 Leche fresca	1.86485
Leche pasteurizada y fresca	1.86485
13 Leche procesada	0.16355
Leche en polvo	0.11612
Leche evaporada, condensada y maternizada	0.04742
14 Derivados de leche	0.96882
Queso fresco	0.24055
Yogurt	0.20483
Queso Oaxaca o asadero	0.14712
Crema de leche	0.10468
Queso manchego o Chihuahua	0.09869
Otros quesos	0.07393
Helados	0.04832
Queso amarillo	0.02749
Mantequilla	0.02321
15 Huevo	0.53375
Huevo	0.53375
1.1.5 Aceites y grasas vegetales comestibles	0.32098
16 Aceites y grasas vegetales comestibles	0.32098
Aceites y grasas vegetales comestibles	0.32098
1.1.6 Frutas y hortalizas	3.67260
17 Frutas frescas	1.29222
Manzana	0.20787
Plátanos	0.17775
Naranja	0.14212
Aguacate	0.13057
Mango	0.09072
Papaya	0.08775
Limón	0.07816
Otras frutas	0.06578
Uva	0.05676
Melón	0.05059
Sandía	0.03930
Pera	0.03865
Durazno	0.03680
Toronja	0.03130
Piña	0.03101
Guayaba	0.02709
18 Hortalizas frescas	1.63082
Jitomate	0.49532
Papa	0.23429

CONCEPTO	NACIONAL
Cebolla	0.13590
Otras legumbres	0.11626
Tomate verde	0.10892
Calabacita	0.07865
Chile serrano	0.06715
Zanahoria	0.06248
Otros chiles frescos	0.05904
Chile poblano	0.05460
Lechuga y col	0.05315
Chícharo	0.04753
Nopales	0.03738
Chayote	0.03038
Pepino	0.02805
Ejotes	0.02173
19 Legumbres secas	0.37287
Frijol	0.31594
Chile seco	0.03197
Otras legumbres secas	0.02497
20 Frutas y legumbres procesadas	0.37668
Jugos o néctares envasados	0.15528
Chiles procesados	0.07877
Verduras envasadas	0.04519
Puré de tomate y sopas enlatadas	0.03723
Otras conservas de frutas	0.03403
Frutas y legumbres preparadas para bebés	0.02619
1.1.7 Azúcar, café y refrescos envasados	2.16495
21 Azúcar	0.20734
Azúcar	0.20734
22 Café	0.15119
Café soluble	0.11826
Café tostado	0.03293
23 Refrescos envasados y agua purificada	1.80642
Refrescos envasados	1.45187
Agua embotellada	0.35455
1.1.8 Otros alimentos	1.85214
24 Condimentos	0.17587
Mayonesa y mostaza	0.09138
Concentrados de pollo y sal	0.04611
Otros condimentos	0.03838
25 Chocolates y golosinas	0.31721
Papas fritas y similares	0.09672
Concentrados para refrescos	0.08261
Chocolate	0.06008
Dulces, cajetas y miel	0.04674
Gelatina en polvo	0.03107
26 Alimentos cocinados fuera de casa	1.35905
Otros alimentos cocinados	0.52718
Camitas	0.25392
Pollos rostizados	0.22964
Barbacoa o birria	0.22119
Pizzas	0.12712
1.2 Bebidas alcohólicas y tabaco	2.79901
1.2.1 Bebidas alcohólicas	2.19880
27 Cerveza	1.46330
Cerveza	1.46330
28 Vinos y licores	0.73550
Tequila	0.27712
Otros licores	0.13751

CONCEPTO	NACIONAL
Ron	0.12242
Brandy	0.10091
Vino de mesa	0.09754
1.2.2 Tabaco	0.60021
29 Cigarrillos	0.60021
Cigarrillos	0.60021
2. Ropa, calzado y accesorios	5.59224
2.1 Ropa	3.51074
2.1.1 Ropa hombre	1.15854
30 Camisas y ropa interior para hombre	0.41056
Camisas	0.32428
Ropa interior para hombre	0.04719
Calcetines	0.03909
31 Pantalones, trajes y otras prendas para hombre	0.74798
Pantalones para hombre base algodón	0.27729
Trajes	0.24752
Pantalones para hombre otros materiales	0.18914
Otras prendas para hombre	0.03403
2.1.2 Ropa mujer	1.20508
32 Blusas y ropa interior para mujer	0.42347
Blusas para mujer	0.24022
Ropa interior para mujer	0.11972
Medias y pantimedias	0.06353
33 Pantalones para mujer	0.39316
Pantalones para mujer base algodón	0.23139
Pantalones para mujer otros materiales	0.16177
34 Vestidos, faldas y conjuntos para mujer	0.38844
Conjuntos y otras prendas para mujer	0.16434
Vestidos para mujer	0.15739
Faldas para mujer	0.06672
2.1.3 Ropa para niños y bebés	0.66186
35 Ropa para niños	0.50847
Pantalones para niño base algodón	0.15939
Pantalones para niño otros materiales	0.10872
Camisas y playeras para niños	0.09767
Vestidos para niña	0.07148
Ropa interior para niño	0.02474
Ropa interior para niña	0.02400
Calcetines y calcetas	0.02247
36 Ropa para bebés	0.15339
Trajes para bebés	0.11154
Camisetas para bebés	0.04185
2.1.4 Ropa de abrigo y uniformes escolares	0.48525
37 Ropa de abrigo	0.23414
Chamarras y abrigos	0.18761
Sombreros	0.02583
Suéter para niño y niña	0.02069
38 Uniformes escolares	0.25111
Uniformes para niño	0.14853
Uniformes para niña	0.10259
2.2 Calzado	1.61553
2.2.1 Calzado	1.61553
39 Calzado	1.61553
Zapatos tenis	0.45526
Zapatos para mujer	0.38213
Zapatos para hombre	0.36721
Zapatos para niños	0.16453
Otros gastos del calzado	0.12985

CONCEPTO	NACIONAL
Zapatos de material sintético	0.11656
2.3 Accesorios y cuidados del vestido	0.46597
2.3.1 Cuidados del vestido y accesorios	0.46597
40 Limpieza del vestido	0.32813
Servicio de tintorería	0.25523
Servicio de lavandería	0.07290
41 Accesorios personales	0.13784
Bolsas, maletas y cinturones	0.08665
Relojes, joyas y bisutería	0.05119
3. Vivienda	26.40994
3.1 Costo de uso de vivienda	16.68326
3.1.1 Costo de uso de vivienda	16.68326
42 Vivienda alquilada	2.52067
Renta de vivienda	2.52067
43 Vivienda propia	14.16258
Vivienda propia	11.97088
Mantenimiento de vivienda materiales	0.84636
Derechos por el suministro de agua	0.74243
Mantenimiento de vivienda servicios	0.42586
Impuesto predial	0.17706
3.2 Electricidad y combustibles	4.11633
3.2.1 Electricidad y combustibles	4.11633
44 Electricidad	2.27215
Electricidad	2.27215
45 Gas doméstico	1.84418
Gas doméstico	1.84418
3.3 Otros servicios relacionados con la vivienda	5.61035
3.3.1 Otros servicios relacionados con la vivienda	5.61035
46 Servicio telefónico	3.51560
Servicio telefónico local	1.84885
Larga distancia nacional	1.00862
Larga distancia internacional	0.44789
Línea telefónica	0.21023
47 Servicios de la vivienda	2.09476
Servicio doméstico	1.49095
Otros servicios para el hogar	0.60381
4. Muebles, aparatos y accesorios domésticos	4.85674
4.1 Muebles y aparatos domésticos	2.55593
4.1.1 Muebles	1.15951
48 Muebles de cocina	0.33673
Muebles para cocina	0.19600
Antecomedores	0.06613
Estufas	0.05403
Calentadores para agua	0.02057
49 Muebles de madera	0.82278
Salas	0.25281
Comedores	0.15110
Colchones	0.15896
Muebles diversos para el hogar	0.15534
Recámaras	0.10457
4.1.2 Aparatos	1.39642
50 Aparatos eléctricos	0.57783
Refrigeradores	0.14231
Lavadoras de ropa	0.14104
Otros aparatos eléctricos	0.13871
Aparatos de aire acondicionado	0.05629
Ventiladores	0.04941
Planchas eléctricas	0.02598

CONCEPTO	NACIONAL
Licadoras	0.02408
51 Aparatos electrónicos	0.81859
Televisores	0.30017
Computadoras	0.22385
Equipos modulares	0.18381
Reproductores de video	0.07833
Radios y grabadoras	0.03242
4.2 Accesorios y artículos de limpieza para el hogar	2.30081
4.2.1 Accesorios y utensilios	0.77425
52 Accesorios domésticos	0.15844
Focos	0.04413
Cerillos	0.04274
Pilas	0.03960
Velas y veladoras	0.03198
53 Utensilios domésticos	0.32970
Escobas	0.12233
Loza y cristalería	0.08610
Baterías de cocina	0.04807
Utensilios de plástico para el hogar	0.04196
Otros utensilios de cocina	0.03124
54 Accesorios textiles de uso en el hogar	0.28611
Colchas	0.07374
Otros blancos para el hogar	0.06322
Sábanas	0.05342
Cobijas	0.04020
Toallas	0.03146
Cortinas	0.02408
4.2.2 Detergentes y productos similares	1.52655
55 Detergentes y productos similares	1.52655
Detergentes	0.67494
Suavizantes y limpiadores	0.37579
Blanqueadores	0.15609
Jabón para lavar	0.14769
Desodorantes ambientales	0.08754
Plaguicidas	0.08450
Salud y cuidado personal	8.57766
5.1 Salud	4.01097
5.1.1 Medicamentos y aparatos	1.46062
56 Medicamentos	1.20400
Antibióticos	0.26930
Analgésicos	0.13790
Cardiovasculares	0.13270
Nutricionales	0.12774
Anticonceptivos y hormonales	0.12101
Gastrointestinales	0.10519
Expectorantes y descongestivos	0.08996
Otros medicamentos	0.08985
Dermatológicos	0.06398
Antigripales	0.04297
Material de curación	0.02339
57 Aparatos médicos	0.25662
Lentes y otros aparatos	0.18879
Prótesis dentales	0.06784
5.1.2 Servicios médicos	2.55034
58 Servicios médicos	2.55034
Consulta médica	0.83185
Hospitalización general	0.55685
Operación quirúrgica	0.28393

CONCEPTO	NACIONAL
Cuidado dental	0.23834
Estudios médicos de gabinete	0.15925
Hospitalización parto	0.16106
Análisis clínicos	0.14455
Consulta médica durante el embarazo	0.07017
Atención médica durante el parto	0.06218
Análisis clínicos durante el embarazo	0.04217
5.2 Cuidado personal	4.56669
5.2.1 Servicios para el cuidado personal	0.55021
59 Servicios para el cuidado personal	0.55021'
Corte de cabello	0.42853
Sala de belleza	0.12167
5.2.2 Artículos para higiene y cuidado personal	4.01649
60 Artículos para el cuidado personal	2.61730
Productos para el cabello	0.49962
Lociones y perfumes	0.46249
Jabón de tocador	0.38850
Pasta dental	0.35190
Desodorantes personales	0.25702j
Cremas para la piel	0.20767
Navajas y máquinas de afeitar	0.18439
Artículos de maquillaje facial	0.10715
Otros artículos de tocador	0.09464
Otros artículos de maquillaje	0.06391
61 Artículos de papel para higiene personal	1.39919
Papel higiénico	0.54421
Pañales	0.39159
Toallas sanitarias	0.25359
Servilletas de papel	0.13234
Pañuelos desechables	0.07745
6. Transporte	13.40559
6.1 Transporte público	4.72601
6.1.1 Transporte público urbano	3.84067
62 Transporte público urbano	3.84067
Colectivo	1.81805
Autobús urbano	1.31665
Taxi	0.57201
Metro o transporte eléctrico	0.13396
6.1.2 Transporte público foráneo	0.88534
63 Transporte público foráneo	0.88534
Autobús foráneo	0.49757
Transporte aéreo	0.38777
6.2 Transporte por cuenta propia	8.67958
6.2.1 Adquisición de vehículos	3.32589
64 Vehículos automotores y de pedal	3.32589
Automóviles	3.30297
Bicicletas	0.02292
6.2.2 Uso de vehículos	5.35369
65 Gasolinas y aceites lubricantes	3.79999
Gasolina de bajo octanaje	3.19002
Gasolina de alto octanaje	0.46658
Aceites lubricantes	0.14339
66 Refacciones y accesorios automotrices	0.21499
Neumáticos	0.13215
Otras refacciones	0.05691
Acumuladores	0.02593
67 Servicios para el automóvil	1.33871
Seguro de automóvil	0.44353

CONCEPTO	NACIONAL
Tenencia de automóvil	0.29752
Reparación de automóvil	0.17209
Lavado y engrasado de automóvil	0.16648
Mantenimiento de automóvil	0.14815
Cuotas de autopistas	0.07721
Estacionamiento	0.03373
7. Educación y esparcimiento	11.53647
7.1 Educación	7.04579
7.1.1 Educación privada	5.20396
68 Educación privada	5.20396
Universidad	1.66319
Primaria	1.15652
Preparatoria	0.90548
Secundaria	0.56826
Preprimaria	0.27853
Carrera corta	0.23810
Enseñanza adicional	0.20346
Jardín de niños y guardería	0.19042
7.1.2 Artículos de educación	1.84183
69 Libros	1.02378
Libros de texto	0.73300
Otros libros	0.29078
70 Material escolar	0.81805
Cuadernos y carpetas	0.68768
Plumas, lápices y otros	0.13037
7.2 Esparcimiento	4.49068
7.2.1 Servicios de esparcimiento	3.04442
71 Hoteles y gastos turísticos	0.86645
Servicios turísticos en paquete	0.77176
Hoteles	0.09468
72 Otros servicios de esparcimiento	2.17798
Cine	0.48777
Centro nocturno	0.48250
Servicio de televisión por cable o satélite	0.40833
Otras diversiones	0.38655
Club deportivo	0.18846
Espectáculos deportivos	0.12626
Servicio de Internet	0.09811
7.2.2 Artículos de esparcimiento	1.44626
73 Periódicos y revistas	0.36536
Periódicos	0.27226
Revistas	0.09310
74 Otros artículos de esparcimiento	1.08089
Juguetes	0.36563
Discos y casetes	0.35369
Alimento para mascotas	0.11260
Material y aparatos fotográficos	0.09976
Renta de películas	0.07530
Instrumentos musicales y otros	0.04993
Artículos deportivos	0.02399
8. Otros servicios	6.87980
8.1 Otros servicios	6.87980
8.1.1 Otros servicios	6.87980
75 Restaurantes, bares y similares	6.50480
Loncherías	3.57568
Restaurantes	2.33179
Cantinas	0.42010
Cafeterías	0.17723

CONCEPTO		NACIONAL
76	Servicios profesionales	0.18196
	Servicios profesionales	0.18196
77	Servicios diversos	0.19304
	Servicios funerarios	0.12555
	Cuotas de licencias y otros documentos	0.06749

La Encuesta Nacional de Ingresos y Gastos en los Hogares.

La Encuesta Nacional de Ingresos y Gastos en los Hogares surge en el año de 1984, y a partir de 1992, se realiza con una periodicidad de levantamiento de cada dos años, tratando de conservar la comparabilidad del mismo marco conceptual, periodos de referencia, unidades de análisis, cobertura geográfica, instrumentos de captación y los procedimientos operativos utilizados en las encuestas de 1984, 1989, 1992, 1994, 1996, 1998 y 2000.

Los resultados de la ENIGH, son utilizados para distintos fines, entre los cuales se puede mencionar los siguientes: como insumo para la generación de los ponderadores para el cálculo del Índice Nacional de Precios al Consumidor y como insumo para el cálculo de estadísticas sobre los niveles de vida. Ambas estadísticas son frecuentemente utilizadas por especialistas nacionales e internacionales; para el estudio del comportamiento de la economía nacional en el ámbito de la economía de los hogares y comparativos con otros países.

La unidad de observación que la ENIGH considera para su estudio es el “Hogar”, el cual se define como el conjunto de personas unidas o no por lazos de parentesco que residen habitualmente en la misma vivienda particular y se sostienen principalmente de un gasto común para comer; una persona que vive sola también constituye un hogar.

La generación de estadísticas de la ENIGH se basa en la aplicación de un esquema de muestreo probabilístico, a su vez el diseño es polietápico, estratificado y por conglomerados, donde la unidad última de selección es la vivienda y la unidad de observación es el hogar y en consecuencia los resultados obtenidos de la encuesta se generalizan a toda la población.

El marco muestral utilizado es el de propósitos múltiples del INEGI, constituido por la información demográfica y cartográfica obtenida a partir del levantamiento del Censo de Población y Vivienda de 1995.

En este marco de muestreo se excluyen a todas las viviendas colectivas y las de diplomáticos extranjeros, ya que para fines de la encuesta no son objeto de estudio.

El procedimiento de selección para conformar la muestra se realiza de la siguiente manera: Se seleccionan las viviendas, las cuales se distinguen entre dos clases, la vivienda particular y la vivienda colectiva, siendo objeto de la encuesta sólo la primera. Se realiza en forma independiente para cada entidad y estrato, el procedimiento varía dependiendo de la zona.

La selección de la muestra, está calculada para dar estimaciones de los siguientes niveles de desagregación: nivel nacional, localidades de 2,500 y más habitantes y localidades de menos de 2,500 habitantes.

La ENIGH 2000 se levantó en todas las entidades del país, en cada una de éstas se asignó una muestra de 230 a 500 viviendas, distribuidas en 5 estratos, con la cual se tuvo una muestra nacional de 10,000 viviendas.

COBERTURA GEOGRÁFICA	VIVIENDAS EN MUESTRA
AREA URBANA Zona con localidades de 2, 500 y más habitantes	6,400
AREA RURAL Zona con localidades de menos de 2,500 habitantes	3,600
TOTAL NIVEL NACIONAL	10,000

Para llevar a cabo la generación de las estadísticas se realiza una serie de tareas que van desde la planeación hasta la presentación de resultados, para lo cual se toma como base la estructura del INEGI que cuenta con oficinas centrales en Aguascalientes, donde se realiza la planeación general, el diseño técnico y metodológico; 10 oficinas regionales, que realizan labores de planeación regional y seguimiento del trabajo de campo y las oficinas del Instituto en las 32 entidades federativas del país, las cuales realizan el operativo de campo con las fuentes informantes y en algunas de ellas también la crítica-codificación y la captura de los cuestionarios.

El marco conceptual de la ENIGH está basado en las recomendaciones internacionales de la Organización de las Naciones Unidas, la Organización Internacional del Trabajo y está articulado al Sistema de Cuentas Nacionales y a las encuestas de hogares que levanta el INEGI.

Las variables incluidas en el marco conceptual de la ENIGH se agrupan de acuerdo a los temas considerados para la captación de datos y son los siguientes: Identificación de hogares en la vivienda, Equipamiento del hogar, Características de la vivienda, características sociodemográficas de los miembros del hogar, Condición de actividad y características ocupacionales de los miembros del hogar de 12 años y más, Gasto total e Ingreso total.

Los resultados de la encuesta se presentan a través de la publicación de la Encuesta Nacional de Ingresos y Gastos en los Hogares (ENIGH) con información sobre el monto, la estructura y la distribución de los ingresos en efectivo o en especie de los hogares, ya sea que provenga del trabajo, de los intereses de sus propiedades y regalos. Por otra parte, se conoce el destino de sus gastos ya sea en bienes de consumo no duradero (alimentos, bebidas y tabaco) o en bienes de consumo duradero (inmuebles, automóviles, etcétera) para el consumo privado del hogar.

También se presenta una publicación independiente bajo el título de Encuesta Nacional de Ingresos y Gastos en los Hogares (ENIGH). Características de las viviendas por niveles de Ingresos de su(s) hogar(es).

Conglomerado de cada genérico por método de agrupación.

#	Grupo	Código	KMEANS	PAM	AGNES	Descripción
1	Alimentos	AAPTO01	2	1	1	Tortilla de maíz
2	Alimentos	AAPTO02	2	2	1	Masa y harinas de maíz
3	Alimentos	AAPTO03	2	3	2	Maíz
4	Alimentos	AAPPA01	1	3	2	Pan dulce
5	Alimentos	AAPPA02	1	2	2	Pan Blanco
6	Alimentos	AAPPA03	2	2	1	Pan de caja
7	Alimentos	AAPPA04	2	2	1	Pastelillos y pasteles
8	Alimentos	AAPGA01	2	2	1	Pasta para sopa
9	Alimentos	AAPGA02	2	1	2	Galletas populares
10	Alimentos	AAPGA03	2	2	1	Tortillas de harina de trigo
11	Alimentos	AAPGA04	2	2	2	Otras galletas
12	Alimentos	AAPGA05	1	2	2	Harinas de trigo
13	Alimentos	AAPAC01	1	3	2	Cereales en hojuela
14	Alimentos	AAPAC02	2	3	2	Arroz
15	Alimentos	AACCA01	2	3	2	Pollo en piezas
16	Alimentos	AACCA02	1	2	2	Pollo entero
17	Alimentos	AACVC01	2	3	2	Pulpa de cerdo
18	Alimentos	AACVC02	1	2	2	Chuletas y manteca de cerdo
19	Alimentos	AACVC03	1	2	2	Lomo
20	Alimentos	AACVC04	1	2	2	Pierna
21	Alimentos	AACVR01	2	3	1	Bistec de res
22	Alimentos	AACVR02	2	3	1	Carne molida de res
23	Alimentos	AACVR03	2	2	2	Chuletas y costillas de res
24	Alimentos	AACVR04	2	3	1	Retazo
25	Alimentos	AACVR05	2	2	2	Cortes especiales de res
26	Alimentos	AACVR06	2	3	1	Hígado de res
27	Alimentos	AACVR07	2	3	2	Otras vísceras de res
28	Alimentos	AACOC01	2	2	2	Otros cortes de carne
29	Alimentos	AACCF01	2	3	2	Jamón
30	Alimentos	AACCF02	2	2	2	Salchichas
31	Alimentos	AACCF03	2	3	2	Chorizo
32	Alimentos	AACCF04	2	3	2	Otros embutidos
33	Alimentos	AACCF05	2	3	2	Carnes secas
34	Alimentos	AACCF06	2	1	1	Tocino
35	Alimentos	AAMPM01	4	1	1	Otros pescados
36	Alimentos	AAMPM02	2	3	2	Camarón
37	Alimentos	AAMPM03	1	2	2	Mojarra
38	Alimentos	AAMPM04	2	2	2	Otros mariscos
39	Alimentos	AAMPM05	4	1	1	Robalo y mero
40	Alimentos	AAMPM06	4	1	1	Huachinango
41	Alimentos	AAMPC01	1	3	2	Atún y sardina en lata
42	Alimentos	AAMPC02	2	2	2	Otros pescados y mariscos
43	Alimentos	AALLF01	2	1	2	Leche pasteurizada y fresca

#	Grupo	Código	KMEANS	PAM	AGNES	Descripción
44	Alimentos	AALLP01	1	3	2	Leche en polvo
45	Alimentos	AALLP02	2	3	2	Leche evaporada, condensada y maternizada
46	Alimentos	AALDL01	2	3	2	Queso fresco
47	Alimentos	AALDL02	2	2	2	Yoghurt
48	Alimentos	AALDL03	4	1	1	Queso Oaxaca o asadero
49	Alimentos	AALDL04	2	3	2	Crema de leche
50	Alimentos	AALDL05	2	3	2	Queso chihuahua o manchego
51	Alimentos	AALDL06	2	3	2	Otros quesos
52	Alimentos	AALDL07	2	2	2	Helados
53	Alimentos	AALDL08	2	2	2	Queso amarillo
54	Alimentos	AALDL09	2	2	2	Mantequilla
55	Alimentos	AALHU01	2	2	2	Huevo
56	Alimentos	AAAAG01	1	2	2	Aceites y grasas vegetales comestibles
57	Alimentos	AAFFF01	1	2	2	Manzana
58	Alimentos	AAFFF02	2	2	2	Plátanos
59	Alimentos	AAFFF03	1	3	2	Naranja
60	Alimentos	AAFFF04	1	2	2	Aguacate
61	Alimentos	AAFFF05	1	2	2	Mango
62	Alimentos	AAFFF06	2	2	2	Papaya
63	Alimentos	AAFFF07	1	2	2	Limón
64	Alimentos	AAFFF08	1	2	3	Otras frutas
65	Alimentos	AAFFF09	1	2	2	Uva
66	Alimentos	AAFFF10	3	4	3	Melón
67	Alimentos	AAFFF11	3	4	3	Sandía
68	Alimentos	AAFFF12	1	2	2	Pera
69	Alimentos	AAFFF13	4	1	1	Durazno
70	Alimentos	AAFFF14	1	2	2	Toronja
71	Alimentos	AAFFF15	1	2	2	Piña
72	Alimentos	AAFFF16	1	2	2	Guayaba
73	Alimentos	AAFHF01	3	4	3	Jitomate
74	Alimentos	AAFHF02	2	2	2	Papa
75	Alimentos	AAFHF03	1	2	2	Cebolla
76	Alimentos	AAFHF04	2	3	2	Otras legumbres
77	Alimentos	AAFHF05	3	4	3	Tomate verde
78	Alimentos	AAFHF06	3	4	3	Calabacita
79	Alimentos	AAFHF07	3	4	3	Chile serrano
80	Alimentos	AAFHF08	1	2	2	Zanahoria
81	Alimentos	AAFHF09	3	4	3	Otros chiles frescos
82	Alimentos	AAFHF10	3	4	3	Chile poblano
83	Alimentos	AAFHF11	1	2	2	Lechuga y col
84	Alimentos	AAFHF12	3	2	3	Chicharo
85	Alimentos	AAFHF13	3	4	3	Nopales
86	Alimentos	AAFHF14	3	4	3	Chayote
87	Alimentos	AAFHF15	3	4	3	Pepino
88	Alimentos	AAFHF16	3	4	3	Ejotes
89	Alimentos	AAFLS01	1	2	2	Frijol
90	Alimentos	AAFLS02	1	2	2	Chile seco

#	Grupo	Código	KMEANS	PAM	AGNES	Descripción
91	Alimentos	AAFLS03	2	3	2	legumbres secas
92	Alimentos	AAFFL01	2	2	2	Jugos o néctares envasados
93	Alimentos	AAFFL02	2	3	2	Chiles procesados
94	Alimentos	AAFFL03	2	3	2	Verduras envasadas
95	Alimentos	AAFFL04	2	3	2	Puré de tomate y sopas enlatadas
96	Alimentos	AAFFL05	1	3	2	Otras conservas de frutas
97	Alimentos	AAFFL06	4	1	1	Frutas y legumbres preparadas para bebida
98	Alimentos	AAZZU01	2	2	2	Azúcar
99	Alimentos	AAZCA01	1	2	2	Café soluble
100	Alimentos	AAZCA02	1	2	2	Café tostado
101	Alimentos	AAZRE01	4	1	1	Refrescos envasados
102	Alimentos	AAZRE02	2	2	2	Agua embotellada
103	Alimentos	AAOCO01	2	3	2	Mayonesa y mostaza
104	Alimentos	AAOCO02	2	3	2	Concentrados de pollo y sal
105	Alimentos	AAOCO03	2	3	1	Otros condimentos
106	Alimentos	AAOCH01	2	3	2	Papas fritas y similares
107	Alimentos	AAOCH02	2	3	2	Concentrados para refrescos
108	Alimentos	AAOCH03	2	2	2	Chocolate
109	Alimentos	AAOCH04	2	2	2	Dulces, cajetas y miel
110	Alimentos	AAOCH05	2	3	2	Gelatina en polvo
111	Alimentos	AAOAC01	2	1	1	Otros alimentos cocinados
112	Alimentos	AAOAC02	2	3	2	Carnitas
113	Alimentos	AAOAC03	2	3	2	Pollos rostizados
114	Alimentos	AAOAC04	2	1	2	Barbacoa o birria
115	Alimentos	AAOAC05	2	1	1	Pizzas
116	Alimentos	ABBCE01	4	3	2	Cerveza
117	Alimentos	ABBVL01	2	3	1	Tequila
118	Alimentos	ABBVL02	2	3	2	Otros licores
119	Alimentos	ABBVL03	2	3	2	Ron
120	Alimentos	ABBVL04	2	1	1	Brandy
121	Alimentos	ABBVL05	2	3	2	Vino de mesa
122	Alimentos	ABTCI01	1	2	2	Cigarrillos
123	Ropa	RRHC01	2	1	1	Camisas
124	Ropa	RRHC02	2	3	2	Ropa interior para hombre
125	Ropa	RRHC03	2	2	2	Calcetines
126	Ropa	RRHP01	2	3	1	Pantalón hombre base algodón
127	Ropa	RRHP02	4	1	1	Trajes
128	Ropa	RRHP03	4	3	1	Pantalón hombre otros materiales
129	Ropa	RRHP04	2	3	1	Otras prendas para hombre
130	Ropa	RRMB01	4	1	1	Blusas para mujer
131	Ropa	RRMB02	2	3	2	Ropa interior para mujer
132	Ropa	RRMB03	4	1	1	Medias y pantimedias
133	Ropa	RRMP01	2	3	2	Pantalón mujer base algodón
134	Ropa	RRMP02	4	1	1	Pantalón mujer otros materiales
135	Ropa	RRMV01	4	1	1	Conjuntos y otras prendas para mujer
136	Ropa	RRMV02	4	1	1	Vestido para mujer
137	Ropa	RRMV03	4	1	1	Falda para mujer

#	Grupo	Código	KMEANS	PAM	AGNES	Descripción
138	Ropa	RRNN01	2	3	1	Pantalón niño base algodón
139	Ropa	RRNN02	4	1	1	Pantalón niño otros materiales
140	Ropa	RRNN03	4	1	1	Camisas y playeras para niño
141	Ropa	RRNN04	2	3	2	Vestido para niña
142	Ropa	RRNN05	4	1	1	Ropa interior para niño
143	Ropa	RRNN06	4	1	1	Ropa interior para niña
144	Ropa	RRNN07	4	1	1	Calcetines y calcetas
145	Ropa	RRNB01	4	1	1	Traje para bebé
146	Ropa	RRNB02	2	1	1	Camiseta para bebé
147	Ropa	RRAA01	4	1	2	Chamarras y abrigos
148	Ropa	RRAA02	2	3	2	Sombreros
149	Ropa	RRAA03	2	3	1	Sueter para niño y niña
150	Ropa	RRAU01	4	1	1	Uniforme para niño
151	Ropa	RRAU02	4	1	1	Uniforme para niña
152	Ropa	RCCC01	2	3	2	Zapatos tenis
153	Ropa	RCCC02	2	1	1	Zapatos para mujer
154	Ropa	RCCC03	1	2	2	Zapatos para hombre
155	Ropa	RCCC04	2	1	1	Zapatos para niños
156	Ropa	RCCC05	2	2	2	Otros gastos del calzado
157	Ropa	RCCC06	4	1	1	Zapatos de material sintético
158	Ropa	RAAL01	2	2	2	Servicio de tintorería
159	Ropa	RAAL02	3	4	3	Servicio de lavandería
160	Ropa	RAAA01	4	1	1	Bolsas, maletas y cinturones
161	Ropa	RAAA02	2	2	1	Relojes, joyas y bisutería
162	Vivienda	VCCA01	1	2	2	Vivienda alquilada Renta de vivienda
163	Vivienda	VCCP01	1	2	2	Vivienda propia Vivienda propia
164	Vivienda	VCCP02	3	2	3	Vivienda propia Mantenimiento de vivienda materiales
165	Vivienda	VCCP03	2	2	1	Vivienda propia Derechos por el suministro de agua
166	Vivienda	VCCP04	1	2	2	Vivienda propia Mantenimiento de vivienda servicios
167	Vivienda	VCCP05	1	2	2	Vivienda propia Impuesto predial
168	Vivienda	VEEE01	4	1	1	Electricidad Electricidad
169	Vivienda	VEEG01	2	2	2	Gas doméstico Gas doméstico
170	Vivienda	VOOT01	2	1	1	Servicio telefónico Servicio telefónico local
171	Vivienda	VOOT02	2	2	2	Servicio telefónico Larga distancia nacional
172	Vivienda	VOOT03	3	4	3	Servicio telefónico Larga distancia internacional
173	Vivienda	VOOT04	4	1	1	Servicio telefónico Línea telefónica
174	Vivienda	VOOD01	1	2	2	Servicio doméstico Servicio doméstico
175	Vivienda	VOOO02	1	2	2	Servicios de la vivienda Otros servicios para el hogar
176	Muebles	MADMC01	2	2	2	Muebles para cocina
177	Muebles	MADMC02	2	3	1	Antecomedores
178	Muebles	MADMC03	2	3	1	Estufas
179	Muebles	MADMC04	2	3	2	Calentadores para agua
180	Muebles	MADMC01	2	3	1	Salas
181	Muebles	MADMC02	2	1	1	Comedores
182	Muebles	MADMC03	4	1	1	Colchones
183	Muebles	MADMC04	3	2	3	Muebles diversos para el hogar
184	Muebles	MADMC05	4	1	1	Recámaras

#	Grupo	Código	KMEANS	PAM	AGNES	Descripción
185	Muebles	MADAE01	2	3	1	Refrigeradores
186	Muebles	MADAE02	2	2	2	Lavadoras de ropa
187	Muebles	MADAE03	4	1	2	Otros aparatos eléctricos
188	Muebles	MADAE04	3	4	3	Aparatos de aire acondicionado
189	Muebles	MADAE05	2	3	2	Ventiladores
190	Muebles	MADAE06	2	3	2	Planchas eléctricas
191	Muebles	MADAE07	2	3	2	Licuadoras
192	Muebles	MADAE01	3	2	1	Televisores
193	Muebles	MADAE02	2	1	1	Computadoras
194	Muebles	MADAE03	2	3	2	Equipos modulares
195	Muebles	MADAE04	4	1	1	Reproductores de video
196	Muebles	MADAE05	2	3	2	Radios y grabadoras
197	Muebles	MLHUD01	2	1	1	Focos
198	Muebles	MLHUD02	2	2	1	Cerillos
199	Muebles	MLHUD03	2	2	2	Pilas
200	Muebles	MLHUD04	2	2	2	Velas y veladoras
201	Muebles	MLHUD01	2	2	1	Escobas
202	Muebles	MLHUD02	2	3	1	Loza y cristalería
203	Muebles	MLHUD03	1	2	2	Baterías de cocina
204	Muebles	MLHUD04	2	3	1	Utensilios de plástico p/el hogar
205	Muebles	MLHUD05	2	1	1	Otros utensilios de cocina
206	Muebles	MLHUT01	4	1	1	Colchas
207	Muebles	MLHUT02	4	1	1	Otros blancos para el hogar
208	Muebles	MLHUT03	2	1	1	Sábanas
209	Muebles	MLHUT04	2	3	1	Cobijas
210	Muebles	MLHUT05	2	3	1	Toallas
211	Muebles	MLHUT06	2	1	2	Cortinas
212	Muebles	MALDS01	2	2	2	Detergentes
213	Muebles	MALDS02	4	1	1	Suavizantes y limpiadores
214	Muebles	MALDS03	3	4	3	Blanqueadores
215	Muebles	MALDS04	1	3	2	Jabón para lavar
216	Muebles	MALDS05	2	3	1	Desodorantes ambientales
217	Muebles	MALDS06	2	3	1	Plaguicidas
218	Salud	SSAMAM01	2	2	2	Antibióticos
219	Salud	SSAMAM02	2	3	2	Analgésicos
220	Salud	SSAMAM03	2	3	1	Cardiovasculares
221	Salud	SSAMAM04	2	3	2	Nutricionales
222	Salud	SSAMAM05	2	3	2	Anticonceptivos y hormonales
223	Salud	SSAMAM06	2	3	2	Gastrointestinales
224	Salud	SSAMAM07	1	2	2	Expectorantes y descongestivos
225	Salud	SSAMAM08	2	1	1	Otros medicamentos
226	Salud	SSAMAM09	3	2	1	Dermatológicos
227	Salud	SSAMAM10	2	3	1	Antigripales
228	Salud	SSAMAM11	4	1	1	Material de curación
229	Salud	SSAMAA01	4	1	1	Lentes y otros aparatos
230	Salud	SSAMAA02	3	4	3	Prótesis dentales
231	Salud	SSCSMS01	2	2	2	Consulta médica

#	Grupo	Código	KMEANS	PAM	AGNES	Descripción
232	Salud	SSCSMS02	3	2	3	Hospitalización general
233	Salud	SSCSMS03	2	2	2	Operación quirúrgica
234	Salud	SSCSMS04	2	3	1	Cuidado dental
235	Salud	SSCSMS05	3	4	3	Estudios médicos de gabinete
236	Salud	SSCSMS06	3	4	3	Hospitalización parto
237	Salud	SSCSMS07	4	2	2	Análisis clínicos
238	Salud	SSCSMS08	4	1	1	Consulta médica durante el embarazo
239	Salud	SSCSMS09	2	3	2	Atención médica durante el parto
240	Salud	SSCSMS10	4	3	1	Análisis clínicos durante el embarazo
241	Salud	SCPSCS01	2	2	2	Corte de cabello
242	Salud	SCPSCS02	2	3	2	Sala de belleza
243	Salud	SCPAHA01	2	3	2	Productos para el cabello
244	Salud	SCPAHA02	2	1	1	Lociones y perfumes
245	Salud	SCPAHA03	2	2	2	Jabón de tocador
246	Salud	SCPAHA04	2	2	2	Pasta dental
247	Salud	SCPAHA05	2	3	2	Desodorantes personales
248	Salud	SCPAHA06	2	3	1	Cremas para la piel
249	Salud	SCPAHA07	2	3	2	Navajas y máquinas de afeitar
250	Salud	SCPAHA08	3	4	3	Artículos de maquillaje facial
251	Salud	SCPAHA09	4	1	1	Otros artículos de tocador
252	Salud	SCPAHA10	3	1	1	Otros artículos de maquillaje
253	Salud	SCPAHH01	2	1	2	Papel higiénico
254	Salud	SCPAHH02	2	3	1	Pañales
255	Salud	SCPAHH03	2	3	1	Toallas sanitarias
256	Salud	SCPAHH04	2	1	2	Servilletas de papel
257	Salud	SCPAHH05	4	1	1	Pañuelos desechables
258	Transporte	TTPUU01	2	2	1	Transporte público urbano Colectivo
259	Transporte	TTPUU02	2	2	1	Transporte público urbano Autobús urbano
260	Transporte	TTPUU03	2	2	1	Transporte público urbano Taxi
261	Transporte	TTPUU04	2	2	2	Transporte público urbano Metro o transporte eléctrico
262	Transporte	TTPFF01	2	1	2	Transporte público foráneo Autobús foráneo
263	Transporte	TTPFF02	1	2	1	Transporte público foráneo Transporte aéreo
264	Transporte	TCPAV01	1	2	2	Vehículos automotores y de pedal Automóviles
265	Transporte	TCPAV02	2	3	1	Vehículos automotores y de pedal Bicicletas
266	Transporte	TCPVG01	3	1	1	Gasolinas y aceites lubricantes Gasolina de bajo octanaje
267	Transporte	TCPVG02	4	1	1	Gasolinas y aceites lubricantes Gasolina de alto octanaje
268	Transporte	TCPVG03	2	2	1	Gasolina y aceites lubricantes Aceites lubricantes
269	Transporte	TCPVR01	2	3	1	Refacciones y accesorios automotrices Neumáticos
270	Transporte	TCPVR02	2	2	1	Refacciones y accesorios automotrices Otras refacciones
271	Transporte	TCPVR03	2	1	1	Refacciones y accesorios automotrices Acumuladores
272	Transporte	TCPVS01	2	3	2	Servicios para automóvil Seguro de automóvil
273	Transporte	TCPVS02	1	2	2	Servicios para automóvil Tenencia de automóvil
274	Transporte	TCPVS03	3	4	3	Servicios para el automóvil Reparación de automóvil
275	Transporte	TCPVS04	3	4	3	Servicios para el automóvil Lavado y engrasado de automóvil
276	Transporte	TCPVS05	2	1	1	Servicios para automóvil Mantenimiento de automóvil
277	Transporte	TCPVS06	4	1	1	Servicios para automóvil Cuotas de autopista
278	Transporte	TCPVS07	2	1	1	Servicios para automóvil Estacionamiento

#	Grupo	Código	KMEANS	PAM	AGNES	Descripción
279	Educación	EEPP01	4	1	1	Universidad
280	Educación	EEPP02	4	1	4	Primaria
281	Educación	EEPP03	4	1	4	Preparatoria
282	Educación	EEPP04	4	1	4	Secundaria
283	Educación	EEPP05	4	1	1	Preprimaria
284	Educación	EEPP06	4	1	1	Carrera corta
285	Educación	EEPP07	3	4	1	Enseñanza adicional
286	Educación	EEPP08	4	1	4	Jardín de niños y guardería
287	Educación	EEAL01	1	2	2	Libros de texto
288	Educación	EEAL02	2	3	1	Otros libros
289	Educación	EEAM01	1	2	2	Cuadernos y carpetas
290	Educación	EEAM02	2	3	2	Plumas, lápices y otros
291	Educación	ESSH01	3	4	3	Servicios turísticos en paquete
292	Educación	ESSH02	4	1	1	Hoteles
293	Educación	ESSO01	2	2	1	Cine
294	Educación	ESSO02	2	2	2	Centro nocturno
295	Educación	ESSO03	4	2	1	Serv de televisión por cable o satélite
296	Educación	ESSO04	4	1	1	Otras diversiones
297	Educación	ESSO05	4	1	1	Club deportivo
298	Educación	ESSO06	2	2	2	Espectáculos deportivos
299	Educación	ESSO07	3	2	3	Servicio de Internet
300	Educación	ESAR01	2	2	2	Periódicos
301	Educación	ESAT02	2	2	2	Revistas
302	Educación	ESAO01	2	2	2	Juguetes
303	Educación	ESAO02	2	3	1	Discos y casetes
304	Educación	ESAO03	2	3	1	Alimento para mascotas
305	Educación	ESAO04	2	2	2	Material y aparatos fotográficos
306	Educación	ESAO05	3	4	3	Renta de películas
307	Educación	ESAO06	4	1	1	Instrumentos musicales y otros
308	Educación	ESAO07	2	1	1	Artículos deportivos
309	Otros	OOOSRB01	2	3	2	Loncherías
310	Otros	OOOSRB02	2	3	2	Restaurantes
311	Otros	OOOSRB03	1	2	2	Cantinas
312	Otros	OOOSRB04	2	2	2	Cafeterías
313	Otros	OOOSSP01	3	4	3	Servicios profesionales
314	Otros	OOOSSD01	4	1	2	Servicios funerarios
315	Otros	OOOSSD02	4	1	1	Licencias y otros doctos

Causalidad

En la sección 3.3 del libro “Time Series: Applications to Finance” de Ngai Hang Chang, en sus páginas 25 a 30 habla acerca de los modelos autorregresivos, profundizando sobre la dualidad que existe entre estacionaridad y causalidad y lo tomamos para este anexo de manera íntegra:

Modelos Autorregresivos

Una clase de modelos ampliamente utilizados son los modelos autorregresivos (AR). Un modelo autorregresivo intuitivamente recuerda al modelo tradicional de regresión.

Cuando reemplazamos al predictor en el clásico modelo de regresión por los valores pasados de las series de tiempo, tenemos un modelo AR.

Es por ello razonable esperar que la mayoría de los resultados estadísticos del modelo clásico de regresión puedan ser generalizados al caso AR con pequeñas modificaciones.

Este es en verdad el caso y es por esta razón que los modelos AR se han convertido en uno de los más utilizados modelos lineales de series de tiempo.

Formalmente un modelo $AR_{(p)}$, Y_t puede escribirse como $\phi(B)Y_t = Z_t$, donde

$$\phi(B) = (1 - \phi_1 B - \dots - \phi_p B^p),$$

$$BY_t = Y_{t-1}, \text{ tal que}$$

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + Z_t$$

Formalmente tenemos las siguientes definiciones:

Definición (1).

Y_t se dice que es un proceso $AR_{(p)}$ si :

- I. Y_t es estacionario.
- II. Y_t satisface $\phi(B)Y_t = Z_t$ para toda t .

Definición (2).

Y_t se dice que es un proceso $AR_{(p)}$ con media μ si $Y_t - \mu$ es un proceso $AR_{(p)}$.

Dualidad entre Causalidad y Estacionaridad

Parece haber confusión entre las nociones de estacionaridad y causalidad para modelos AR ($ARMA$ en general).

La pregunta es: ¿Es verdad que siempre existe un $AR_{(p)}$?

Para responder a esta pregunta, consideremos el caso $AR_{(1)}$, donde

$$Y_t = \phi Y_{t-1} + Z_t, \quad Z_t \square WN(0, \sigma^2)$$

Iterando esta ecuación,

$$Y_t = Z_t + \phi Z_{t-1} + \dots + \phi^{k+1} Z_{t-k-1}$$

Nos lleva a la siguiente pregunta:

¿Podemos encontrar un proceso estacionario que satisfaga $Y_t = \phi Y_{t-1} + Z_t$?

Primero, si tal proceso Y_t existe, ¿cómo se verá?

Dado que Y_t satisface $Y_t = \phi Y_{t-1} + Z_t$, debe tener la siguiente forma:

$$Y_t = \sum_{i=0}^k \phi^i Z_{t-i} + \phi^{k+1} Y_{t-k-1}$$

Asumamos de momento que $|\phi| < 1$. Dado que Y_t es estacionario, $EY_t^2 = a$ una constante para toda t . En particular denotamos $\|Y_t\|^2 = EY_t^2$, tenemos

$$\|Y_t - \sum_{j=0}^k \phi^j Z_{t-j}\|^2 = \phi^{2k+2} \|Y_{t-k-1}\|^2 \rightarrow 0 \text{ cuando } k \rightarrow \infty$$

Así $Y_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$ en L^2 .

Para este proceso nuevamente definido $Y_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$, tenemos las siguientes propiedades.

- I. Y_t satisface $Y_t = \phi Y_{t-1} + Z_t$, para toda t .
- II. $EY_t = 0$, $\text{var} Y_t = \sigma^2 / (1 - \phi^2)$.
- III. $\text{cov}(Y_t, Y_{t+k}) = \text{cov}\left(\sum_{j=0}^{\infty} \phi^j Z_{t-j}, \sum_{l=0}^{\infty} \phi^l Z_{t+k-l}\right) = \sigma^2 \sum_{j=0}^{\infty} \phi^{2j+k} = \sigma^2 \phi^k / (1 - \phi^2)$.

Por lo tanto, el recientemente definido Y_t es estacionario. La respuesta a la pregunta ¿Podemos encontrar un proceso estacionario que satisfaga $Y_t = \phi Y_{t-1} + Z_t$?

Es que existe un proceso estacionario $AR_{(1)}$, Y_t que satisface $Y_t = \phi Y_{t-1} + Z_t$.

Pregunta 2. ¿Qué pasa con el supuesto de que $|\phi| > 1$?

Este supuesto es inmaterial, dado que no es necesario dado que hemos establecido la forma correcta de Y_t . Aún cuando $|\phi| > 1$, el proceso Y_t no continúa siendo convergente, podemos reescribir $Y_t = \phi Y_{t-1} + Z_t$, como sigue. Dado que $Y_{t+1} = \phi Y_t + Z_{t+1}$, dividiendo ambos lados entre ϕ , tenemos

$$Y_t = \frac{1}{\phi} Y_{t+1} - \frac{1}{\phi} Z_{t+1}$$

Remplazamos t por $t+1$ llegamos a $Y_{t+1} = \frac{Y_{t+2} - Z_{t+2}}{\phi}$

Sustituyendo e iterando adelante sobre t , tenemos

$$\begin{aligned} Y_t &= -\frac{1}{\phi} Z_{t+1} + \frac{1}{\phi} Y_{t+1} \\ &= -\frac{1}{\phi} Z_{t+1} + \frac{1}{\phi} \left(\frac{1}{\phi} Y_{t+2} - \frac{1}{\phi} Z_{t+2} \right) \\ &= \dots = -\frac{1}{\phi} Z_{t+1} - \frac{1}{\phi^2} Z_{t+2} - \dots + \frac{1}{\phi^{k+1}} Y_{t+k+1} \end{aligned}$$

Por lo tanto, $Y_t = -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j}$ es la solución estacionaria de $Y_t = \phi Y_{t-1} + Z_t$. Este proceso Z_t es de cualquier modo no natural dado que depende de valores futuros de Y_t , los cuales son inobservables. Hemos impuesto una condición a futuro.

Condición Causal.

Un proceso útil AR debería depender únicamente de su historia, es decir, $Z_k : k = -\infty, \dots, t$ no de valores futuros. Formalmente, si existe una secuencia de constantes ψ_i con $\sum_{i=0}^{\infty} |\psi_i| < \infty$ tal que $Y_t = \sum_{i=0}^{\infty} \psi_i Z_{t-i}$, el proceso Y_t se dice que es causal (estacionario según algunos textos).

Pregunta 3. ¿Será la condición de causalidad demasiado restrictiva?

Sea Y_t la solución estacionaria del modelo no causal $AR_{(1)}$, $Y_t = \phi Y_{t-1} + Z_t$, $|\phi| > 1$. Sabemos que $Y_t = \sum_{j=1}^{\infty} -\phi^{-j} Z_{t+j}$ es una solución estacionaria de $Y_t = \phi Y_{t-1} + Z_t$, aunque no causal. Sin embargo, puede mostrarse que Y_t también satisface

$$Y_t = \phi^{-1} Y_{t-1} + \bar{Z}_t, \quad \bar{Z}_t \stackrel{i.i.d.}{\sim} 0, \bar{\sigma}^2$$

Para un nuevamente definido ruido $\bar{Z}_t \stackrel{i.i.d.}{\sim} 0, \bar{\sigma}^2$ consecuentemente, sin pérdida de generalidad, podemos considerar simplemente el proceso causal. Para el caso $AR_{(1)}$ la expresión causal es $Y_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$.

Estacionariedad Asintótica

Existe otra sutileza acerca del proceso $AR_{(1)}$. Suponga que el proceso no regresa al pasado remoto pero parte de un valor inicial Y_0 . Entonces

$$Y_t = Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \dots + \phi^{t-1} Z_1 + \phi^t Y_0$$

Si Y_0 es una variable aleatoria que es independiente de la secuencia Z_t tal que $EY_0 \neq 0$, entonces $EY_t = \phi^t EY_0$. En este caso, el proceso Y_t no es siquiera estacionario. Para dar la vuelta a este problema, supongamos que Y_0 es independiente de la secuencia Z_t con $EY_0 = 0$. Consideramos la varianza de Y_t :

$$\text{var } Y_t = \sigma^2 (1 + \phi^2 + \dots + \phi^{2(t-1)}) + \phi^{2t} \text{var } Y_0$$

$$= \frac{\sigma^2 (1 - \phi^{2t})}{1 - \phi^2} + \phi^{2t} \text{var } Y_0$$

$$\rightarrow \frac{\sigma^2}{(1 - \phi^2)} \text{ como } t \rightarrow \infty, |\phi| < 1$$

Incluso cuando $EY_0 = 0$, el proceso Y_t no es estacionario dado que la varianza esta cambiando en el tiempo. Este únicamente es estacionario cuando t es grande (es decir, es estacionario en un sentido asintótico). Con valores fijos iniciales, el modelo AR no es estacionario en un sentido riguroso; este únicamente es asintóticamente estacionario. Es por esta razón que cuando un modelo AR es simulado, tenemos que descartar el inicio de los datos a fin de que el efecto de Y_0 sea insignificante.

Teorema de Causalidad

Un proceso se dice que es causal si puede ser expresado como valores presentes y pasados de procesos de ruido $Z_t, Z_{t-1}, Z_{t-2}, \dots$. Formalmente, tenemos la siguiente definición:

Un proceso Y_t se dice que es causal si existe una secuencia de constantes ψ_i 's tal que

$$Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \text{ con } \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

Para un modelo $AR_{(p)}$, $\phi(B)Y_t = Z_t$, escribimos

$$Y_t = \phi^{-1}(B)Z_t = \psi(B)Z_t = \sum_{i=0}^{\infty} \psi_i Z_{t-i}, \text{ donde } \psi_0 = 1. \text{ Bajo qué condiciones será esta expresión}$$

bien definida (es decir, ¿bajo qué condiciones el modelo $AR_{(p)}$ será causal?). La respuesta está en el siguiente teorema.

Teorema

Un proceso $AR_{(p)}$ será causal si las raíces del polinomio característico $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$ caen todas fuera del círculo unitario (es decir, $z: \phi(z) = 0 \subseteq z: |z| > 1$).

Estructura de covarianza de modelos AR .

Dado un modelo causal $AR_{(p)}$, tenemos

$$\gamma(k) = E Y_t Y_{t+k} = E \left(\sum_{i=0}^{\infty} \psi_i Z_{t-i} \right) \left(\sum_{l=0}^{\infty} \psi_l Z_{t+k-1-l} \right) = \sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{k+1-i}$$

Ejemplo

Para un modelo $AR_{(1)}$, $Y_t = \phi Y_{t-1} + Z_t$, tenemos $\psi_i = \phi^i$, tal que $\gamma(k) = \frac{\sigma^2 \phi^k}{(1-\phi^2)}$ y $\rho(k) = \phi^k$.

Aunque podemos usar la ecuación

$$\gamma(k) = E Y_t Y_{t+k} = E \left(\sum_{i=0}^{\infty} \psi_i Z_{t-i} \right) \left(\sum_{l=0}^{\infty} \psi_l Z_{t+k-1-l} \right) = \sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{k+1-i}$$

Para encontrar la función de covarianza de un modelo $AR_{(p)}$ dado, esta requiere resolver ψ 's en términos de ϕ 's y en ocasiones es difícil encontrar una fórmula explícita. Podemos reducir esta dificultad con la siguiente observación.

Sea Y_t un modelo causal estacionario $AR_{(p)}$. Multiplicamos Y_t por Y_{t-k} :

$$Y_t Y_{t-k} = \phi_1 Y_{t-1} Y_{t-k} + \dots + \phi_p Y_{t-p} Y_{t-k} + Z_t Y_{t-k}$$

Tomamos la esperanza y obtenemos.

$$\phi_1 \gamma(k-1)$$

Dividiendo entre $\gamma(0)$, obtenemos

$$\rho(k) = \phi_1 \rho(k-1) + \dots + \phi_p \rho(k-p) \text{ para toda } k.$$

Hemos llegado a un conjunto de ecuaciones de diferencia, las ecuaciones de Yule-Walker, cuyas soluciones generales están dadas por:

$$\rho(k) = A_1 \pi_1^{-|k|} + \dots + A_p \pi_p^{-|k|},$$

Donde π_i son las soluciones de la correspondiente ecuación característica de un proceso $AR_{(p)}$,

$$1 - \phi_1 z^{p-1} - \dots - \phi_p z^p = 0.$$

Error de pronóstico.

El capítulo cinco del libro “Time Series Analysis: Univariate and Multivariate Methods” de William W. S. Wei, en sus páginas 86 a 91 habla acerca del error de pronóstico y lo tomamos para este anexo de manera íntegra:

Pronóstico

La Incertidumbre es un hecho en la vida de los individuos y las organizaciones. El pronóstico es esencial para la planeación y control de operación de muchas áreas. Desarrollaremos el error cuadrático medio mínimo para el pronóstico de series estacionarias y no estacionarias. Estos modelos pueden ser utilizados para actualizar pronósticos cuando se dispone de nueva información. También discutiremos las implicaciones del modelo de series de tiempo construido en términos de su eventual función de pronóstico.

Uno de los objetivos más importantes en el análisis de series de tiempo es el pronóstico de valores futuros. Incluso si el propósito final del modelado de series de tiempo es para el control de un sistema, su operación está usualmente basada en el pronóstico. El termino pronosticar es usado con mayor frecuencia en la literatura reciente que el término predicción. De cualquier modo la mayoría de los resultados de pronóstico se derivan de la teoría general de predicción lineal desarrollada por Kolmogorov (1939, 1941), Weiner(1949), Kalman (1960), Yaglom(1962) y Whittle(1983), entre otros.

Consideremos el modelo general ARIMA(p,d,q).

$$\phi(B)(1-B)^d Z_t = \theta(B)a_t$$

Donde

$$\phi(B) = (1 - \phi_1 B - \dots - \phi_p B^p)$$

$$\theta(B) = (1 - \theta_1 B - \dots - \theta_q B^q)$$

y las series a_t es un proceso de ruido blanco Gaussiano $N(0, \sigma_a^2)$.

El parámetro de tendencia determinístico θ_0 se omite por simplicidad, sin pérdida de generalidad.

Discutiremos el mínimo error cuadrático medio de pronóstico para ambos casos, cuando $d = 0$ y cuando $d \neq 0$.

Mínimo error cuadrático medio de pronóstico para modelos ARMA.

Cuando $d = 0$ y $\mu = 0$ tenemos el modelo estacionario ARMA:

$$\phi(B)Z_t = \theta(B)a_t$$

Dado que el modelo es estacionario podemos reescribirlo en una representación de promedios móviles.

$$Z_t = \psi(B)a_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots$$

Donde

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j = \frac{\theta(B)}{\phi(B)}$$

y $\psi_0 = 1$.

Para $t = n+l$ tenemos

$$Z_{n+l} = \sum_{j=0}^{\infty} \psi_j a_{n+l-j} \dots\dots\dots(1)$$

Suponemos que en el tiempo $t = n$, tenemos las observaciones $Z_n, Z_{n-1}, Z_{n-2}, \dots$ y deseamos pronosticar l pasos adelante el valor futuro Z_{n+l} como combinación lineal de las observaciones $Z_n, Z_{n-1}, Z_{n-2}, \dots$

Dado que Z_t para $t = n, n-1, n-2, \dots$ puede escribirse como $Z_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots$.

Podemos escribir el mínimo error cuadrático medio de pronóstico $\hat{Z}_n(l)$ de Z_{n+l} como

$$\hat{Z}_n(l) = \psi_l^* a_n + \psi_{l+1}^* a_{n-1} + \psi_{l+2}^* a_{n-2} + \dots$$

Donde las ψ_j^* serán determinadas.

La media del error cuadrático de es:

$$E(Z_{n+l} - \hat{Z}_n(l))^2 = \sigma_a^2 \sum_{j=0}^{l-1} \psi_j^2 + \sigma_a^2 \sum_{j=0}^{\infty} [\psi_{l+j} - \psi_{l+j}^*]^2$$

El cuál puede verse fácilmente que se minimiza cuando $\psi_{l+j}^* = \psi_{l+j}$, de ahí

$$\hat{Z}_n(l) = \psi_l a_n + \psi_{l+1} a_{n-1} + \psi_{l+2} a_{n-2} + \dots$$

Utilizando (1) y el hecho de que

$$E(a_{n+j} | Z_n, Z_{n-1}, \dots) = \begin{cases} 0, & j > 0 \\ a_{n+j}, & j \leq 0 \end{cases}$$

Tenemos

$$E(Z_{n+l} | Z_n, Z_{n-1}, \dots) = \psi_l a_n + \psi_{l+1} a_{n-1} + \psi_{l+2} a_{n-2} + \dots$$

Así el mínimo error cuadrático medio de pronóstico de Z_{n+l} está dado por la esperanza condicional. Esto es:

$$\hat{Z}_n(l) = E(Z_{n+l} | Z_n, Z_{n-1}, \dots)$$

El error de pronóstico es:

$$e_n(l) = Z_{n+l} - \hat{Z}_n(l) = \sum_{j=0}^{l-1} \psi_j a_{n+l-j}$$

Dado que $E(e_n(l) | Z_t, t \leq n) = 0$ el pronóstico es insesgado con varianza

$$Var(e_n(l)) = \sigma_a^2 \sum_{j=0}^{l-1} \psi_j^2$$

Para un proceso Normal con $(1 - \alpha)$ 100% los límites del pronóstico son:

$$\hat{Z}_n(l) \pm N_{\alpha/2} \left[1 + \sum_{j=1}^{l-1} \psi_j^2 \right]^{1/2} \sigma_a$$

Donde $N_{\alpha/2}$ es la desviación normal estándar tal que $P(N > N_{\alpha/2}) = \alpha/2$.

El error de pronóstico $e_n(l)$ es una combinación lineal de valores futuros aleatorios entrando al sistema después del tiempo n . Especialmente, el error de pronóstico en el tiempo 1 es:

$$e_n(1) = Z_{n+1} - \hat{Z}_n(1) = a_{n+1}$$

Entonces los errores de pronóstico de un tiempo son independientes. Esto implica que $\hat{Z}_n(1)$ es en verdad el mejor pronóstico de Z_{n+1} . De otra manera, si los errores de pronóstico al tiempo uno están correlacionados, entonces podemos calcular el pronóstico \hat{a}_{n+1} de a_{n+1} de los errores disponibles $a_n, a_{n-1}, a_{n-2}, \dots$ y de ahí mejorar el pronóstico de Z_{n+1} simplemente utilizando $\hat{Z}_n(1) + \hat{a}_{n+1}$ como el pronóstico. De cualquier forma los errores de pronóstico de tiempos más alejados están correlacionados. Esto es cierto para los errores de pronóstico:

$$e_n(l) = Z_{n+l} - \hat{Z}_n(l) = a_{n+l} + \psi_l a_{n+l-1} + \dots + \psi_{l-1} a_{n+1}$$

Y

$$e_{n-j}(l) = Z_{n+l-j} - \hat{Z}_{n-j}(l) = a_{n+l-j} + \psi_l a_{n+l-j-1} + \dots + \psi_{l-1} a_{n-j+1}$$

Los cuales se producen en el mismo tiempo de retraso l pero con diferentes orígenes n y $n-j$ para $j < l$. Esto también es cierto para errores de pronóstico de diferentes tiempos de retraso generados a partir del mismo origen. Por ejemplo:

$$\text{Cov } e_n(2), e_n(1) = E \left[\begin{matrix} a_{n+2} + \psi_1 a_{n+1} & a_{n+1} \end{matrix} \right] = \psi_1 \sigma_a^2$$

Mínimo error cuadrático medio de pronóstico para modelos ARIMA.

Ahora consideremos el modelo general no estacionario ARIMA(p,d,q) con $d \neq 0$, es decir:

$$\phi(B)(1-B)^d Z_t = \theta(B)a_t$$

Donde

$$\phi(B) = (1 - \phi_1 B - \dots - \phi_p B^p)$$
 es un operador estacionario AR y

$$(1 - \theta_1 B - \dots - \theta_q B^q)$$
 es un operador MA invertible.

Es interesante notar que aunque para este proceso la media y los momentos de segundo orden tales como las funciones de varianza y autocovarianza varían en el tiempo, la completa evolución del proceso está completamente determinada por un número finito de parámetros fijos. De ahí podemos ver el pronóstico del proceso como la estimación de una función de estos parámetros y obtener el mínimo error cuadrático medio de pronóstico utilizando un argumento Bayesiano. Esto es bien conocido que utilizando la aproximación respecto al criterio del error cuadrático medio, el cuál corresponde a una función cuadrática de pérdida, cuando la serie es conocida hasta el tiempo n el pronóstico óptimo de Z_{n+l} está dado por su esperanza condicional

$E(Z_{n+l} | Z_n, Z_{n-1}, \dots)$. El mínimo error cuadrático medio de pronóstico para el modelo estacionario ARMA que discutimos es un caso especial del pronóstico para el modelo ARIMA (p, d, q) con $d = 0$.

Para obtener la varianza del pronóstico para el modelo general ARIMA reescribimos el modelo en el tiempo $t+l$ en una representación AR que existe porque el modelo es invertible. Así,

$$\pi(B)Z_{t+l} = a_{t+l}$$

Donde

$$\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j = \frac{\phi(B)(1-B)^d}{\theta(B)}$$

O equivalentemente

$$Z_{t+l} = \sum_{j=1}^{\infty} \pi_j Z_{t+l-j} + a_{t+l}$$

Siguiendo a Wegman(1986), aplicamos el operador $1 - \psi_1 B + \dots + \psi_{l-1} B^{l-1}$

Obtenemos

$$\sum_{j=0}^{\infty} \sum_{k=0}^{l-1} \pi_j \psi_k Z_{t+l-j-k} + \sum_{k=0}^{l-1} \psi_k a_{t+l-k} = 0$$

Donde $\pi_0 = -1$ y $\psi_0 = 1$. Puede mostrarse fácilmente que:

$$\sum_{j=0}^{\infty} \sum_{k=0}^{l-1} \pi_j \psi_k Z_{t+l-j-k} = \pi_0 Z_{t+l} + \sum_{m=1}^{l-1} \sum_{i=0}^m \pi_{m-i} \psi_i Z_{t+l-m} + \sum_{j=1}^{\infty} \sum_{i=0}^{l-1} \pi_{l-1+j-i} \psi_i Z_{t-j+1}$$

Escogiendo los ponderadores ψ tales que

$$\sum_{i=0}^m \pi_{m-i} \psi_i = 0 \text{ para } m = 1, 2, \dots, l-1$$

Tenemos

$$Z_{t+l} = \sum_{j=1}^{\infty} \pi_j^{(l)} Z_{t-j+1} + \sum_{i=0}^{l-1} \psi_i a_{t+l-i}$$

Donde

$$\pi_j^{(l)} = \sum_{i=0}^{l-1} \pi_{l-1+j-i} \psi_i$$

Así, dado Z_t para $t \leq n$, tenemos

$$\hat{Z}_n(l) = E Z_{n+l} | Z_t, t \leq n = \sum_{j=1}^{\infty} \pi_j^{(l)} Z_{n-j+1}$$

Porque $E a_{n+j} | Z_t, t \leq n = 0$ para $j > 0$. El error de pronóstico es

$$e_n(l) = Z_{n+l} - \hat{Z}_n(l) = \sum_{j=0}^{l-1} \psi_j a_{n+l-j}$$

Donde los ponderadores ψ_j pueden ser calculados recursivamente a partir de π_j como sigue:

$$\psi_j = \sum_{i=0}^{j-1} \pi_{j-i} \psi_i \quad \text{para } j = 1, \dots, l-1$$

Para un proceso estacionario existe $\lim_{l \rightarrow \infty} \sum_{j=0}^{l-1} \psi_j^2$. De ahí eventualmente se pueden aproximar los límites para el pronóstico por dos líneas horizontales paralelas.

Para el proceso no estacionario $\lim_{l \rightarrow \infty} \sum_{j=0}^{l-1} \psi_j^2$ no existe. De hecho $\sum_{j=0}^{l-1} \psi_j^2$ se incrementa indefinidamente cuando $l \rightarrow \infty$. Así los límites para el pronóstico en este caso son más y más amplios a medida que l crece. La implicación práctica es que el pronóstico es menos certero a medida que el tiempo es más distante.

Código

```
# rm(list=ls(all=TRUE))

# Lectura y acomodo de Base de Datos

library(xlsReadWrite)

#ruta <- "C:\\JR\\Tesis\\Jackeline\\Datos\\"

ruta <- "D:\\TESIS_MAESTRIA\\RAMON3\\"

NomArch <- c("1 ALIMENTOS","2 ROPA","3 VIVIENDA","4 MUEBLES","5 SALUD",
"6 TRANSPORTE","7 EDUCACION","8 OTROS")

for(i in 1:8)
{# i <- 1

  DAT <- NULL

  DAT <- read.xls(paste(ruta,NomArch[i],".xls",sep=""))

  DAT <- DAT[,-1]

  if(i == 1)

    dat <- DAT

  else

    dat <- cbind(dat,DAT)

}

NOMCOL <- read.xls(paste(ruta,"9 CODIGOS.xls",sep=""))

nomcol <- as.character(NOMCOL[,2])

dat <- as.matrix(dat)

cc <- which(dat == "N/E")

dat[cc] <- NA

dat <- matrix(as.numeric(dat),nrow(dat),ncol(dat))

dimnames(dat) <- list(NULL,nomcol)

dim(dat)
```

```

dimnames(dat)[[2]]

# Trabajando sólo con columnas independientes

ss <- as.numeric(substring(nomcol,nchar(nomcol)-1,nchar(nomcol)))
cc <- which(!is.na(ss))
length(cc)
NOMCOL <- NOMCOL[cc,]
mat <- dat[,cc]
nomcol <- dimnames(mat)[[2]]
nr <- nrow(mat)
nc <- ncol(mat)
Fechalni <- 1989

plot.ts( ts(mat[,1],Fechalni,freq=12),ylim=c(0,300),ylab="Índice",main="Índice de precios
para los 315 productos genéricos" )

axis(1,1989:2007)

for(i in 1:nc)
  lines.ts(ts(mat[,i],Fechalni,freq=12),col=2)

# La matriz de trabajo será mat.

# Todas las series relativas al mes anterior
dmat <- (mat[2:nr,]/mat[1:(nr-1),] - 1)*100
par(mfrow=c(1,1))
plot.ts( ts(dmat[,1],Fechalni,freq=12),ylim=c(-50,100),ylab="Relativo",main="Índice de
precios relativo al mes anterior para los 315 productos genéricos" )

axis(1,1989:2007)

```

```

for(i in 1:nc)
  lines.ts(ts(dmat[,i],FechaIni,freq=12),col=2)

aa <- mat[,1]
bb <- dmat[,1]
par(mfrow=c(2,1))
plot.ts( ts(aa,FechaIni,freq=12), ylab="Índice",main="Tortilla de maíz" ,col=2)
axis(1,1989:2007)
plot.ts( ts(bb,FechaIni+1/12,freq=12), ylab="Relativo al mes anterior",col=3 )
axis(1,1989:2007)

##
# Inicia Estimación de un AR(24)
##

np    <- 24
MATCOEF <- matrix(0,nc,np)
for(i in 1:nc)
{ # i <- 2
  ss <- dmat[,i]

# Esto se hace porque no todas las series estan completas\
  cc <- which(is.na(ss))
  if(length(cc) > 0)
    ss <- ss[-cc]
  ns <- length(ss)

```

```

ii <- (np+1):ns
li <- length(ii)
yy <- ss[ii]

# xx <- matrix(0,li,np)
# for(j in 1:np)
#   xx[,j] <- ss[ii-j]
# MATCOEF[,i] <- as.numeric(solve(t(xx)%*%xx)%*%t(xx)%*%yy)
  aa <- ar(yy,aic=FALSE,order.max=24,method="yule-walker")
  MATCOEF[i,] <- aa$ar
}
##
# Agrupando
##

library(cluster)

TIPO <- 3
if(TIPO == 1)
{
# Kmeans
metodo <- "Kmeans"
clus <- kmeans(MATCOEF,4)
grupos <- clus$clus
}
if(TIPO == 2)

```

```

{
# PAM
metodo <- "PAM"
clus <- pam(MATCOEF,4)
par(mfrow=c(1,1))
plot(clus)
grupos <- clus$clustering
}
if(TIPO == 3)
{
# AGNES
metodo <- "AGNES"
clus <- agnes(MATCOEF,metric = "euclidean",method="ward")
par(mfrow=c(1,1))
plot(clus)
grupos <- cutree(clus, k = 4)
}

table(grupos)

##
# Graficando las interpretaciones de los grupos
##

ng <- max(grupos)
ym <- range(MATCOEF)

```

```

par(mfrow=c(ng,1))
for(i in 1:ng)
{ # i <- 1

cc <- which(grupos == i)
lc <- length(cc)

plot(1:np,MATCOEF[cc[1],],type="l",ylim=ym,ylab="Coef",xlab="Lag",main=paste(metodo,"
: Grupo =",i,sep=""))

axis(1,1:24)

for(j in 1:lc)

lines(1:np,MATCOEF[cc[j],],)

mm <- colMeans(MATCOEF[cc,])

lines(1:np,mm,lwd=3,col=2)

}

pc <- princomp(MATCOEF)

summary(pc)

sco <- pc$sco[,1:2]

par(mfrow=c(1,1))

plot(sco[,1],sco[,2],pch=16,col=2,xlab="Score 1",ylab="Score 2",main="Componentes
Principales")

for(i in 1:ng)

{ # i <- 1

cc <- which(grupos == i)

points(sco[cc,1],sco[cc,2],pch=15+i,col=i)

```

```

}
text(sco[,1],sco[,2]-0.02,substring(nomcol,1,3),cex=0.5)
legend(min(sco[,1]),max(sco[,2]),paste("Grupo",1:ng),col=1:ng,pch=15+1:ng)

par(mfrow=c(ng,1))
for(i in 1:ng)
{# i <- 1
  cc <- which(grupos == i)
  aa <- dmat[,cc]
  bb <- rowMeans(aa,na.rm=T)
  matplot(1:nrow(aa),aa,type="l",lwd=1,lty=1,col=1,ylim=c(-50,100),main=paste(metodo,"":
Grupo =",i,sep=""")
  ,xaxt="n",xlab="",ylab="Relativo")
  axis(1,seq(1,nrow(aa),12),1989:2008)
  lines(1:nrow(aa),bb,lwd=3,col=2)
}

par(mfrow=c(ng,1))
for(i in 1:ng)
{# i <- 1
  cc <- which(grupos == i)
  aa <- mat[,cc]
  bb <- rowMeans(aa,na.rm=T)
  matplot(1:nrow(aa),aa,type="l",lwd=1,lty=1,col=1,ylim=c(0,300),main=paste(metodo,"":
Grupo =",i,sep=""")
  ,xaxt="n",xlab="",ylab="Indice")
  axis(1,seq(1,nrow(aa),12),1989:2008)
  lines(1:nrow(aa),bb,lwd=3,col=2)
}

```



```

}
aa <- substring(nomcol,1,1)
ta <- table(aa)
an <- names(ta)
ta <- as.numeric(ta)

TABLA <- NULL

for(i in 1:ng)
{# i <- 1
  tt <- table(aa[grupos == i])
  nt <- names(tt)
  cc <- match(nt,an)
  tt <- as.numeric(tt)
  mm <- data.frame(Id.Grupo=i,Id.Des=nt,Tamano=tt,
  Rel.Grupo=round(tt/sum(tt)*100,1),Rel.Des=round(tt/ta[cc]*100,1))

  TABLA <- rbind(TABLA,mm)
}
a1 <- c("A","E","M","O","R","S","T","V")
a2 <- c("Alimentos","Educación","Mebles","Otros","Ropa","Salud","Transporte","Vivienda")
TABLA[,2] <- a2[match(TABLA[,2],a1)]

write.xls(TABLA,paste(ruta,metodo,"_1.xls",sep=""))

TABLA2 <-
data.frame(Agregado=aa,Desagregado=nomcol,Grupo=grupos,Descripcion=NOMCOL[,4])
write.xls(TABLA2,paste(ruta,metodo,"_2.xls",sep=""))

```