

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
ESCUELA DE POSGRADO



Análisis de componentes principales evolutivos frente a segmentado:
una comparación

Tesis para obtener el grado académico de Maestro en Estadística que
presenta:

Luis Guillermo Purizaca Rosillo

ASESOR:

Sergio Camiz

Lima, 2023

Informe de Similitud

Yo, ...**Prof. Sergio Camiz**..., docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor(a) de la tesis/el trabajo de investigación titulado

Análisis de componentes principales evolutivos frente a segmentado: una comparación,
del autor **Luis Guillermo Purizaca Rosillo**,
dejo constancia de lo siguiente:


El mencionado documento tiene un índice de puntuación de similitud de **7%**. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el **11/11/2023**.

He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio.

Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:

.....Lima el 12-11-2023.....

Apellidos y nombres del asesor / de la asesora: <u>Sergio Camiz</u>	
Pass: YC2564122	Firma 
ORCID: 0000-0002-2566-5207	



Resumen

La reducción de dimensionalidad, fundamental en el análisis estadístico multivariado, busca comprender estructuras complejas de datos extensos de forma reducida. El Análisis de Componentes Principales (PCA) es comúnmente empleado para estos fines, pero en series de tiempo, sus limitaciones surgen debido a que el PCA no considera la variación de la estructura de correlación entre las series de tiempo, así como la potencial autocorrelación existente. Este estudio se enfoca en presentar el PCA Evolutivo, una propuesta metodológica desarrollada por Camiz et al. (2011), diseñada específicamente para capturar las fluctuaciones en niveles de correlación presentes en este tipo de datos. Se procede a comparar los resultados obtenidos a través del PCA Evolutivo con aquellos generados por el método PCA Segmentado, desarrollado por Banko et al. (2011). Los resultados revelan que el PCA Evolutivo complementa la descripción de datos en comparación con el PCA y ofrece perspectivas distintas respecto al PCA Segmentado en la identificación de segmentos homogéneos, relacionados con cambios en correlación en el tiempo. Estas diferencias se atribuyen a las estrategias de construcción: el PCA Evolutivo sigue un enfoque basado en el Índice de Calinski-Harabász, mientras que el PCA Segmentado se guía por el método bottom-up. La estabilidad del PCA Evolutivo depende de la ventana de tiempo, determinando la amplitud de los intervalos homogéneos. Por otro lado, el PCA Segmentado es más sensible a la cantidad de dimensiones retenidas, proporcionando mayor robustez al ajustar segmentos mediante una función de costos que disminuye con la retención de más componentes principales. Aplicamos los métodos en datos simulados y en series de tiempo de materias primas.

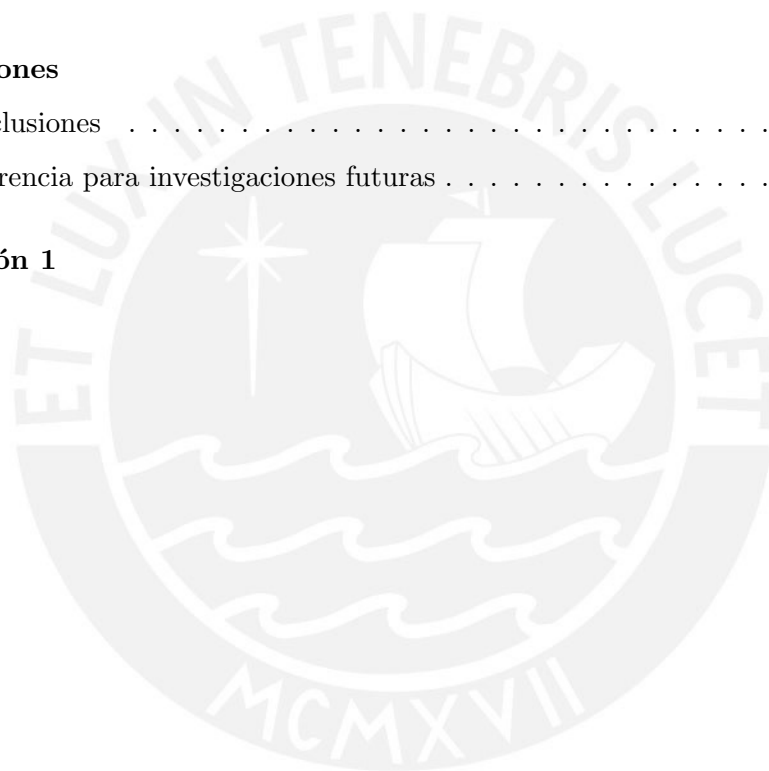
Palabras clave

Análisis de componentes principales, segmentación, análisis evolutivo, correlación, series de tiempo.

Índice general

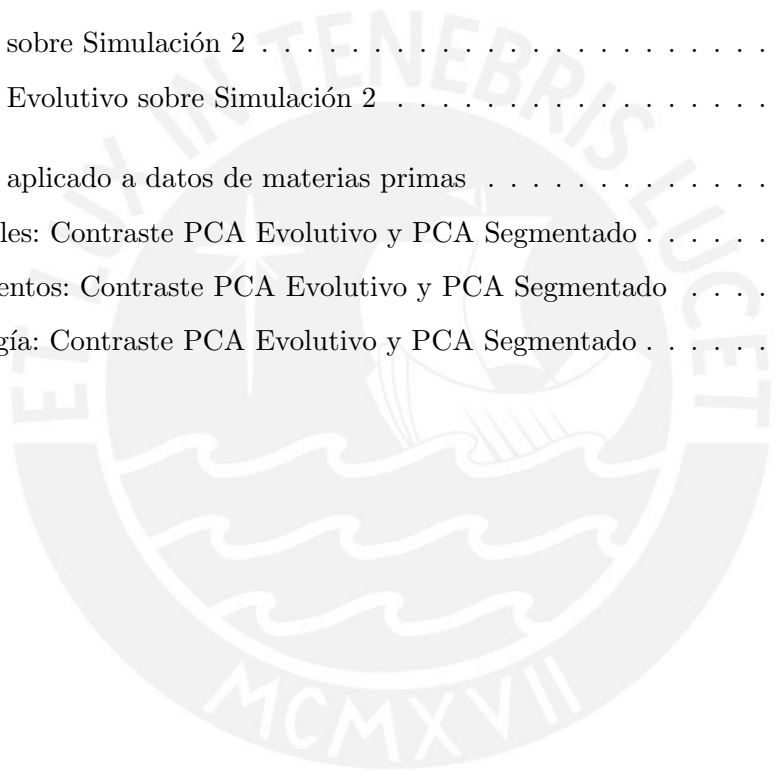
Índice de cuadros	1
Índice de figuras	2
1. Introducción	4
1.1. Consideraciones preliminares	4
1.2. Objetivos	7
1.3. Organización del trabajo	8
2. Conceptos y modelos	9
2.1. Componentes principales con datos autocorrelacionados	13
2.2. Componentes principales evolutivos - PCA Evolutivo	16
2.3. Componentes principales segmentados - PCA Segmentado	17
3. Simulaciones	20
3.1. Simulación 1: Datos autocorrelacionados con una estructura de correlación predefinida	20
3.1.1. Generación	20
3.1.2. Aplicación	21
3.1.3. Desempeño	22
3.2. Simulación 2: Datos autocorrelacionados con más de una estructura de corre- lación predefinida	25
3.2.1. Generación	25
3.2.2. Aplicación	26
3.2.3. Desempeño	26
4. Aplicación	29
4.1. Descripción de los datos	29

4.2. PCA Clásico	30
4.3. Metales	31
4.3.1. PCA Evolutivo	31
4.3.2. PCA Segmentado	32
4.4. Alimentos	32
4.4.1. PCA Evolutivo	32
4.4.2. PCA Segmentado	32
4.5. Energía	33
4.5.1. PCA Evolutivo	33
4.5.2. PCA Segmentado	33
5. Conclusiones	35
5.1. Conclusiones	35
5.2. Sugerencia para investigaciones futuras	36
A. Simulación 1	37
Bibliografía	72



Índice de cuadros

3.1. PCA sobre Simulación 1	22
3.2. PCA Evolutivo sobre Simulación 1	23
3.3. PCA sobre Simulación 2	26
3.4. PCA Evolutivo sobre Simulación 2	28
4.1. PCA aplicado a datos de materias primas	31
4.2. Metales: Contraste PCA Evolutivo y PCA Segmentado	32
4.3. Alimentos: Contraste PCA Evolutivo y PCA Segmentado	33
4.4. Energía: Contraste PCA Evolutivo y PCA Segmentado	34



Índice de figuras

4.1. Evolución histórico de precios estandarizados de materias primas por categorías	30
A.1. Método PCA Evolutivo - Modelo A: Evolución del primer autovalor.	37
A.2. Método PCA Evolutivo - Índice de Calisnki-Harabász - Modelo A	38
A.3. Método PCA Evolutivo - Modelo A: Correlación del primer autovalor con las variables simuladas.	39
A.4. Método PCA Evolutivo - Modelo B: Evolución del primer autovalor.	40
A.5. Método PCA Evolutivo - Índice de Calisnki-Harabász - Modelo B	41
A.6. Método PCA Evolutivo - Modelo B: Correlación del primer autovalor con las variables simuladas.	42
A.7. Método PCA Evolutivo - Modelo B4: Evolución del primer autovalor.	43
A.8. Método PCA Evolutivo - Índice de Calisnki-Harabász - Modelo B4	44
A.9. Método PCA Evolutivo - Modelo B4: Correlación del primer autovalor con las variables simuladas.	45
A.10. Método PCA Evolutivo - Modelo B2: Evolución del primer autovalor.	46
A.11. Método PCA Evolutivo - Índice de Calisnki-Harabász - Modelo B2	47
A.12. Método PCA Evolutivo - Modelo B2: Correlación del primer autovalor con las variables simuladas	48
A.13. Método PCA Evolutivo - Modelo B3: Evolución del primer autovalor.	49
A.14. Método PCA Evolutivo - Índice de Calisnki-Harabász - Modelo B3	50
A.15. Método PCA Evolutivo - Modelo B3: Correlación del primer autovalor con las variables simuladas.	51
A.16. Método PCA Segmentado para los modelos de la Simulación 1	52
A.17. Método PCA Evolutivo - Modelo C: Evolución del primer autovalor	53
A.18. Método PCA Evolutivo - Índice de Calisnki-Harabász - Modelo C	54
A.19. Método PCA Evolutivo - Modelo C: Correlación del primer autovalor con las variables simuladas	55

A.20.Método PCA Evolutivo - Modelo D: Evolución del primer autovalor	56
A.21.Método PCA Evolutivo - Índice de Calisnki-Harabász - Modelo D	57
A.22.Método PCA Evolutivo - Modelo D: Correlación del primer autovalor con las variables simuladas	58
A.23.Método PCA Evolutivo - Modelo D4: Evolución del primer autovalor	59
A.24.Método PCA Evolutivo - Índice de Calisnki-Harabász - Modelo D4	60
A.25.Método PCA Evolutivo - Modelo D4: Correlación del primer autovalor con las variables simuladas.	61
A.26.Método PCA Evolutivo - Modelo D2: Evolución del primer autovalor.	62
A.27.Método PCA Evolutivo - Índice de Calisnki-Harabász - Modelo D2	63
A.28.Método PCA Evolutivo - Modelo D2: Correlación del primer autovalor con las variables simuladas.	64
A.29.Método PCA Evolutivo - Modelo D3: Evolución del primer autovalor.	65
A.30.Método PCA Evolutivo - Índice de Calisnki-Harabász - Modelo D3	66
A.31.Método PCA Evolutivo - Modelo D3: Correlación del primer autovalor con las variables simuladas.	67
A.32.Método PCA Segmentado para los modelos de la Simulación 2	68
A.33.Método PCA Evolutivo - Materias Primas de Metales	69
A.34.Método PCA Evolutivo - Materias Primas de Alimentos	70
A.35.Método PCA Evolutivo - Materias Primas de Energía	71

Capítulo 1

Introducción

1.1. Consideraciones preliminares

Con el fin de reducir la dimensionalidad de grandes cantidades de información de datos que permita realizar análisis simplificados sin perder cantidad de información relevante de los datos originales, el análisis de componentes principales (PCA) aparece como una de las técnicas de la estadística multivariada más populares.

Este método permite conocer de manera estructural los datos hallando una cantidad reducida de nuevas variables no correlacionadas entre si como una combinación lineal de las variables originales.

En el caso que las variables sean series de tiempo, es posible aún aplicar PCA, sin embargo, el método no considera la variación de la estructura de correlación entre las series a lo largo del tiempo mismo, así como la potencial autocorrelación existente o la correlación entre series con tiempos diferentes que se puedan presentar.

Frente a ello, diversas técnicas han sido propuestas en base al PCA para atender estos escenarios en diversas áreas de estudio. Por ejemplo, Ku et al. (1995) desarrollaron un método similar al PCA denominado análisis de componentes principales dinámico (DPCA). Este método consiste en incorporar rezagos de las variables mismas de tal forma que los componentes principales capturan información presente como histórica de las variables originales asumiendo estacionariedad. Sin embargo, el método no define objetivamente la cantidad de componentes principales ni la cantidad de rezagos a considerar.

Por su parte, Liu et al. (2004) aplicaron el método DPCA en un conjunto de datos de monitoreo de variables industriales definiendo métodos apropiados para la selección en el número de rezagos a incorporar a través de los criterios AIC y BIC. Sin embargo, este proceso no asegura la selección de una cantidad de rezagos óptimo.

La implicancia de obviar este aspecto fue abordado por Vanhatalo and KulaHCI (2016) y Zamprogno et al. (2020) quienes encontraron que el número de componentes principales retenidos para explicar una proporción relevante de los datos aumenta cuando una estructura de autocorrelación es incorporada. Asimismo, cuando la estructura de autocorrelación es muy parecida para todas las variables de un conjunto de datos, estos no se ven influenciados tan significativamente como cuando la estructura de autocorrelación es de distinto grado para cada variable y con signos opuestos.

Sobre estos aspectos, en Peña and Yohai (2016) y Peña et al. (2020) fue generalizado el método de DPCA donde las componentes principales son caracterizadas por (i) ser no necesariamente una combinación lineal de las observaciones, (ii) construirse usando diversas funciones de pérdida de reconstrucción, (iii) ser adaptable a series no estacionarias y (iv) ser flexible cuando las series presentan outliers. El método permite seleccionar de manera automática el número de rezagos a usar así como el número de componentes principales a considerar.

El método generalizado DPCA no solo aborda la subjetividad sobre la cantidad de rezagos y componentes principales seleccionados sino también la posible no estacionariedad. Al respecto, otros métodos abordaron estas características.

En el caso del comportamiento de series de tiempo de precios de commodities, diversos estudios encuentran que son caracterizadas por ser estacionales o cíclicas, ser correlacionadas con variables económicas (inflación, producción industrial, crecimiento económico, entre otros), presentar variaciones poco recurrentes pero explosivas, presentar correlación variante en el tiempo, evidenciar la existencia de superciclos, etc (Deaton and Laroque (1992) y Erten and Ocampo (2021)). Al respecto, identificar reglas y patrones sobre estas series de forma reducida posibilitaría un mejor análisis para la identificación de eventos especiales (booms o crush), movimientos extremos, cambios en la correlación y tendencia (régimen) o períodos estables, es decir brindaría un mejor entendimiento del pasado que puedan usarse para me-

jores predicciones a futuro con aportes en la gestión de riesgo financiero, políticas públicas, gestión de portafolios de cobertura, entre otros.

Por este motivo, han sido propuestos métodos para la identificación de cambios estructurales como los modelos econométricos orientados más a la inferencia (Markov Switching, por ejemplo) o métodos orientados más al análisis exploratorio (segmentación, ventanas móviles, entre otros).

Para estudiar la variación en los niveles de correlación Camiz et al. (2011) proponen un nuevo método para identificar estos cambios a lo largo del tiempo denominado Análisis de Componentes Principales Evolutivos (PCA Evolutivo) basado en la aplicación del método tradicional PCA en ventanas de tiempo móviles y así estudiar períodos de tiempo homogéneos que sean de interés. Del mismo modo, Banko et al. (2011) proponen de forma similar la identificación y construcción de segmentos de datos homogéneos basado en funciones de costos a partir de la aplicación del método tradicional (PCA Segmentado).

Banko et al. (2011) propuso la segmentación de la base de datos para así considerar el tratamiento de la estacionariedad; el método PCA es aplicado de forma paralela a la construcción de cada segmento con el fin de identificar cambios en los procesos dinámicos de los datos y así encontrar períodos homogéneos. Para ello, usaron las medidas estadísticas del T^2 de Hotelling y Q -Error de reconstrucción como herramientas para definir los segmentos según alcancen el nivel de límite definido. Según Banko et al. (2011) esta especificación del método permite identificar modelos localmente lineales directamente y definir fácilmente la convergencia dentro de cada segmento construido.

Con el mismo objetivo, Camiz et al. (2011) proponen un método evolutivo del PCA, el cual se basa en considerar ventanas móviles consecutivas de la base de datos y aplicar PCA de manera paralela. El método tiene como objetivo estudiar la evolución de correlación entre los datos originales y las componentes principales. De acuerdo con Camiz et al. (2011), el método aplicado en series de tiempo puede ser útil para reconocer la existencia de homogeneidad de subgrupos en intervalos de tiempos y a la vez identificar variaciones importantes entre intervalos.

Atoche Murrieta (2022) también aplicó el método PCA Evolutivo apoyado también del

método de clasificación factorial jerárquica (HFC, por sus siglas en inglés) en un conjunto de datos de variables financieras del Perú para identificar patrones y características del ciclo financiero entre los años 2000-2020. A través del método HFC, identificó tres grupos de variables que representaban la evolución del ciclo financiero en el corto, mediano y largo plazo durante el período de estudio; y con los métodos PCA y PCA Evolutivo identificó las variables más importantes de cada grupo. Como parte de sus hallazgos a partir del análisis exploratorio realizado planteó monitorear el precio de las acciones para un entendimiento de las fluctuaciones en el corto plazo, las tasas de interés para el mediano plazo y el sistema financiero para el largo plazo.

La relevancia de estos últimos enfoques consiste en que permiten extraer tramos de tiempo con observaciones similares o diferentes al resto de datos. De esta forma, las series de tiempo son reducidas a períodos homogéneos para la identificación de patrones, tendencias, reglas, duración de shocks externos o puntos de quiebre de manera simplificada con una aplicación interesante para considerar frente a conjuntos de datos multivariados con gran amplitud, diversidad y complejidad. Por ello, la presente tesis tiene como objetivo estudiar las diferencias metodológicas de los métodos PCA Evolutivo y PCA Segmentado en una aplicación real a datos de materias primas de forma exploratoria.

Actualmente, la evolución futura de los precios depende de cambios en factores estructurales como respuesta a eventos con efectos inciertos aún, tales como la evolución o fin de la pandemia por el Covid-19 o el desarrollo del conflicto Ucrania-Rusia más recientemente.

1.2. Objetivos

El objetivo general es estudiar de manera comparativa los métodos PCA Evolutivo y PCA Segmentado para identificar períodos con estructura de correlación similares e instantes donde hay cambios estructurales a lo largo del tiempo.

De forma específica se busca:

1. Estudiar las propiedades de ambos métodos.
2. Implementar los métodos en código R.
3. Realizar simulaciones de diversos conjuntos de datos, estudiar la sensibilidad de los métodos a cambios en los parámetros de entrada y comparar resultados.

4. Aplicar los métodos a datos reales.

1.3. Organización del trabajo

En adelante, en el Capítulo 2 se presentará la derivación matemática del método PCA de forma matricial y los diversos teoremas necesarios para su demostración. Se desarrollará el impacto de la correlación serial sobre el método, sus implicancias e impactos, para así presentar las variaciones sobre el método PCA propuestas por Camiz et al. (2011) y Banko et al. (2011).

Luego, en el Capítulo 3, se realizarán ejercicios de simulación sobre conjuntos de datos con diversas estructuras simuladas, como correlación constante, cambiante, con presencia y ausencia de correlación serial. En este capítulo, se prueba el desempeño de los métodos estudiados para la reducción de dimensionalidad, la generación de segmentos homogéneos y cambios de correlación a lo largo del tiempo.

El Capítulo 4 contiene la aplicación de los métodos sobre datos de series de tiempo reales. Los datos cubren alrededor de 21 commodities individuales que son parte del Índice de Precios de Commodities del Fondo Monetario Internacional (FMI). Se analizarán los resultados de cada método para concluir y brindar recomendaciones para posteriores estudios en el Capítulo 5.

Capítulo 2

Conceptos y modelos

Sea \mathbf{X} una matriz de n observaciones y p variables centradas en su valor medio de cada variable y varianzas unitarias, donde $X = [X_1 \ X_2 \ \dots \ X_p]$ tal que $X_i = \{X_{ti}\}$, $\forall i = 1, 2, \dots, p$ es el vector i medido a lo largo de los puntos de tiempo $t = 1, 2, \dots, n$; el método PCA consiste en determinar una base del espacio generado para las columnas de \mathbf{X} formada de vectores ortogonales y ordenados de tal manera que cada uno resume la mayor parte de la inercia de la tabla de datos no resumida por los precedentes. De esta forma, los componentes principales, o sea los valores atribuidos a las unidades en esta nueva tabla de datos, son las combinaciones lineales de las variables originales que son máximamente correlacionadas con estas, siempre bajo el vínculo de ortogonalidad con las precedentes.

Para derivar la construcción matemática del método PCA se inicia con la definición de autovalores y autovectores complementado con el uso de algunos teoremas, cuyas demostraciones pueden encontrarse en Ford (2015):

Autovalores y autovectores

1. El autovector de una matriz \mathbf{X} es un vector distinto de cero $\boldsymbol{\nu}$ en \mathbb{R}^n tal que $\mathbf{X}\boldsymbol{\nu} = \lambda\boldsymbol{\nu}$, para algún escalar λ .
2. El autovalor de \mathbf{X} es un escalar de λ tal que la ecuación $\mathbf{X}\boldsymbol{\nu} = \lambda\boldsymbol{\nu}$ tiene una solución no trivial.

Entonces, si $\mathbf{X}\boldsymbol{\nu} = \lambda\boldsymbol{\nu}$ para $\boldsymbol{\nu} \neq 0$, diremos que $\boldsymbol{\nu}$ es un autovector correspondiente al autovalor λ .

Teorema 2.0.1 (Teorema Espectral). *Si una matriz real \mathbf{X} es simétrica $n \times n$, entonces puede ser factorizada como sigue:*

$$\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$$

donde:

1. \mathbf{V} es una matriz ortogonal que contiene los n autovectores ortogonales de \mathbf{X} , tal que $\mathbf{V}'\mathbf{V} = \mathbf{I}$;
2. $\mathbf{\Lambda}$ es una matriz diagonal que contiene los correspondientes autovalores.

Teorema 2.0.2 (Descomposición en valores singulares). *Cualquier matriz \mathbf{X} de $n \times p$ puede ser factorizada como sigue:*

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$$

donde:

1. \mathbf{U} es una matriz ortogonal real con dimensiones $n \times n$, tal que $\mathbf{U}'\mathbf{U} = \mathbf{I}$;
2. \mathbf{V} es una matriz ortogonal real con dimensiones $p \times p$ tal que $\mathbf{V}'\mathbf{V} = \mathbf{I}$;
3. $\mathbf{\Sigma}$ es una matriz real no negativa con dimensiones $n \times p$, cuyas entradas diagonales σ_i están ordenadas de forma decreciente.

Los valores en las diagonales de $\mathbf{\Sigma}$ son conocidos como valores singulares de \mathbf{X} y corresponden a la raíz cuadrada de los autovalores de $\mathbf{X}'\mathbf{X}$ o $\mathbf{X}\mathbf{X}'$.

Esto es, si se premultiplica por \mathbf{X}' se obtiene:

$$\mathbf{X}'\mathbf{X} = (\mathbf{V}\mathbf{\Sigma}'\mathbf{U}')(\mathbf{U}\mathbf{\Sigma}\mathbf{V}')$$

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{\Sigma}'\mathbf{\Sigma}\mathbf{V}'$$

$$\mathbf{X}'\mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{\Sigma}'\mathbf{\Sigma}$$

Por lo que $\mathbf{\Sigma}'\mathbf{\Sigma}$ contiene los autovalores de $\mathbf{X}'\mathbf{X}$ entonces $\mathbf{\Sigma}$ contiene la raíz cuadrada de los autovalores de $\mathbf{X}'\mathbf{X}$ denominados valores singulares.

Por lo tanto, aplicar el Teorema 2.0.2 a \mathbf{X} es equivalente a aplicar el Teorema 2.0.1 a $\mathbf{X}'\mathbf{X}$ donde \mathbf{V} contiene la base ortonormal de los autovectores de $\mathbf{X}'\mathbf{X}$ y los valores singulares de \mathbf{X} también son equivalentes a la raíz cuadrada de los autovalores de $\mathbf{X}\mathbf{X}'$. De la

misma manera, U contiene la base ortonormal de los autovectores de $\mathbf{X}\mathbf{X}'$.

En ambos casos, si los autovalores o los valores singulares son únicos (no múltiple), dicha descomposición es única, a menos del signo de los autovectores. Además, se suelen ordenar los autovalores, valores singulares y autovectores por orden decreciente en correspondencia con los autovalores o los valores singulares, lo cual será útil en adelante.

Además:

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$$

$$\mathbf{V}'(\mathbf{X}'\mathbf{X})\mathbf{V} = \mathbf{\Lambda}$$

Dado que \mathbf{V} es ortogonal entonces la varianza de los datos es la suma de los autovalores $\mathbf{\Lambda}$ de su matriz de covarianzas $\mathbf{X}'\mathbf{X}$.

Norma de Frobenius La norma de Frobenius de una matriz \mathbf{X} de dimensiones $n \times p$, denotada por $\|\mathbf{X}\|_F$ está definida por:

$$\|\mathbf{X}\|_F := \sqrt{\sum_{i=1}^n \sum_{j=1}^p x_{ij}^2} = \sqrt{\text{tr}(\mathbf{X}'\mathbf{X})}$$

Una propiedad importante es que la norma de Frobenius es la invarianza bajo la multiplicación de matrices ortogonales. Además, si el Teorema 2.0.2 es aplicado, la norma de Frobenius de \mathbf{X} puede calcularse usando los valores singulares de \mathbf{X} o los autovalores de $\mathbf{X}'\mathbf{X}$ si el Teorema 2.0.1 es aplicado. Es decir:

$$\|\mathbf{X}\|_F^2 = \text{tr}(\mathbf{X}'\mathbf{X}) = \text{tr}((\mathbf{V}\mathbf{\Sigma}'\mathbf{U}')(\mathbf{U}\mathbf{\Sigma}\mathbf{V}')) = \text{tr}(\mathbf{\Sigma}'\mathbf{\Sigma}) = \text{tr}(\mathbf{\Lambda}) = \lambda_1 + \dots + \lambda_p = \sum_{i=1}^p \lambda_i$$

Aproximación Low-Rank Sea una matrix $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$ donde $\mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_p) = \text{diag}(\lambda_1^{1/2}, \dots, \lambda_p^{1/2})$ con $|\sigma_1| > \dots > |\sigma_p|$ del mismo modo que $|\lambda_1^{1/2}| > \dots > |\lambda_p^{1/2}|$. Entonces, dado $1 \leq k \leq p$, se obtiene una matriz \mathbf{X}_k con rango k si se consideran solo k términos mayores de los valores singulares de \mathbf{X} o los k términos mayores de los autovalores de $\mathbf{X}'\mathbf{X}$.

Usando la norma de Frobenius, la aproximación genera autovalores omitidos en el pro-

ceso definido por $\|\mathbf{X} - \mathbf{X}_k\|_F$ y dado que es invariante bajo la multiplicación de matrices ortogonales, el problema se reduce a comparar Σ y Σ_k o lo que resulta equivalente a observar la sumatoria de autovalores no considerados en la aproximación:

$$\|\mathbf{X} - \mathbf{X}_k\|_F^2 = \|\mathbf{X}\|_F^2 - \|\mathbf{X}_k\|_F^2 = \sum_{i=1}^p \lambda_i - \sum_{j=1}^k \lambda_j = \sum_{i=k+1}^p \lambda_i$$

Es importante notar que al aproximarse a la matriz original \mathbf{X} con un rango menor \mathbf{X}_k , se considera la suma de los k autovalores de la matriz de covarianzas $\mathbf{X}'\mathbf{X}$. Dado que se excluye información durante el proceso, implica que al buscar maximizar la varianza de los datos con una aproximación de menor rango, simultáneamente se busca minimizar la diferencia expresada como $\|\mathbf{X} - \mathbf{X}_k\|_F^2$.

Este resultado es conocido como el Teorema de Eckart-Young, cuya demostración puede encontrarse en Eckart and Young (1936), y permite expresar el error relativo de una aproximación de bajo rango en términos de valores singulares de \mathbf{X} , esto es:

$$\frac{\|\mathbf{X} - \mathbf{X}_k\|_F^2}{\|\mathbf{X}\|_F^2} = \frac{\sigma_{k+1}^2 + \dots + \sigma_p^2}{\sigma_1^2 + \dots + \sigma_r^2}$$

De tal forma que si el objetivo es garantizar un límite determinado de error relativo se puede encontrar un valor adecuado de k examinando los valores singulares de forma directa en lugar de hacer diversas iteraciones.

Por lo tanto, en base al Teorema 2.0.2 y la aproximación de menor rango surge una relación directa con el método PCA. Esto es, dada una dimensión objetivo $k \leq p$, el método consiste en encontrar una nueva matriz de datos de menor dimensionalidad a partir de la matriz de datos original tal que el error de reconstrucción sea minimizado.

Si el error de reconstrucción es definido a partir de la norma de Frobenius $\|\mathbf{X} - \mathbf{X}_k\|_F^2$, Eckart and Young (1936) plantea que existe una elección óptima $\mathbf{X}_k = \mathbf{U}_k \Sigma \mathbf{V}_k'$ utilizando los k primeros valores singulares de \mathbf{X} equivalentes a la raíz cuadrada de los autovalores de Λ conforme el Teorema 2.0.2. Donde la matriz \mathbf{U}_k asociada contiene los k primeros autovectores de $\mathbf{X}\mathbf{X}'$. De esta forma, los componentes principales son $\mathbf{C}_k = \mathbf{U}_k' \mathbf{X}_k = \Sigma \mathbf{V}_k'$.

En cuánto a la elección óptima de k , en el análisis exploratorio no se necesita una regla de retención del número de componentes principales, que si puede ser útil en otro tipo de

métodos (análisis factorial, por ejemplo). Si bien existen varias reglas en la literatura, ninguna resulta ser realmente exitosa (vease Camiz and Pillar (2018)). En nuestro caso, ya que se necesita para tratar el análisis dinámico, utilizaremos la regla propuesta en referencia a Jolliffe and Cadima (2016), la cual consiste en limitar la atención a las componentes principales con autovalores mayores de 0.7, lo que significa que, como las variables son estandarizadas, se retienen las componentes principales que corresponden por lo menos al 70% de una variable estandarizada.

2.1. Componentes principales con datos autocorrelacionados

Sea X_i una serie de tiempo, $\forall i = 1, 2, \dots, p$ donde cada X_i sigue un proceso autorregresivo (AR) y el conjunto de datos puede ser expresado como un proceso de vectores autorregresivos (VAR):

$$X_i = \{X_{ti}\} = \left\{ \sum_{j=1}^h \Phi_j X_{(t-j)i} + u_{ti} \right\}$$

donde h es el orden del modelo VAR, Φ_j es una matriz diagonal que muestra la estructura de autocorrelación asociada al rezago j -ésima y u_{ti} es ruido blanco.

Sin pérdida de generalidad, el proceso VAR de orden q puede ser estudiado como un proceso VAR de orden 1. Entonces:

$$X_i = \{X_{ti}\} = \{\Phi_1 X_{(t-1)i} + u_{ti}\}$$

Donde $u_{ti} \sim N(0, \Sigma)$ con correlación serial cero y para fines prácticos en las siguientes formulaciones $\Phi_1 = \Phi$.

De forma que la varianza de \mathbf{X} es afectado como sigue:

$$XX' = Cov(X, X) = E(X_{ti}X'_{ti})$$

$$XX' = E((\Phi X_{(t-1)i} + u_{ti})(\Phi X_{(t-1)i} + u_{ti})')$$

$$XX' = E((\Phi X_{(t-1)i} + u_{ti})(X'_{(t-1)i}\Phi' + u'_{ti}))$$

$$XX' = \Phi E(X_{(t-1)i}X'_{(t-1)i})\Phi' + E(u_{ti}u'_{ti})$$

$$XX' = \Phi XX' \Phi' + \Sigma$$

Es decir, la varianza del conjunto de datos \mathbf{X}_i es igual a Σ , covarianza de u_{ti} , más el impacto que genera el efecto de la correlación serial sobre la misma matriz covarianza de \mathbf{X} .

En ese sentido, se puede derivar lo siguiente:

1. Cuando no existe correlación serial para cada \mathbf{X}_i , la covarianza de u_{ti} reflejados en Σ contiene valores en la diagonal equivalentes a la varianza de cada \mathbf{X}_i y fuera de la diagonal la covarianza entre las variables.
2. Cuando existe correlación serial y todos los elementos Φ son los mismos se puede mostrar que la varianza del conjunto de datos \mathbf{X} está dada por:

$$XX' = \Phi XX' \Phi' + \Sigma$$

$$XX' = \phi I XX' \phi I + \Sigma$$

$$XX' = \phi^2 XX' + \Sigma$$

$$XX' = \frac{1}{1 - \phi^2} \Sigma$$

Mientras que la covarianza del conjunto de datos \mathbf{X} :

$$X_{ti} = \Phi X_{(t-1)i} + u_{ti}$$

$$X_{(t-1)i} = \Phi X_{(t-2)i} + u_{(t-1)i}$$

⋮

Entonces:

$$X_{ti} = \Phi(\Phi X_{(t-2)i} + u_{(t-1)i}) + u_{ti}$$

$$X_{ti} = \Phi^2 X_{(t-2)i} + \Phi u_{(t-1)i} + u_{ti}$$

⋮

O expresado de otra forma:

$$X_{(t+h)i} = \Phi^h X_{ti} + \sum_{j=1}^{h-1} \Phi^j u_{t+h-i}$$

Entonces,

$$XX'_h = (X_t)(\Phi^h X_t)'$$

$$XX'_h = XX'(\Phi^h)'$$

Por lo que $XX'_h = XX'(\Phi^h)' = \Sigma \frac{(\Phi^h)'}{1-\phi^2}$ para todo $h \geq 0$ y sugiere que la matriz de covarianzas de \mathbf{X} ante la presencia de correlación serial es equivalente a una ponderación de matriz de covarianzas Σ .

En tanto los autovectores de $\mathbf{X}\mathbf{X}'$ como de Σ son los mismos mientras los autovalores de XX'_h son ponderados por $\frac{(\Phi^h)'}{1-\phi^2}$.

Por lo tanto, los componentes principales pueden ser obtenidos a partir de Σ dado que tanto la varianza como la covarianza del conjunto de datos \mathbf{X}_i son proporcionales a Σ .

3. Cuando cada uno de los elementos Φ no son iguales entonces la correlación entre variables es distorsionada y origina grandes cambios cuando incluso existen signos opuestos en cada elemento de Φ (ver Vanhatalo and Kulaheci (2016)).

Del mismo modo, los componentes principales $Y = U'X$ son afectados por la correlación temporal de \mathbf{X} , como sigue: $YY'_h = \Lambda U'(\Phi^h)'U$ para todo $h \geq 0$.

Esto es:

Si $Y = U'X$ entonces $YY' = U'XX'U$ y del mismo modo $YY'_h = U'XX'_hU$. Entonces:

$$YY'_h = U'XX'(\Phi^h)'U$$

$$YY'_h = U'U\Lambda U'(\Phi^h)'U$$

$$YY'_h = \Lambda U'(\Phi^h)'U$$

De modo que para todo $h \geq 0$, los autovalores Λ de $\mathbf{X}\mathbf{X}'$ y los autovectores correspondientes U también son afectados por (Φ^h) .

2.2. Componentes principales evolutivos - PCA Evolutivo

Para analizar la evolución de la correlación entre un conjunto de variables, el método PCA Evolutivo es introducido en Camiz et al. (2011). El método tiene como objetivo observar los resultados de PCA aplicados a datos de tipo series de tiempo sobre ventanas móviles consecutivas, específicamente la serie temporal de los primeros autovalores emitidos por cada PCA y la serie temporal de las correlaciones entre las series originales y las primeras componentes principales.

El enfoque evolutivo sigue el siguiente proceso paso a paso:

1. Definir la longitud m de las ventanas móviles sobre la serie de tiempo de cada variable del conjunto de datos originales \mathbf{X} .
2. Elegir pesos w_i para cada observación $i = 1, 2, \dots, m$ de la ventana móvil a partir de una función gaussiana. Donde la media es igual a $\mu = \frac{(m+1)}{2}$ y la desviación a $\sigma = \frac{(m-1)}{\sqrt{2}}$ son usados para calcular el parámetro $v_i = e^{-\frac{1}{2}(\frac{(i-\mu)}{\sigma})^2}$, el cual vuelve a ser escalado para obtener los pesos $w_i = \frac{v_i}{\sum_i v_i}$ aplicados a todas las ventanas. De esta manera, se calculan matrices de correlación subsiguientes a través de diferentes pesos a las unidades que contribuyan de manera diferente debido a la ponderación variada.
3. Calcular PCA en cada una de las ventanas móviles considerando la matriz de correlaciones calculada con dichos pesos.
4. Observar los resultados graficando la serie temporal de los primeros autovalores emitidos por el PCA y la serie temporal de las correlaciones entre las series originales y las primeras componentes principales. De esta manera se podrán apreciar las fluctuaciones de los autovalores, variaciones de signos en la estructura de correlaciones, que esta mejor presentada para el correspondiente grafico de las correlaciones entre variables y componentes principales.
5. Aplicar el algoritmo descrito en Fisher (1958) a la serie de tiempo, en particular al primer valor propio emitido por PCA Evolutivo. El algoritmo identifica como puntos de

corte en la serie los intervalos entre dos fechas posteriores minimizando la varianza dentro de los grupos formados. Como el algoritmo puede proporcionar varias particiones en un número creciente de clases, a la secuencia se aplica el Índice *Calinski – Harabasz*, buscando los máximos locales del ratio $\frac{BSS/(k-1)}{WSS/(n-k)}$, donde *BSS* es la suma de cuadrado entre grupos, *WSS* es la suma de cuadrado dentro de cada grupo y *k* es el número de grupos que se van a formar. Dado que no se conoce el valor de *k*, primero se consideran $k = 2$, luego $k = 3$, y así sucesivamente. En cada etapa se calcula el ratio indicado anteriormente y el valor de *k* con el ratio de mayor valor es elegido. En caso *k* sea creciente con el ratio y no encuentre un valor máximo, puede considerarse los casos donde hay un salto fuerte en *BSS* o *WSS* como puntos de cortes a lo largo de la evolución de la serie.

Para definir el parámetro *m* en el paso 1, es importante recordar que para aplicar el método PCA es requerido tener más observaciones que variables. Por lo tanto, al momento de definir la ventana móvil debe ser un intervalo de tiempo más grande que el número de variables.

Por otra parte para el paso 2, cuando la matriz de correlación es calculada en dos ventanas consecutivas, estos diferirían en la primera observación de la primera ventana que se elimina y la última observación de la segunda ventana que ingresa. Entonces, la variación entre dos ventanas subsiguientes dependería de solo dos veces las unidades $n - 1$ distantes entre si, y eso contribuye solo a $1/n$ para el cálculo de la correlación. Por este motivo las unidades son ponderadas con pesos con distribución gaussiana, con el fin de aumentar la importancia de las unidades centrales con respecto a las colas de esta forma todas las unidades tienen un peso diferente, correspondiente a su posición debido al desplazamiento de la ventana.

2.3. Componentes principales segmentados - PCA Segmentado

Los cambios en la correlación de los datos generan segmentos específicos dentro de un mismo conjunto de datos multivariado, donde cada segmento tiene sus propias características y que en un análisis total de los datos con estas características no sería posible identificar. Para ello, Bankó and Abonyi (2012) propone lo siguiente:

1. Crear segmentos iniciales bajo el enfoque *Bottom – Up*, el cual está basado en la re-

presentación lineal por partes que consiste en definir cada uno de los elementos del conjunto de datos como segmentos individuales para realizar la unión iterativa de los pares de segmentos consecutivos tanto a la derecha como la izquierda que generen el menor incremento de la función de costos definida, ver Lovrić et al. (2014). Si bien existen métodos alternativos basados en el mismo concepto, como el *Top – Down* o *SlidingWindow*, el método usado en el PCA Segmentado produce mejores resultados y se adapta al conjunto de datos de forma lineal, ver Keogh et al. (2001).

2. Unir pares de segmentos adyacentes y calcular *PCA*, determinar el número de componentes principales retenidos.
3. Hallar el costo de unir pares de segmentos consecutivos. Se propone usar las medidas T^2 o el error de reconstrucción como funciones de costo para la construcción de segmentos homogéneos, a partir del resultado del *PCA* del ítem anterior. La primera función puede ser utilizada para segmentar los datos basados en cambios de régimen, mientras que la segunda puede ser usada en datos con cambios directos en la correlación entre las variables.
4. Encontrar el par de segmentos con menor costo de unión.
5. Unir los segmentos y actualizar los límites de los segmentos.
6. Por definición el enfoque *Bottom – Up*, el proceso se repite hasta crear la cantidad de segmentos deseados.

Las funciones de costos propuestas en el método están dadas por:

1. Distancia Euclidiana al cuadrado entre los puntos de datos originales y los datos reconstruidos a partir de los componentes principales más significativos.
2. T^2 de Hotelling el cual muestra la distancia de cada elemento desde el centro de la data.

Otros parámetros a seleccionar son la cantidad de componentes principales retenidos así como la cantidad de segmentos a crear por lo cual:

1. Bankó and Abonyi (2012) sugiere seleccionar los primeros autovalores que contengan más del 90% de la inercia del conjunto de datos.

2. En tanto la cantidad de segmentos a elegir para fines comparativos será en función a lo dispuesto por el índice *Calinski – Harabasz* propuesto y usado por Camiz et al. (2011).



Capítulo 3

Simulaciones

En esta sección aplicamos los métodos desarrollados anteriormente para diferentes estructuras de conjuntos de datos. Para ello definimos $n = 500$ como la cantidad de observaciones y $p = 5$ como número de variables usando la función *rnorm*. En la primera simulación los datos son simulados bajo una estructura de correlación definida y afectados por estructuras de autocorrelación estacionarias en los rezagos 1,5 y 12 con la misma estructura de correlación predefinida. Cada estructura de autocorrelacion representa casos específicos y busca reflejar el impacto de la correlación serial en el corto, mediano y largo plazo. En la segunda simulación se añade verticalmente dos conjuntos de datos adicionales de las mismas dimensiones con dos estructuras de correlación diferentes.

En cada una de las simulaciones los métodos son aplicados en el siguiente orden: PCA, PCA Evolutivo y PCA Segmentado. La atención en el resultado de las simulaciones estará enfocada en el porcentaje de varianza retenida por el primer componente principal y la construcción de períodos homogéneos entre el PCA Evolutivo y PCA Segmentado.

3.1. Simulación 1: Datos autocorrelacionados con una estructura de correlación predefinida

3.1.1. Generación

Los 500 datos son generados de forma aleatoria a través de la función *rnorm* con media 0 y varianza 1. Los datos son divididos en 5 columnas y posteriormente son afectados por una estructura de autocorrelación en los rezagos 1, 5 y 12.

Además, sea Φ , la estructura de correlación predefinida para la simulación está dada por:

$$A = \begin{bmatrix} 1 & 0.50 & 0.60 & 0.70 & 0.80 \\ 0.50 & 1 & 0.70 & 0.75 & 0.90 \\ 0.60 & 0.70 & 1 & 0.75 & 0.80 \\ 0.70 & 0.75 & 0.75 & 1 & 0.90 \\ 0.80 & 0.90 & 0.80 & 0.90 & 1 \end{bmatrix} \quad (3.1)$$

En tanto la estructura estacionaria de autocorrelación está dada por los siguientes casos:

1. Caso B: $Lag1 = diag(0.9)$, $Lag5 = diag(0.2)$ y $Lag12 = diag(0.3)$
2. Caso B2: $Lag2 = diag(-0.9)$, $Lag5 = diag(-0.2)$ y $Lag12 = diag(-0.3)$
3. Caso B3: $Lag3 = diag(-0.9)$, $Lag5 = diag(0.2)$ y $Lag12 = diag(0.3)$
4. Caso B4: $Lag3 = diag(0.9)$, $Lag5 = diag(-0.2)$ y $Lag12 = diag(-0.3)$

Las matrices diagonales en 3.1.1 son matrices diagonales de orden 5×5 y muestran el nivel de correlación serial con cada variable misma. Para cada rezago, todas las variables muestran el mismo nivel. Sin embargo, en cada rezago el nivel es diferente. Exploramos el caso de autocorrelación positiva (Caso B), autocorrelación negativa (Caso B2), autocorrelación mixta y fuertemente negativa (Caso B3); y autocorrelación mixta y débilmente negativa (Caso B4).

Por el lado del PCA Evolutivo se parametrizó la estandarización de las ventanas móviles bajo distintos tamaños de ventanas móviles referenciales: 12, 24 y 36. Las particiones serán derivadas de los resultados brindados por Índice de Calinski-Harabász dependiendo de los máximos alcanzados para diversas ventanas móviles referenciales.

En el caso del PCA Segmentado, la cantidad de segmentos incorporados en el modelo coincidirá con la solución de particiones considerado en el PCA Evolutivo. En tanto el criterio del costo de segmentación será el error de reconstrucción.

Los métodos son probados con una retención número de componentes principales que permitan alcanzar el 90% de la inercia de los datos para fines comparativos, ver Bankó and Abonyi (2012).

3.1.2. Aplicación

Los métodos PCA, PCA Evolutivo y PCA Segmentado son aplicados a cada conjunto de datos simulados bajo la misma estructura de correlación y considerando un conjunto sin

correlación serial (Modelo A) y otros con correlación serial en cada uno de sus casos (Modelo B, Modelo B2, Modelo B3 y Modelo B4) a partir de las matrices definidas anteriormente.

Para ello se usó una función de desarrollo propio de PCA basado en el marco teórico de conceptos y modelos. Los cálculos fueron realizados en R.

Por el lado del PCA Evolutivo del mismo modo, la aplicación se realizó en R usando el código desarrollado por Camiz et al. (2011).

En tanto el método PCA Segmentado está basado en el algoritmo descrito en Bankó and Abonyi (2012) disponible en Matlab y trasladado a R para fines de este trabajo de investigación.

3.1.3. Desempeño

El método PCA muestra una reducción en la primera dimensión en presencia autocorrelación positiva. Sin embargo, cuando el nivel de autocorrelación es negativo, el impacto es inverso, se observa una expansión del valor. En el Cuadro 3.1 se observa que la varianza explicada pasó desde 79.6 % hasta 75.9 % cuando la estructura de autocorrelación positiva es incorporada. En el caso inverso, la varianza concentrada en la primera dimensión aumentó hasta 82 %.

		Modelo A	Modelo B	Modelo B2	Modelo B3	Modelo B4
Dim 1	Autovalor	3.97	3.79	4.10	4.11	3.86
	Inercia	79.6	75.9	82.2	82.2	77.2
Dim 2	Autovalor	0.51	0.62	0.43	0.44	0.58
	Inercia	10.3	12.3	8.7	8.7	11.5
Dim 3	Autovalor	0.31	0.37	0.29	0.27	0.35
	Inercia	6.2	7.5	5.9	5.4	7.0
Dim 4	Autovalor	0.19	0.21	0.17	0.17	0.21
	Inercia	3.9	4.2	3.3	3.5	4.2
Dim 5	Autovalor	0.01	0.01	0.004	0.01	0.01
	Inercia	0.1	0.1	0.1	0.1	0.1

Cuadro 3.1: PCA sobre Simulación 1

Por otra lado, cuando hay presencia de autocorrelación mixta, el impacto sobre la primera dimensión dependerá de la representatividad del impacto, ya sea si predomina una autocorrelación positiva o negativa.

De esta forma, sin correlación serial, para estos datos son necesarios dos dimensiones para alcanzar el 90 % de la varianza total de datos mientras que en presencia de correlación serial se necesitan hasta tres dimensiones cuando hay presencia total o mayoritaria de autocorrelación negativa. Además, se mantienen en dos dimensiones cuando la autocorrelación es positiva.

En el Cuadro 3.2, para el caso del PCA Evolutivo, la varianza retenida en promedio es de 79.8 % ante la ausencia de correlación serial y de 76.8 % ante la presencia de esta característica cuando el signo es positivo y cuando es negativo en promedio el valor de 81.7 %. La situación también se refleja en el nivel correlación promedio de los datos con el primer autovalor, es decir, no hay cambios importantes en los niveles, lo cual guarda relación con el proceso de generación de datos.

Modelo	Tamaño	Prom Eigen	Prom % Inercia	Prom Corr	
Modelo A	Dim 1	12	3.99	79.9	0.79
	Dim 1	24	3.99	79.8	0.80
	Dim 1	36	3.99	79.7	0.80
Modelo B	Dim 1	12	3.83	76.7	0.71
	Dim 1	24	3.82	76.4	0.74
	Dim 1	36	3.82	76.4	0.75
Modelo B2	Dim 1	12	4.08	81.5	0.82
	Dim 1	24	4.09	81.9	0.83
	Dim 1	36	4.10	82.1	0.84
Modelo B3	Dim 1	12	4.07	81.5	0.80
	Dim 1	24	4.10	81.9	0.82
	Dim 1	36	4.11	82.1	0.83
Modelo B4	Dim 1	12	3.85	76.9	0.75
	Dim 1	24	3.85	77.0	0.75
	Dim 1	36	3.86	77.3	0.76

Cuadro 3.2: PCA Evolutivo sobre Simulación 1

En el Gráfico A.1 del Anexo se observa que los autovalores muestran estabilidad a lo largo del tiempo y las ventanas móviles seleccionadas no tienen impacto como se muestra. Los resultados son consistentes con el resultado otorgado en el método PCA tradicional. La variación experimentada en la ventana de tiempo de menor tamaño es estacionaria y se vuelve más suavizada a medida que aumenta el tamaño la variación mientras que el patrón de estacionariedad persiste alrededor de un valor en promedio.

En tanto a la evaluación del Índice de Calinski-Harabász de los Gráficos A.2 y A.5 para

la sugerencia de particiones, los resultados de los Modelos A y B indican que para ventanas móviles cortas (12 datos) se alcanza un valor relativamente mayor al nivel de tres particiones. Dado que hacia adelante los incrementos son marginales, la cantidad de grupos homogéneos propuesta por el índice en una ventana móvil de 12 puede ser considerada para explorar la evolución de la correlación de los autovalores comparados con sus respectivos conjuntos de datos definidos de donde fueron obtenidos.

La correlación del primer autovalor dentro del Modelo A en el Gráfico A.3, indican que la definición de segmentos homogéneos se ve impactada por el tamaño de la ventana móvil considerada, esto es, los segmentos generados por el método PCA Evolutivo son diferentes si se comparan las ventanas móviles de 12 respecto a 24 y 36. Este escenario coincide con los valores del Índice de Calinski-Harabász del Gráfico A.2 para esas ventanas móviles dado que se observó que constantemente el valor crecía para esas ventanas móviles con lo cual los períodos homogéneos encontrados en estos escenarios no están bien definidos.

Sin embargo, en presencia de correlación serial positiva (Modelo B y Modelo B4), los períodos homogéneos sugeridos no se ven afectados por el tamaño de ventana de tiempo elegida (ver Gráficos A.6 y A.9) indistintamente de lo sugerido por el índice en (ver Gráficos A.5 y A.8). Contrario a lo que ocurre en presencia de correlación serial negativa (Modelo B2 y Modelo B3 en los Gráficos A.12 y A.15, respectivamente); lo que también coincide con la dificultad de definir correctamente un período homogéneo debido al constante crecimiento del Índice de Calinski-Harabász (ver Gráficos A.11 y A.14).

Con lo cual, para conjuntos de datos con correlación negativa y mixta (Modelo B2, B3 y B4) para cada una de las ventanas móviles referenciales se encuentra que el Índice de Calinski-Harabász no alcanza un máximo. Lo cual dificulta la definición de particiones en estos escenarios. Sin embargo, aún si alcanzara un nivel máximo para alguna de las ventanas, como si ocurre con cuando la correlación serial es positiva, es posible que la calidad de este grupo homogéneo sea baja, dado que por construcción de los modelos en esta simulación no esperaríamos particiones pues se trata de una estructura constante en el tiempo.

Por otro lado, para el método PCA Segmentado, los segmentos generados para cada modelo son de amplitud diferente (Modelo A, B, B2, B3 y B4 - ver Gráfico A.16). Esto es, la construcción de cada segmento es sensible a la estructura de autocorrelación serial de

la series donde el método es aplicado. Para este ejercicio se utilizaron la misma cantidad de componentes principales que en el método PCA Evolutivo con el objetivo de que los resultados sean comparables.

3.2. Simulación 2: Datos autocorrelacionados con más de una estructura de correlación predefinida

3.2.1. Generación

En esta sección son simulados dos conjuntos de datos adicionales de las mismas dimensiones con sus respectivas estructuras de correlación para cada uno de los casos mencionados anteriormente (en esta sección denominadas D, D2, D3 y D4). Las estructuras son las siguientes:

$$M1 = \begin{bmatrix} 1 & 0.00 & 0.00 & 0.00 & -0.85 \\ 0.00 & 1 & -0.55 & -0.75 & 0.00 \\ 0.00 & -0.55 & 1 & 0.05 & 0.15 \\ 0.00 & -0.75 & 0.05 & 1 & -0.30 \\ -0.85 & 0.00 & 0.15 & -0.30 & 1 \end{bmatrix} \quad (3.2)$$

$$M2 = \begin{bmatrix} 1 & 0.65 & 0.10 & 0.35 & 0.25 \\ 0.65 & 1 & 0.55 & 0.75 & 0.65 \\ 0.10 & 0.55 & 1 & 0.95 & 0.15 \\ 0.35 & 0.75 & 0.95 & 1 & 0.30 \\ 0.25 & 0.65 & 0.15 & 0.30 & 1 \end{bmatrix} \quad (3.3)$$

El Modelo C será considerado como la unión de los tres conjuntos de datos (Modelo A y los conjuntos de datos derivados de M1 y M2). Las matrices 3.1.1 en cada uno de los casos de la sección anterior serán las mismas para definir la correlación serial en cada variable misma y sus casos respectivos (D, D2, D3 y D4).

Por el lado del PCA Evolutivo el método del mismo modo será aplicado para las mismas ventanas móviles de 12, 24 y 36 períodos. El número de particiones será considerado a partir de los valores máximos alcanzados por el Índice de Calinski-Harabász.

En tanto el método PCA Segmentado heredará las particiones sugeridas por el método

PCA Evolutivo para fines comparativos.

3.2.2. Aplicación

Del mismo modo que en la sección anterior, el Modelo C considera datos sin correlación serial pero ahora está constituido por tres correlaciones entre variables cada 500 datos mientras que los Modelos D, D2, D3 y D4 consideran la misma estructura de correlación serial en los mismos rezagos definidos y con los mismos coeficientes para cada uno de los casos definidos anteriormente.

3.2.3. Desempeño

En el Cuadro 3.3 se observa que la aplicación del modelo PCA sobre un conjunto de datos con más de una estructura de correlación predefinida genera una mayor dispersión en los autovalores. Sobretodo en el primer autovalor, se puede observar un menor valor que cuando el conjunto de datos mantiene una estructura de correlación ya sea ante la presencia o no de correlación serial.

Del mismo modo se observa un menor valor en la primera dimensión cuando hay presencia de correlación serial positiva y un mayor valor cuando es negativo.

		Modelo C	Modelo D	Modelo D2	Modelo D3	Modelo D4
Dim 1	Autovalor	2.33	2.20	2.38	2.38	2.28
	Inercia	46.6	43.5	47.6	47.6	45.6
Dim 2	Autovalor	0.99	1.05	0.98	0.98	1.10
	Inercia	19.8	21.0	19.5	19.7	20.5
Dim 3	Autovalor	0.93	0.97	0.92	0.93	0.90
	Inercia	18.7	19.4	18.5	18.6	17.8
Dim 4	Autovalor	0.40	0.43	0.38	0.37	0.44
	Inercia	8.0	8.5	7.6	7.5	8.8
Dim 5	Autovalor	0.35	0.37	0.38	0.33	0.36
	Inercia	7.0	7.5	6.8	6.6	7.1

Cuadro 3.3: PCA sobre Simulación 2

Por otro lado, los resultados del PCA Evolutivo sugieren la partición en 3 para todo el conjunto de datos simulado a través del Índice de Calinski-Harabász (ver Gráficos A.18, A.21, A.27, A.30 y A.24) tanto en el conjunto de datos con correlación serial como el conjunto con ausencia de esta característica. Además, se muestra que el nivel del tamaño de las ventanas móviles no afecta la sugerencia de particiones brindada por el índice. Sin embargo, puede

apreciarse un mayor nivel del índice en el conjunto de datos del Modelo C comparado el resto de modelos, lo que sugiere que el índice define de una mejor manera y de forma más coherente grupos de datos, a diferencia de cuando hay correlación serial.

En tanto a los autovalores, estos muestran los cambios esperados a lo largo del tiempo por las estructuras de correlación predefinidas (ver Gráficos A.17, A.20, A.26, A.29 y A.23). En tanto las ventanas móviles tienen impacto en la una correcta definición de los tres grupos de datos que han sido unidos. Estos tres grupos son identificados por el Índice de Calinski-Harabász alcanzado un valor máximo en ese nivel de particiones y aumenta a medida que la ventana móvil es más grande; una vez alcanzado el valor máximo hacia adelante el valor del índice decrece.

Una vez definidos la cantidad de particiones se observa también que los plazos donde se realizan las particiones se mantiene constantes para cualquier ventana de tiempo (ver Gráficos A.19, A.22, A.28, A.31 y A.25). No se observa una ampliación o reducción en la construcción de los plazos. Lo que confirma nuevamente que la selección de las ventanas móviles no tienen impacto en la estabilidad de los plazos homogéneos.

Lo resultados en el Cuadro 3.4 sugieren que en promedio el primer autovalor bajo el método PCA Evolutivo es menor que el método tradicional y además contiene menor inercia. También puede apreciarse que en este caso la correlación del primer autovalor con cada una de las variables es más bajo que en el caso de la Simulación 1 debido a la mayor dispersión del nivel de correlación a lo largo del tiempo y por los cambios existentes (ver Gráficos A.19, A.22, A.28, A.31 y A.25).

Por otro lado, para el método PCA Segmentado (Modelo C, D, D2, D3 y D4 - ver Gráfico A.32), en cuanto a los segmentos y la amplitud de cada uno, estos coincide con la generación de datos, indistintamente del efecto de la autocorrelación serial.

Modelo		Tamaño	Prom Eigen	Prom % Inercia	Prom Corr
Modelo C	Dim 1	12	3.27	65.4	0.27
	Dim 1	24	3.12	62.5	0.01
	Dim 1	36	3.01	61.3	0.17
Modelo D	Dim 1	12	3.24	64.9	0.16
	Dim 1	24	3.09	61.8	0.20
	Dim 1	36	3.02	60.4	0.16
Modelo D2	Dim 1	12	3.34	66.9	0.29
	Dim 1	24	3.19	63.8	0.11
	Dim 1	36	3.12	62.5	0.08
Modelo D3	Dim 1	12	3.35	67.1	0.18
	Dim 1	24	3.22	64.4	0.42
	Dim 1	36	3.16	63.2	-0.07
Modelo D4	Dim 1	12	3.25	64.9	0.26
	Dim 1	24	3.11	62.2	0.13
	Dim 1	36	3.05	61.0	0.30

Cuadro 3.4: PCA Evolutivo sobre Simulación 2

Capítulo 4

Aplicación

4.1. Descripción de los datos

En este capítulo aplicaremos el marco teórico desarrollado en el Capítulo 2 para analizar un conjunto de datos de tipo series de tiempo que cubre alrededor de 21 materias primas individuales que son parte del Índice de Precios de Commodities del Fondo Monetario Internacional (FMI). Los datos están disponibles en el Sistema de Precios de Commodity Primarios del FMI y obtenidos de frecuencia mensual en el período de Enero 1990 hasta Diciembre 2021 para este ejercicio. De acuerdo con Gruss and Kebhaj (2019), estas materias primas pueden organizarse en las siguientes categorías: (a) energía (petróleo west, petróleo Brent, petróleo UK, petróleo Dubai y gas natural), (b) metales (oro, aluminio, cobre, plomo, nickel, estaño y zinc) y (c) alimentos (café, maíz, trigo, azúcar, harina de soya, aceite de soya y grano de soya).

Las materias primas tuvieron en los años 2008-2009 un comportamiento generalizado de alta volatilidad, explicado en gran medida por la crisis financiera. Por el lado de la energía, la volatilidad en los niveles de precios se mantuvo con picos y caídas muy recurrentes. Por el lado de los metales, luego de la expansión vivida en la década de los años 2000-2010, posteriormente presentaron movimientos laterales, es decir en un rango de precios estable sin variaciones extraordinarias. De igual forma por el lado de las materias primas agrícolas. Por el lado de los alimentos, debido a que están más directamente asociados a los movimientos de la inflación, luego de la energía presentaron la mayor volatilidad en los últimos años. En general, todas las materias primas experimentaron presiones hacia la alza al cierre de 2021 explicado principalmente por la reactivación económica y presiones sobre la demanda por parte de China tras la pandemia por la Covid-19.

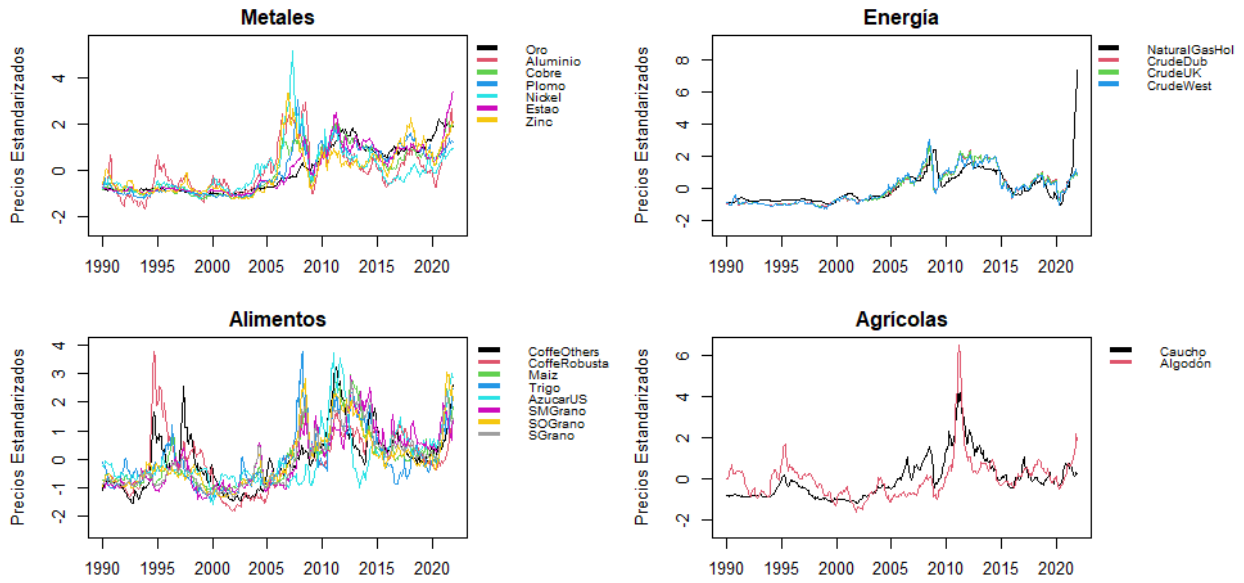


Figura 4.1: Evolución histórico de precios estandarizados de materias primas por categorías

Se observan correlaciones positivas en todos los sentidos, en mayor o menor medida. Cuando se trata de commodities de un mismo tipo, estos muestran correlación fuerte. Cuando se trata de diversos tipos, efectos mixtos son apreciados, es decir, más o menos correlacionados lo cual indica que durante el período estudiado los commodities mostraron comportamiento distintos, aunque en la misma línea.

Esto soportaría la idea que la relevancia de los commodities varía a lo largo de la historia y como tal la correlación existente entre ellos pues dependerá del cada ciclo económico atravesado así como crisis o eventos internacionales de gran relevancia que vale la pena mencionar.

4.2. PCA Clásico

Para cada uno de los conjuntos de datos, el método clásico fue aplicado y se encuentra que para los metales aproximadamente se necesitarían 2 dimensiones para concentrar 90% de la inercia. En tanto para alimentos es necesario hasta 3 dimensiones mientras en energía solo se necesita 1. Por lo cual, limitamos el análisis exploratorio a esta cantidad de componentes principales para cada materia prima.

Agrupación		Eigen	% Inercia	Acumulado
Metales	Dim 1	5.59	79.93	79.93
	Dim 2	0.84	12.05	91.99
	Dim 3	0.22	3.21	95.20
	Dim 4	0.16	2.29	97.49
	Dim 5	0.10	1.36	98.85
Alimentos	Dim 1	5.98	74.69	74.69
	Dim 2	0.85	10.60	85.30
	Dim 3	0.61	7.64	92.93
	Dim 4	0.28	3.44	96.37
	Dim 5	0.13	1.57	97.94
Energía	Dim 1	3.65	91.29	91.29
	Dim 2	0.33	8.34	99.62
	Dim 3	0.01	0.35	99.97
	Dim 4	0.00	0.03	100.00

Cuadro 4.1: PCA aplicado a datos de materias primas

4.3. Metales

4.3.1. PCA Evolutivo

Para la aplicación del método es necesario definir la ventana móvil de tiempo sobre la cuál serán calculado el PCA. Para ello, se usó el Índice Calinski-Harabasz.

Se observa que el índice aumenta a medida que se amplía la ventana de tiempo, ver Gráfico A.33. En la ventana de 12 se observa que el índice alcanza un nivel máximo y luego no muestra tendencia, lo que sugiere que existen 6 grupos homogéneos en toda la longitud de los datos.

En el caso de las ventanas 24 y 36 se observa un mayor nivel del índice como se esperaría pero no se alcanzan niveles máximos que permitan definir adecuadamente las particiones.

Por ello, se usó como ventana móvil 12 datos y una cantidad de 6 particiones.

El subgráfico (d) del Gráfico A.33 muestra los autovalores con las particiones que corresponden con cambios significativos de la serie a lo largo del tiempo. Sin embargo, se observan cambios de correlación dentro de cada partición realizada. Los segmentos identificados bajo el método pueden observarse en la Tabla 4.2.

4.3.2. PCA Segmentado

Las particiones sugeridas por el PCA Segmentado se muestran en la Tabla 4.2; considerando la cantidad de componentes principales retenidos del PCA Evolutivo realizado anteriormente. Se observa que los puntos de corte difieren entre modelos e incluso la amplitud comparado al PCA Evolutivo.

Particiones	PCA Evolutivo		PCA Segmentado	
	Fecha	Duración	Fecha	Duración
Segmento 1	Ene 1990 - Abr 2002	147	Ene 1990 - Feb 2002	205
Segmento 2	May 2002 - Set 2009	17	Mar 2002 - Set 2009	7
Segmento 3	Oct 2009 - Ene 2008	52	Oct 2008 - Set 2009	24
Segmento 4	Feb 2008 - Mar 2013	62	Oct 2009 - Jun 2014	57
Segmento 5	Abr 2013 - Mar 2014	12	Jul 2004 - Feb 2020	68
Segmento 6	Abr 2014 - Dic 2021	93	Mar 2020 - Dic 2021	22

Cuadro 4.2: Metales: Contraste PCA Evolutivo y PCA Segmentado

4.4. Alimentos

4.4.1. PCA Evolutivo

En el Gráfico A.34 se observa que el índice aumenta a medida la ventana de tiempo es de mayor longitud. En la ventana de 72 se observa que alcanza un nivel máximo y sugiere la construcción de 6 particiones. A pesar que el índice con este tamaño de ventana móvil aún crece, el crecimiento es marginal y podría entenderse como una construcción adecuada de los grupos homogéneos.

Por ello, se usó como ventana móvil 72 datos y una cantidad de 6 particiones.

Se observa que en el gráfico evolutivo de los autovalores del Gráfico A.34 que las particiones corresponden a cambios significativos de la serie a lo largo del tiempo. A nivel de correlación se observan cambios ocasionales, lo que explicaría la generación de estos segmentos. Los segmentos identificados bajo el método pueden observarse en la Tabla 4.3.

4.4.2. PCA Segmentado

Las particiones sugeridas por el PCA Segmentado se muestran en la Tabla 4.3; considerando la cantidad de componentes principales retenidos del PCA Evolutivo realizado

anteriormente, ver Gráfico A.34. Se observa que los puntos de corte difieren entre modelos e incluso en la amplitud.

Particiones	PCA Evolutivo		PCA Segmentado	
	Fecha	Duración	Fecha	Duración
Segmento 1	Ene 1990 - Oct 1994	57	Ene 1990 - Dic 1998	107
Segmento 2	Nov 1994 - Abr 1997	30	Ene 1999 - Set 2008	117
Segmento 3	May 1997 - Ago 2002	64	Oct 2008 - Dic 2009	15
Segmento 4	Set 2002 - Oct 2008	74	Ene 2010 - Ene 2014	49
Segmento 5	Nov 2008 - Jul 2013	57	Feb 2014 - Set 2019	68
Segmento 6	Ago 2013 - Dic 2021	101	Oct 2019 - Dic 2021	27

Cuadro 4.3: Alimentos: Contraste PCA Evolutivo y PCA Segmentado

4.5. Energía

4.5.1. PCA Evolutivo

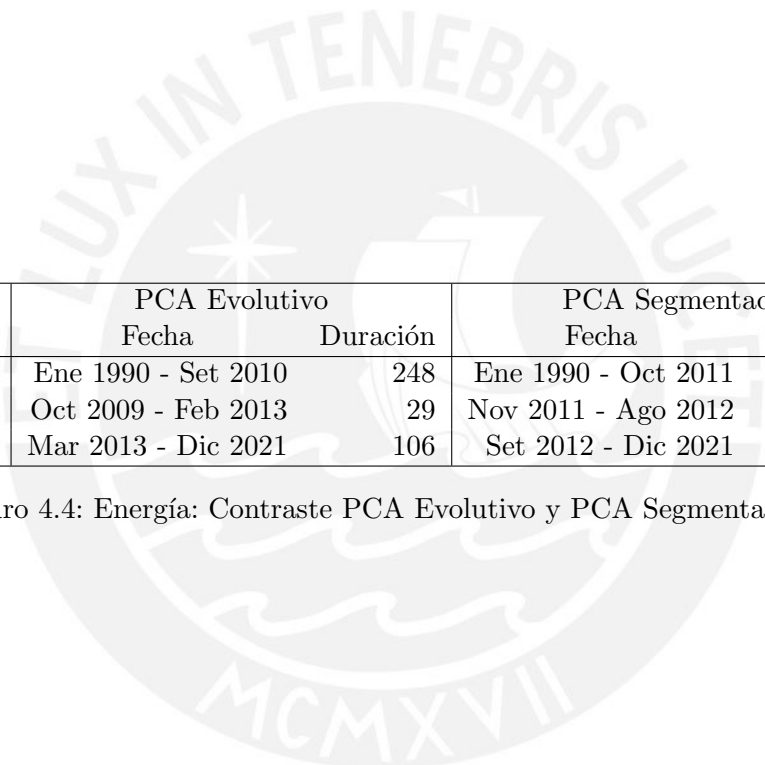
En el Gráfico A.35 se observa que el índice crece de forma desmedida conforme se amplía la ventana de tiempo. Sin embargo, puede observarse que para una ventana de tiempo reducida de 12, el valor llega aun máximo y luego retrocede para cualquier nivel de particiones. El valor máximo con ventana móvil de 12 se alcanza con un nivel de particiones de 3 a diferencia del conjunto de datos de metales y alimentos.

Por ello, se usó como ventana móvil 12 datos y una cantidad de 3 particiones.

En el gráfico evolutivo de los autovalores en el Gráfico A.35, se observa que las particiones corresponden con cambios significativos de la serie a lo largo del tiempo. A diferencia de los otros grupos de datos, el método identifica correctamente un segmento distinto al resto. Los segmentos identificados bajo el método pueden observarse en la Tabla 4.4.

4.5.2. PCA Segmentado

Las particiones sugeridas por el PCA Segmentado se muestran en la Tabla 4.4; considerando la cantidad de componentes principales retenidos del PCA Evolutivo realizado anteriormente, ver Gráfico A.34. Se observa que los puntos de corte difieren entre modelos e incluso la amplitud.



Particiones	PCA Evolutivo		PCA Segmentado	
	Fecha	Duración	Fecha	Duración
Segmento 1	Ene 1990 - Set 2010	248	Ene 1990 - Oct 2011	261
Segmento 2	Oct 2009 - Feb 2013	29	Nov 2011 - Ago 2012	10
Segmento 3	Mar 2013 - Dic 2021	106	Set 2012 - Dic 2021	112

Cuadro 4.4: Energía: Contraste PCA Evolutivo y PCA Segmentado

Capítulo 5

Conclusiones

5.1. Conclusiones

Se ha observado que el método PCA clásico en conjunto con el método PCA Evolutivo logran complementarse, permitiendo realizar análisis históricos de las series de tiempo de materias primas. Ambos métodos permitieron explorar la tendencia común de todas las series, así como el comportamiento particular que tendrían los grupos de materias primas en momentos específicos de la historia, a través de la evolución de la correlación, autovalores y particiones sugeridas.

En tanto los métodos PCA Evolutivo y PCA Segmentado otorgan resultados ligeramente distintos al momento de encontrar segmentos homogéneos. Esto en parte por el criterio de construcción de los segmentos.

En un análisis de simulación realizado en el presente trabajo, se encontró que la construcción de segmentos bajo el método PCA Evolutivo está influenciada por el nivel de autocorrelación serial existente. Cuando este nivel es positivo, indistintamente de la ventana móvil sobre la cual se calcula el método, las particiones sugeridas son las mismas. Sin embargo, esto puede cambiar cuando hay presencia de autocorrelación negativa.

Asimismo, se encuentra que, a pesar que el Índice de Calinski-Harabasz es la herramienta para definir las particiones, ocasionalmente el índice no alcanza un valor máximo absoluto, sino que a medida que la ventana móvil es ampliada, el valor del índice también lo es. Esto conlleva a subjetividades al momento de definir las particiones y representa una dificultad debido a que la calidad de particiones no podrían ser buenas.

A diferencia de ello, en el PCA Segmentado, la construcción de segmentos no es alterada por el tamaño de la ventana móvil sobre la cual se construye el PCA Evolutivo. Asimismo, si se observan distintos segmentos frente a diversas estructuras de datos, nivel de autocorrelación u otras características.

Entonces, el PCA Evolutivo es más volátil a la ventana de tiempo definida, ya que dependiendo de ello definirá la amplitud de plazos homogéneos, mientras que el PCA Segmentado es más sensible a la cantidad de dimensiones retenidas debido a que usa una función de costos (error de reconstrucción - distancia euclidiana entre el valor reducido y el valor original; o el $T^2_{deHotelling}$). Es decir, a medida que la cantidad de componentes principales retenidos aumenta, el error de reconstrucción (criterio para generar los segmentos) es menor, lo cual le da mayor robustez a los segmentos creados.

5.2. Sugerencia para investigaciones futuras

Para investigaciones futuras, es posible considerar otros criterios de clusterización adicionales al Índice de Calinski-Harabász, como K-Means, por ejemplo. Esto es para asegurar estabilidad en la segmentación, dado que el índice no está relacionado con la convergencia en si misma, sino con la calidad de los grupos definidos. Es decir, puede brindar información acerca de la definición de grupos y si están en mayor o menor medida correctamente definidos.

Esta alternativa sugiere extender el marco teórico abordado en esta investigación y permitiría robustecer el propuesta del PCA Evolutivo.

Apéndice A

Simulación 1

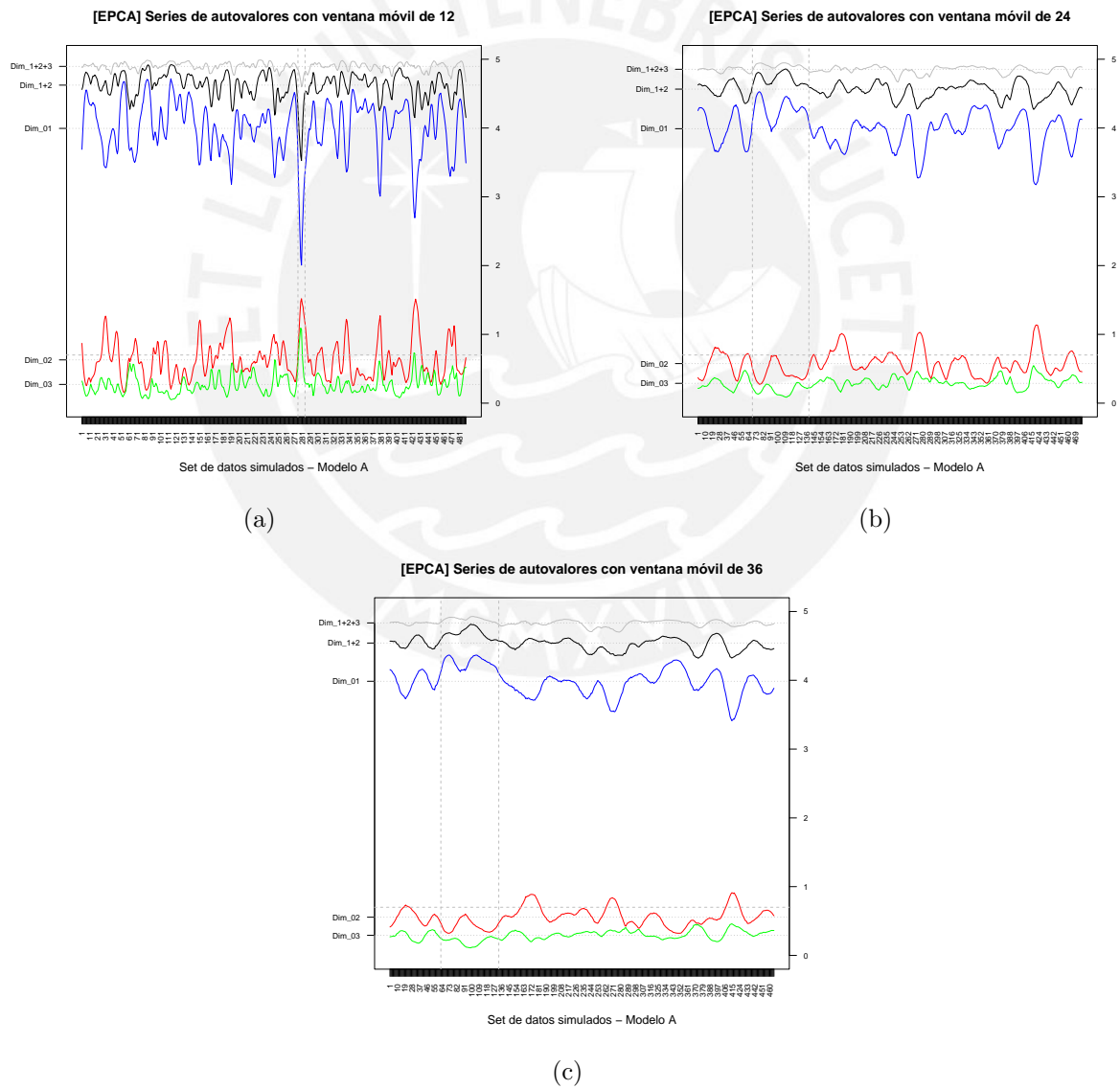


Figura A.1: Método PCA Evolutivo - Modelo A: Evolución del primer autovalor.

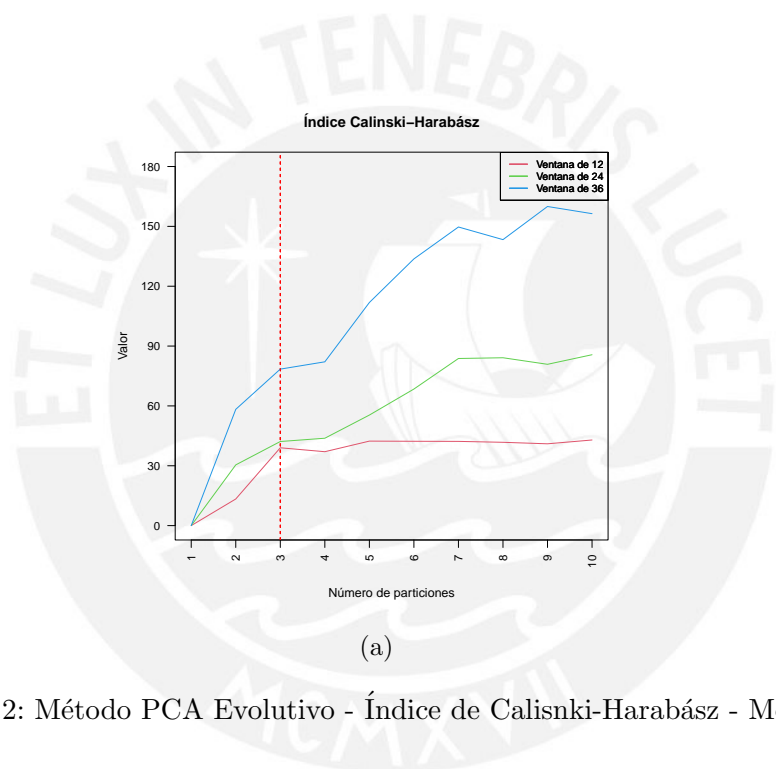


Figura A.2: Método PCA Evolutivo - Índice de Calinski-Harabász - Modelo A

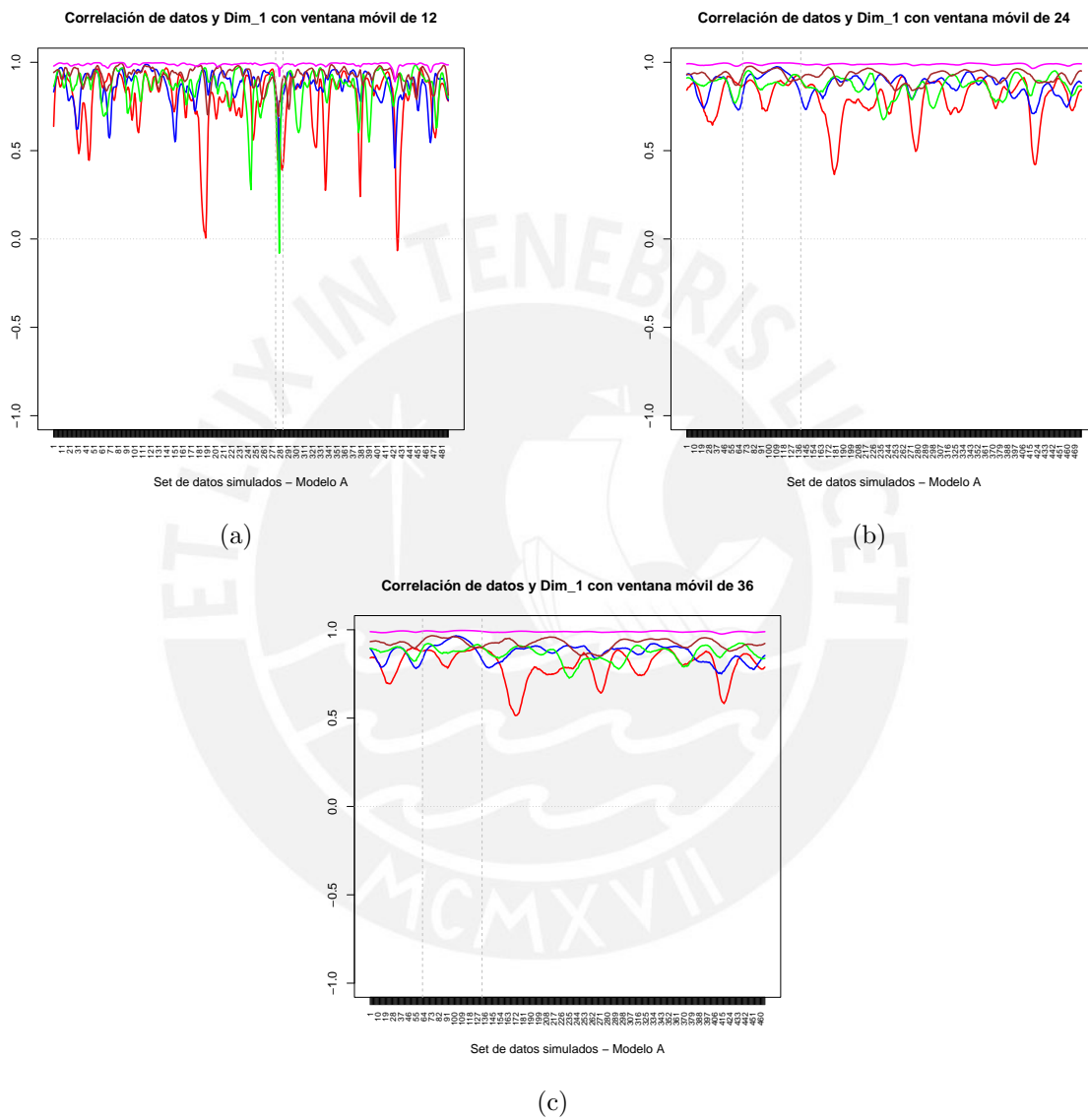


Figura A.3: Método PCA Evolutivo - Modelo A: Correlación del primer autovalor con las variables simuladas.

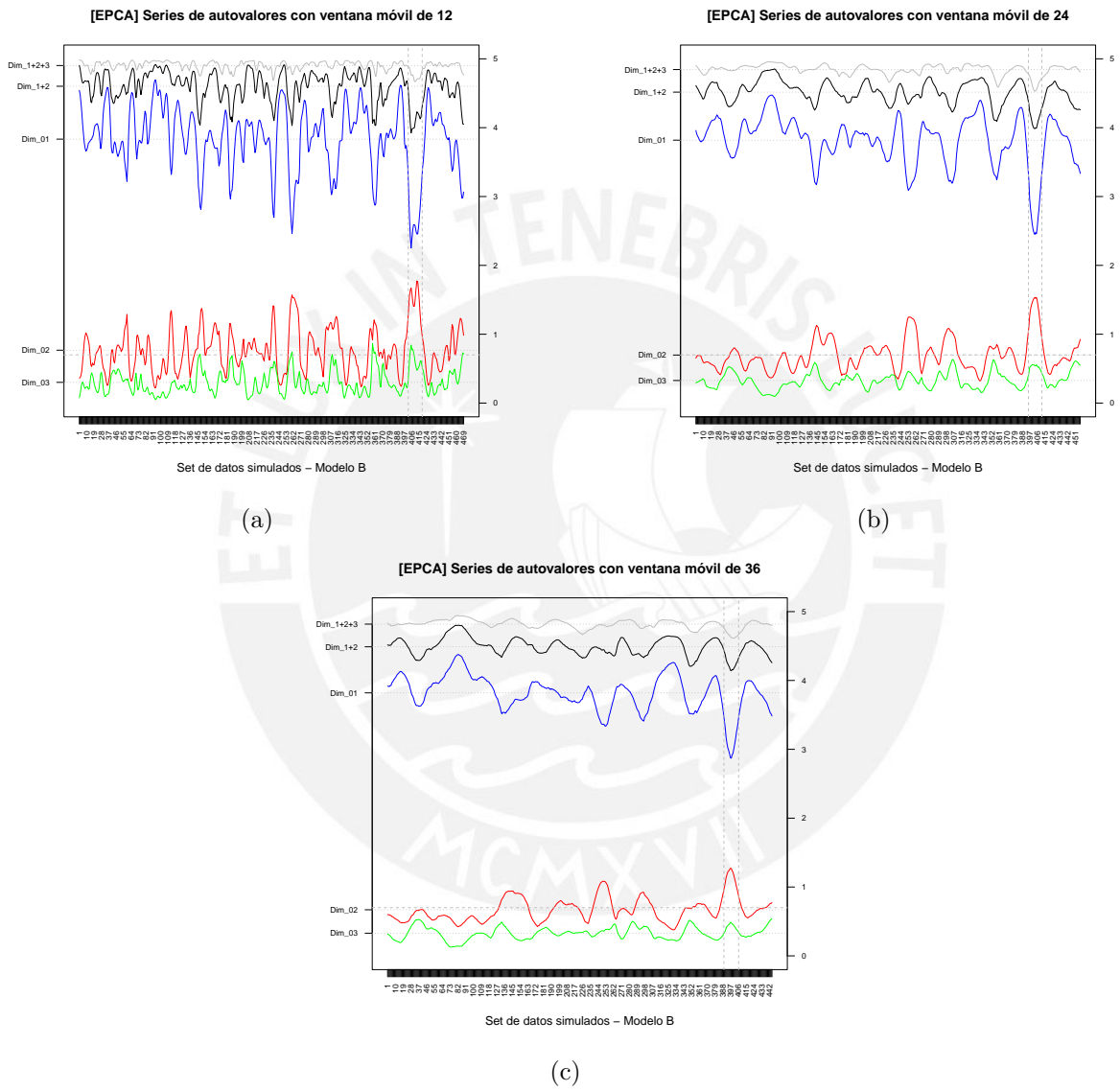


Figura A.4: Método PCA Evolutivo - Modelo B: Evolución del primer autovalor.

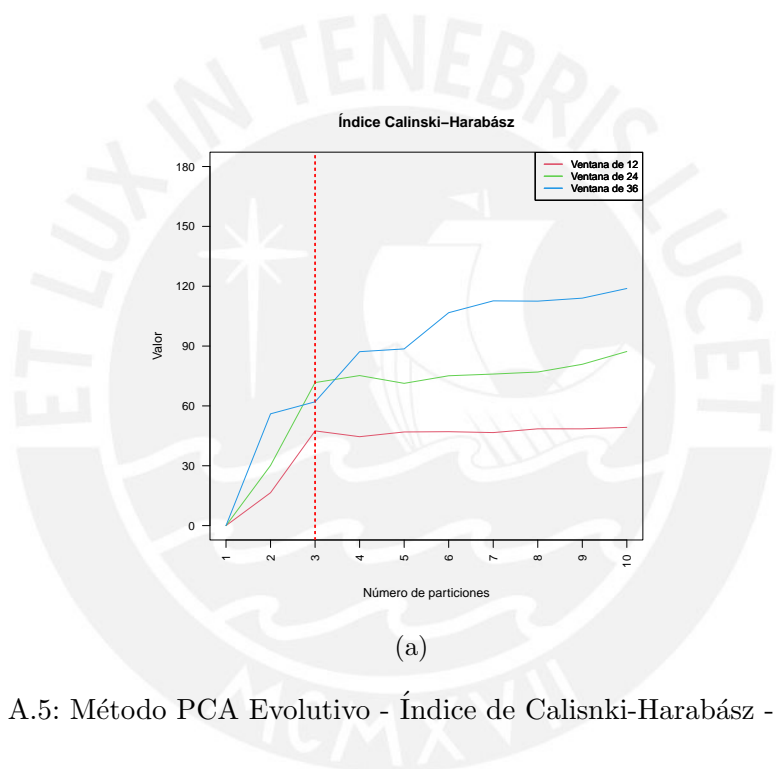


Figura A.5: Método PCA Evolutivo - Índice de Calinski-Harabász - Modelo B

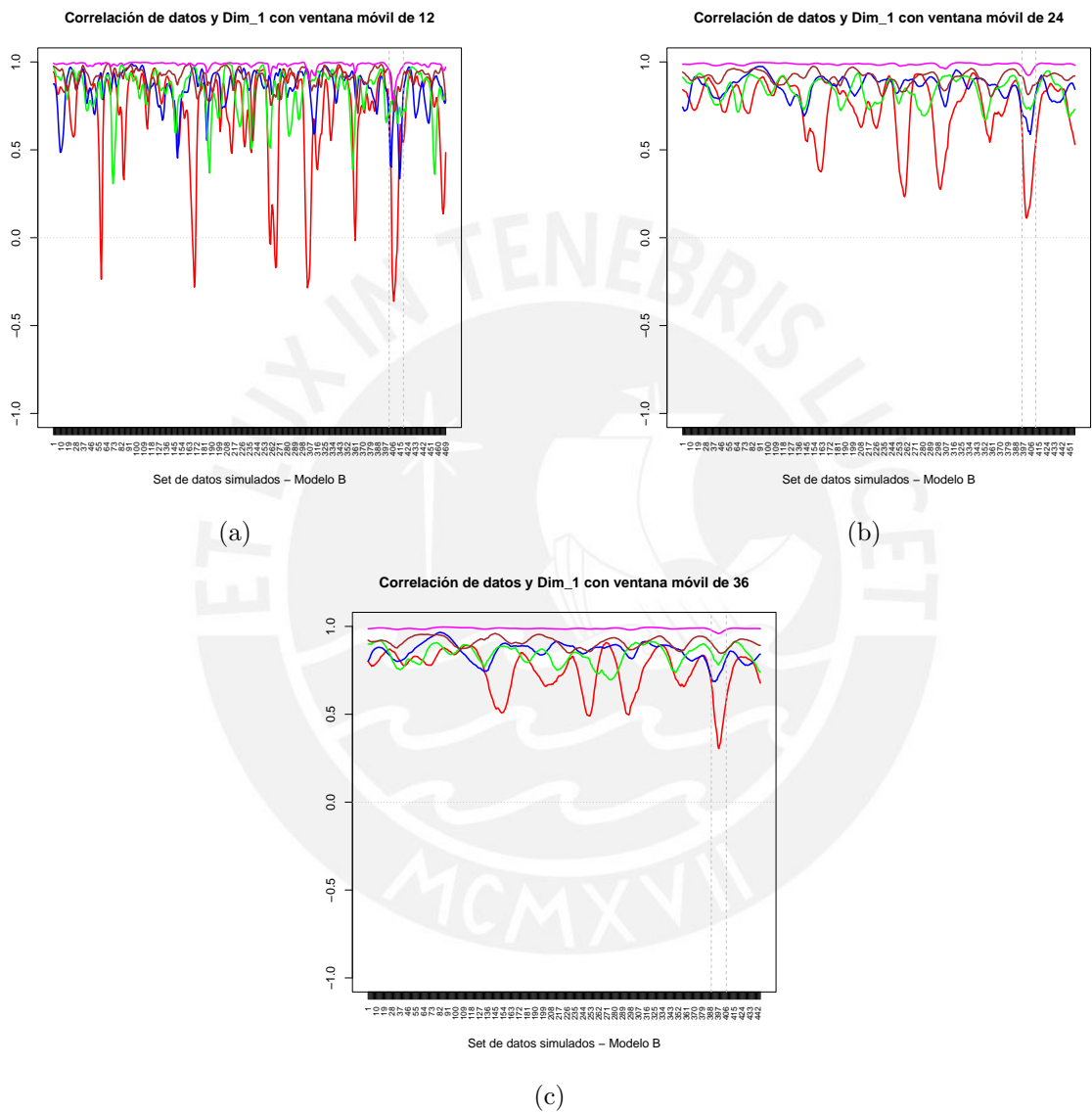


Figura A.6: Método PCA Evolutivo - Modelo B: Correlación del primer autovalor con las variables simuladas.



Figura A.7: Método PCA Evolutivo - Modelo B4: Evolución del primer autovalor.

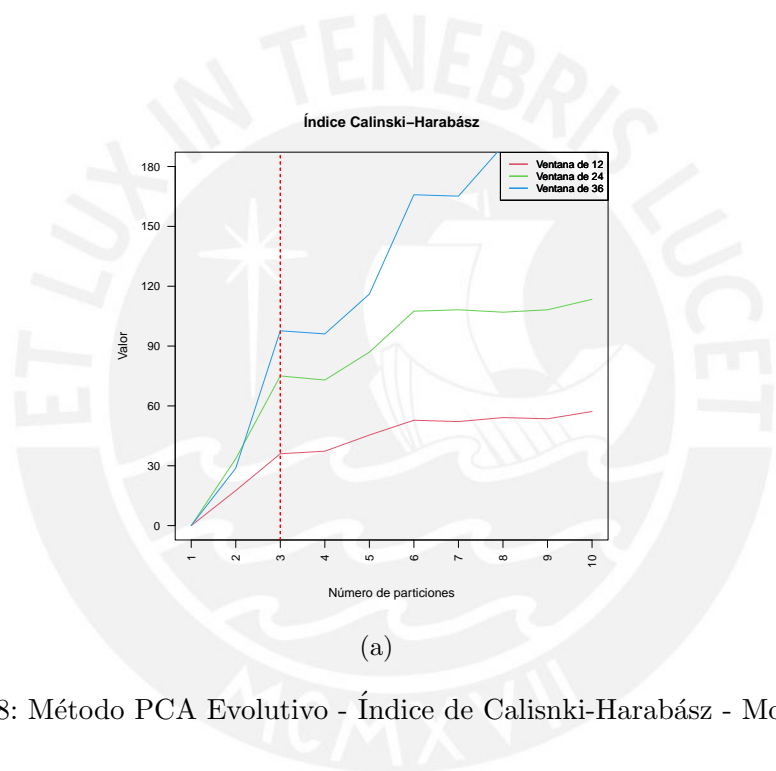


Figura A.8: Método PCA Evolutivo - Índice de Calinski-Harabász - Modelo B4

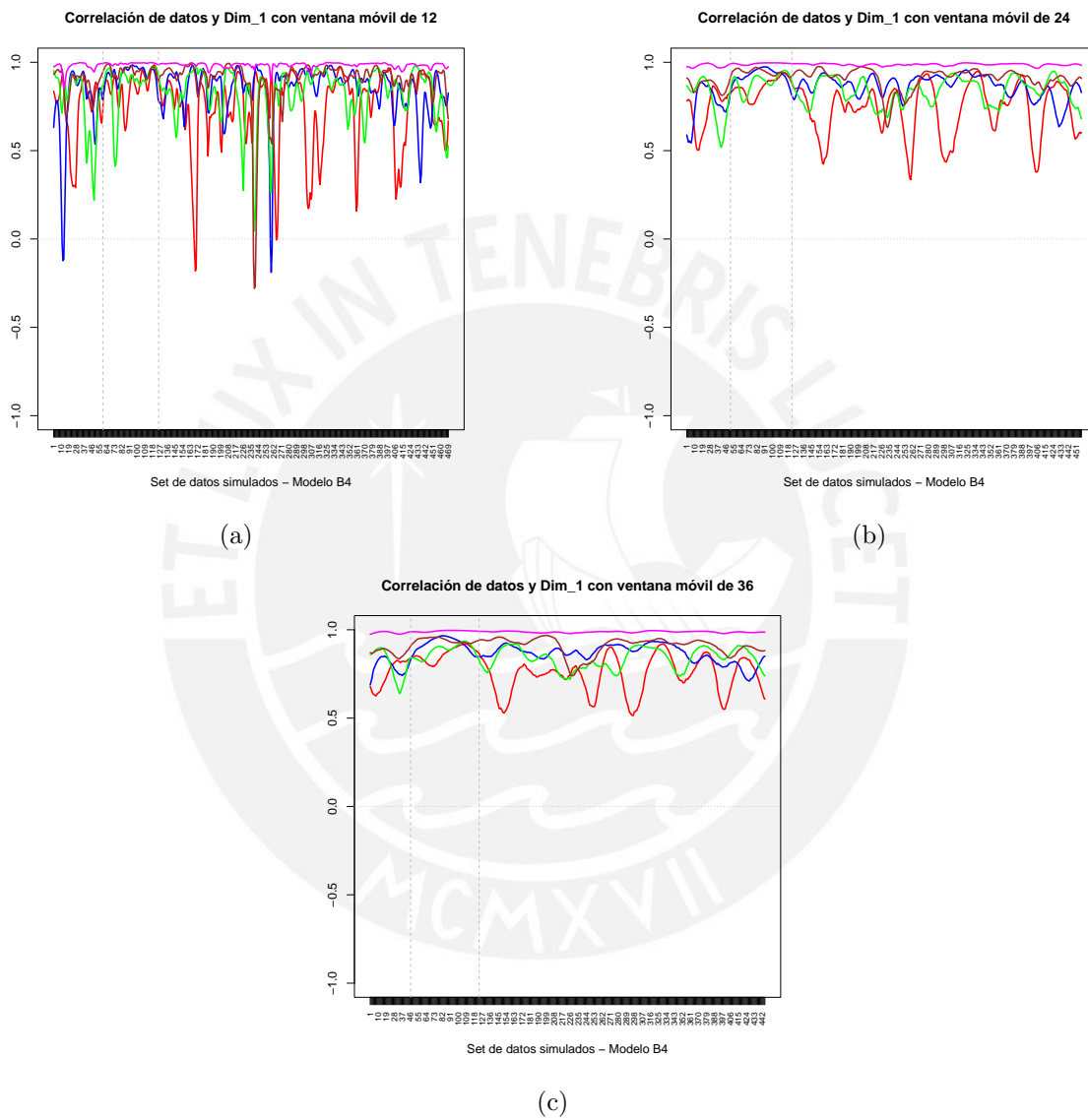


Figura A.9: Método PCA Evolutivo - Modelo B4: Correlación del primer autovalor con las variables simuladas.



Figura A.10: Método PCA Evolutivo - Modelo B2: Evolución del primer autovalor.

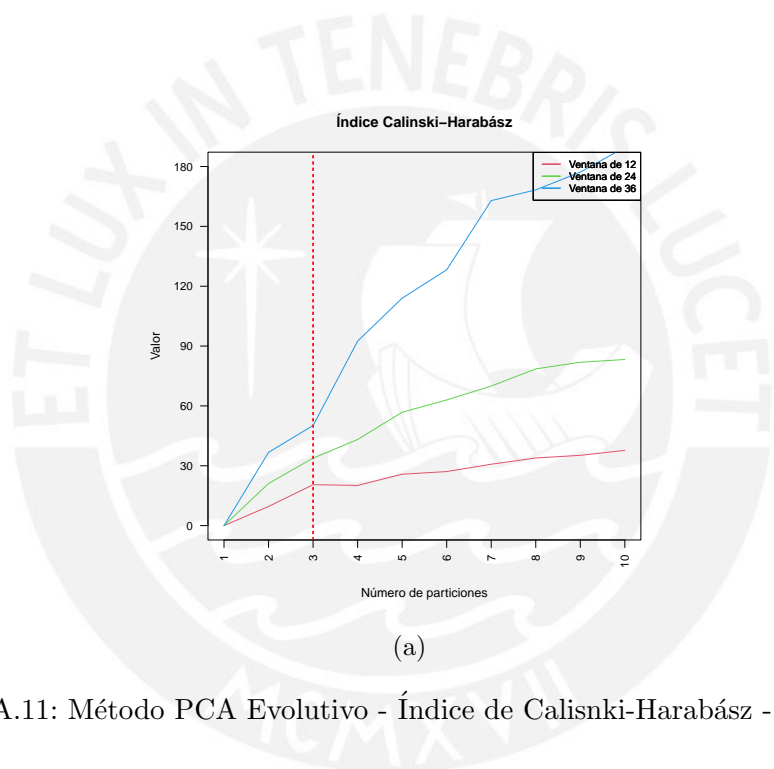


Figura A.11: Método PCA Evolutivo - Índice de Calinski-Harabász - Modelo B2

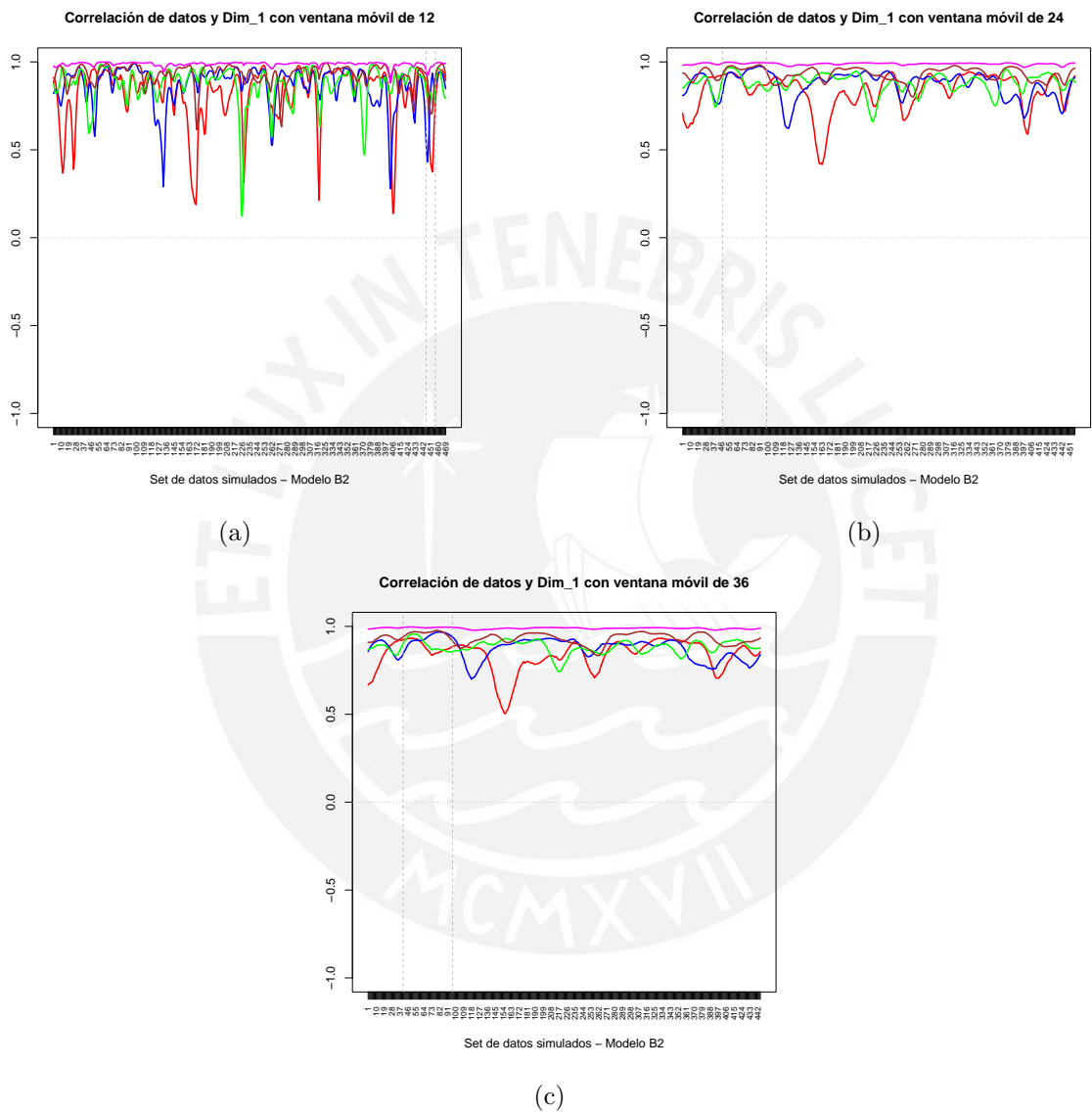


Figura A.12: Método PCA Evolutivo - Modelo B2: Correlación del primer autovalor con las variables simuladas

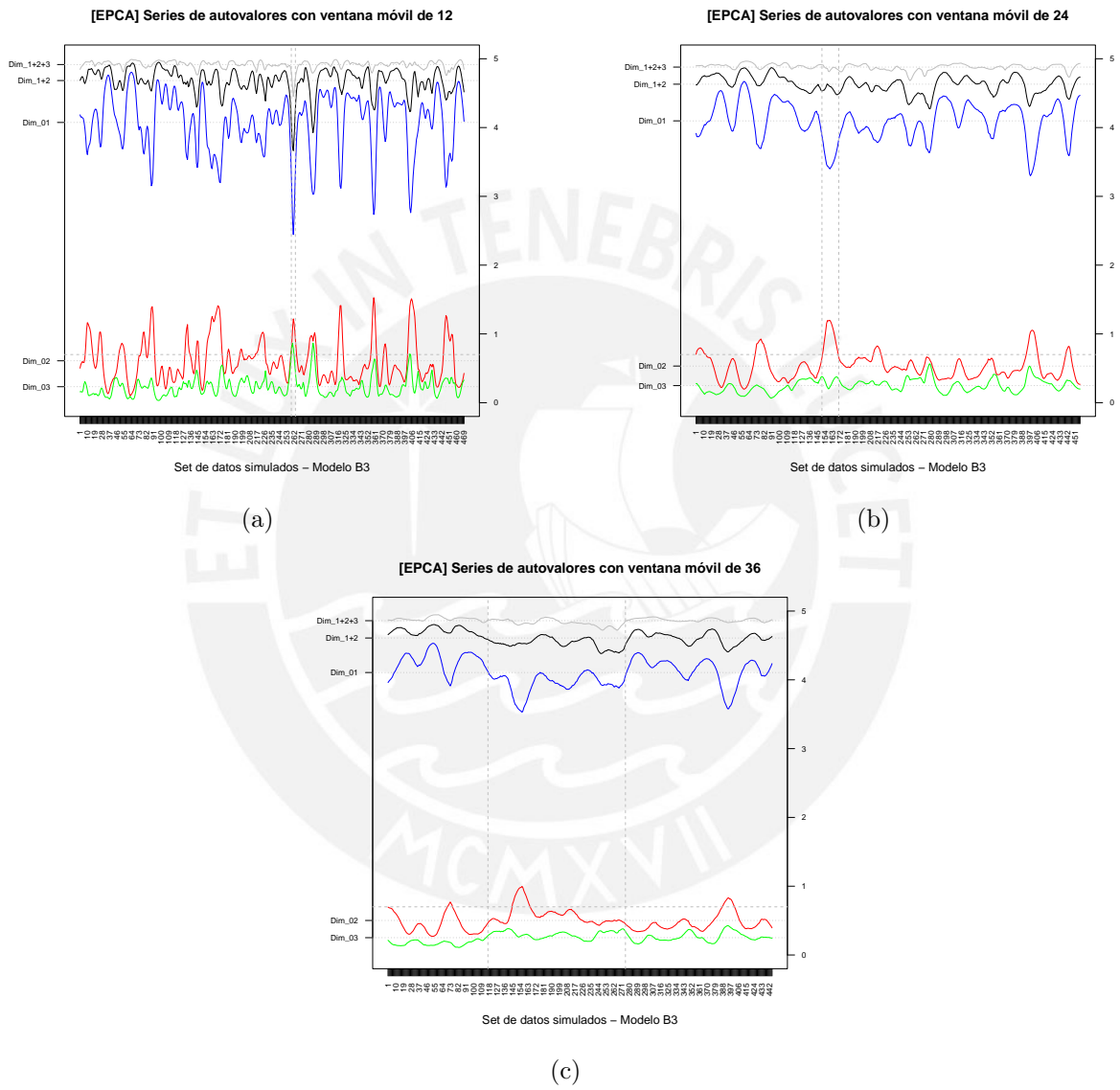


Figura A.13: Método PCA Evolutivo - Modelo B3: Evolución del primer autovalor.

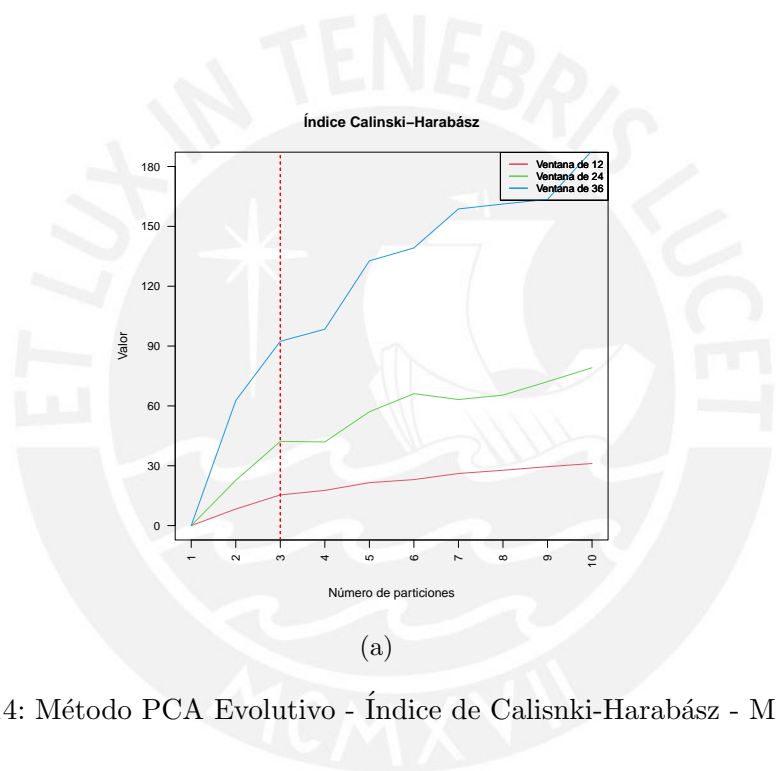
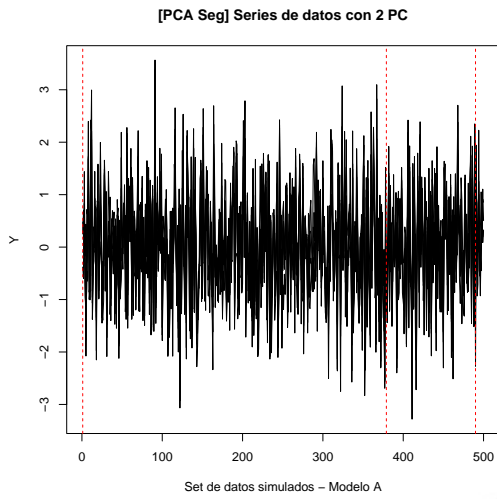


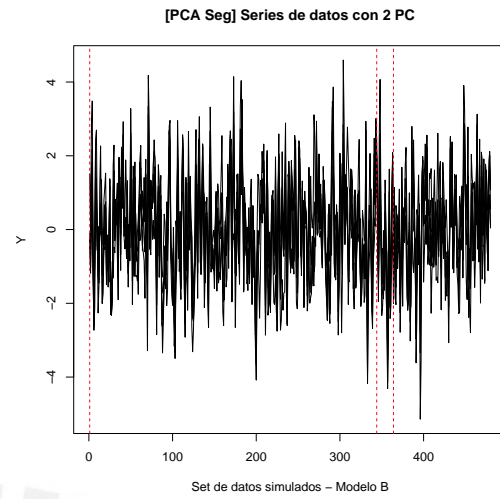
Figura A.14: Método PCA Evolutivo - Índice de Calinski-Harabász - Modelo B3



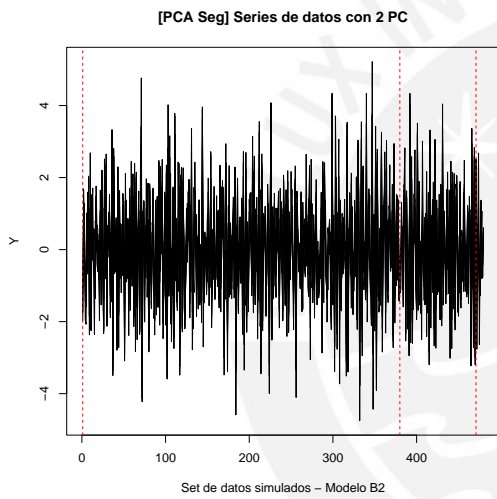
Figura A.15: Método PCA Evolutivo - Modelo B3: Correlación del primer autovalor con las variables simuladas.



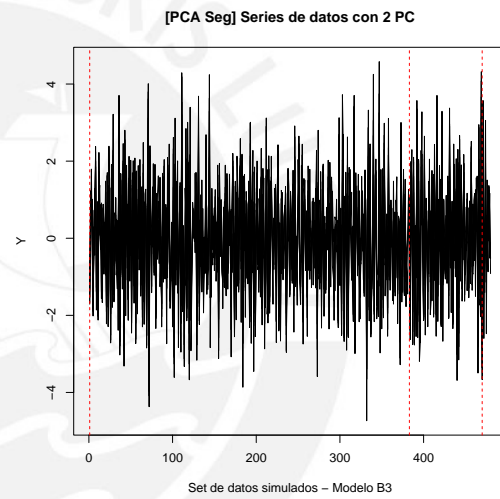
(a)



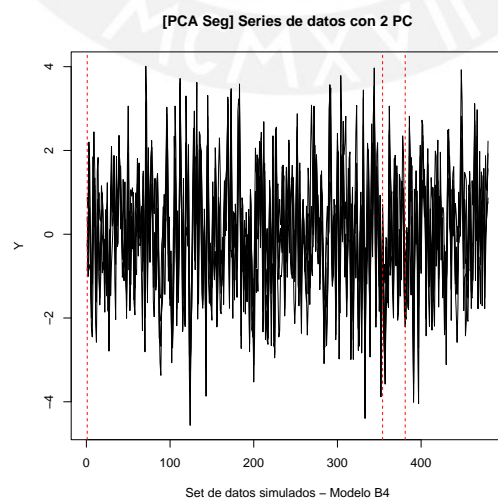
(b)



(c)



(d)



(e)

Figura A.16: Método PCA Segmentado para los modelos de la Simulación 1

Simulación 2

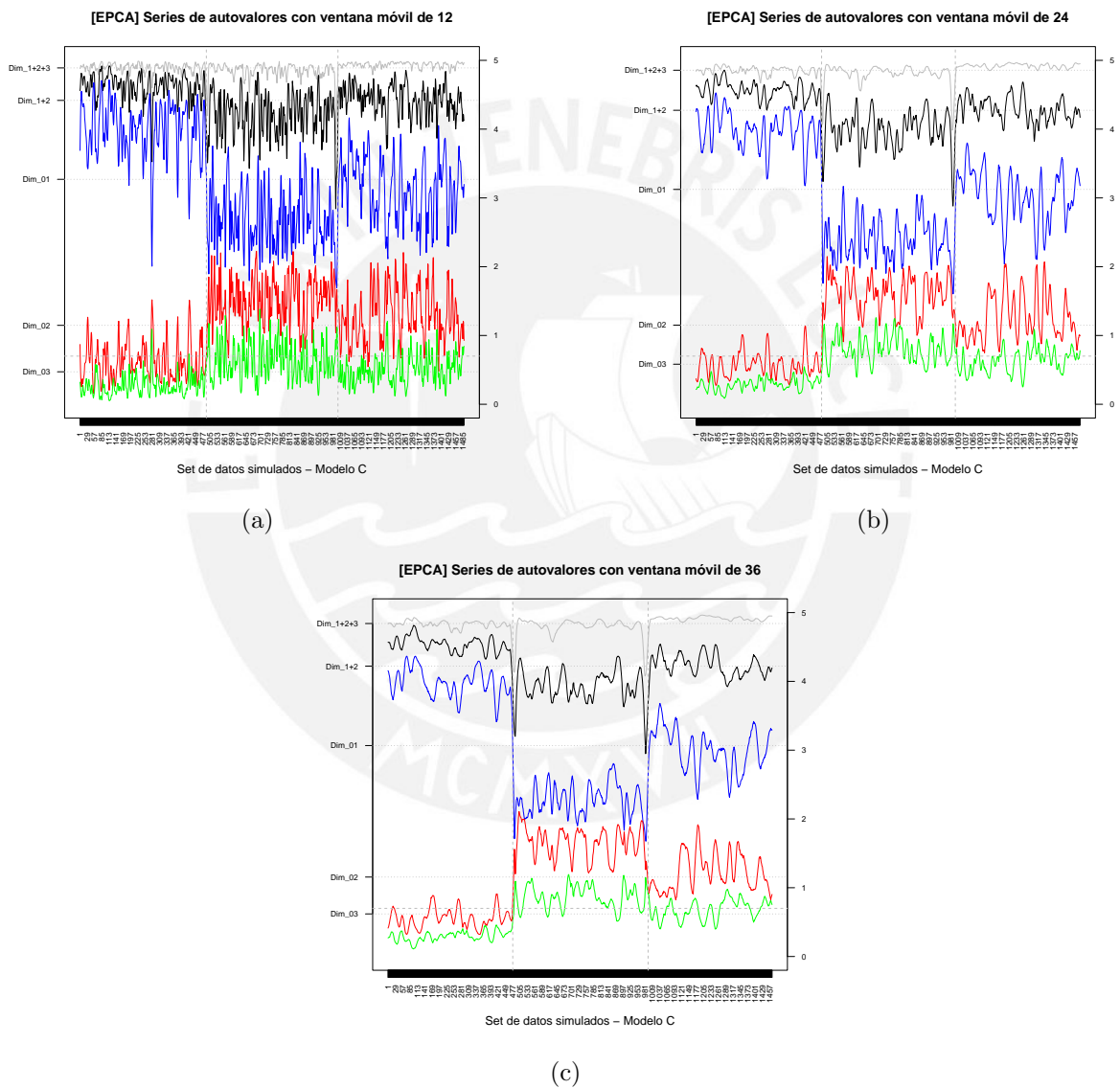


Figura A.17: Método PCA Evolutivo - Modelo C: Evolución del primer autovalor

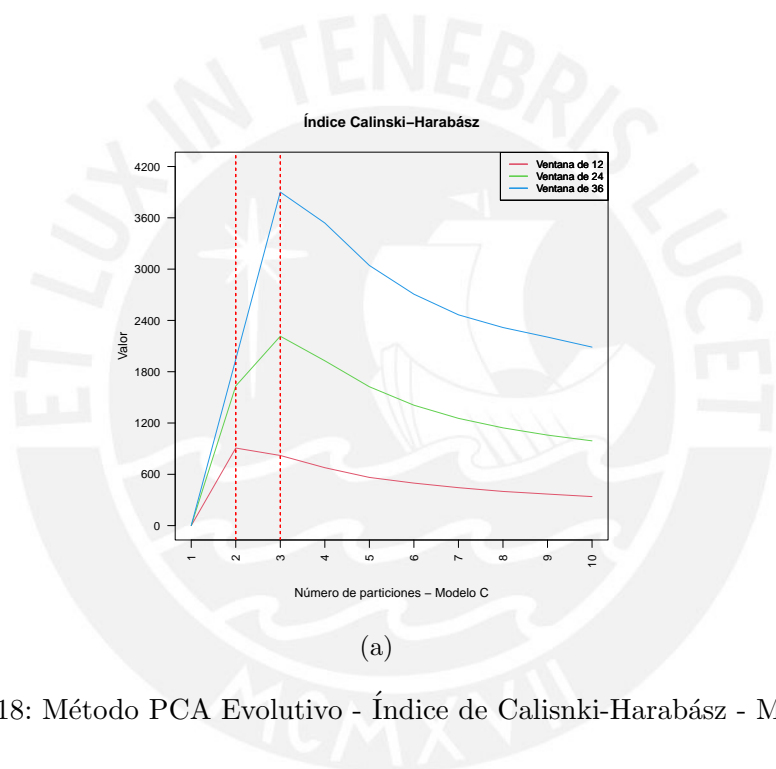


Figura A.18: Método PCA Evolutivo - Índice de Calinski-Harabász - Modelo C

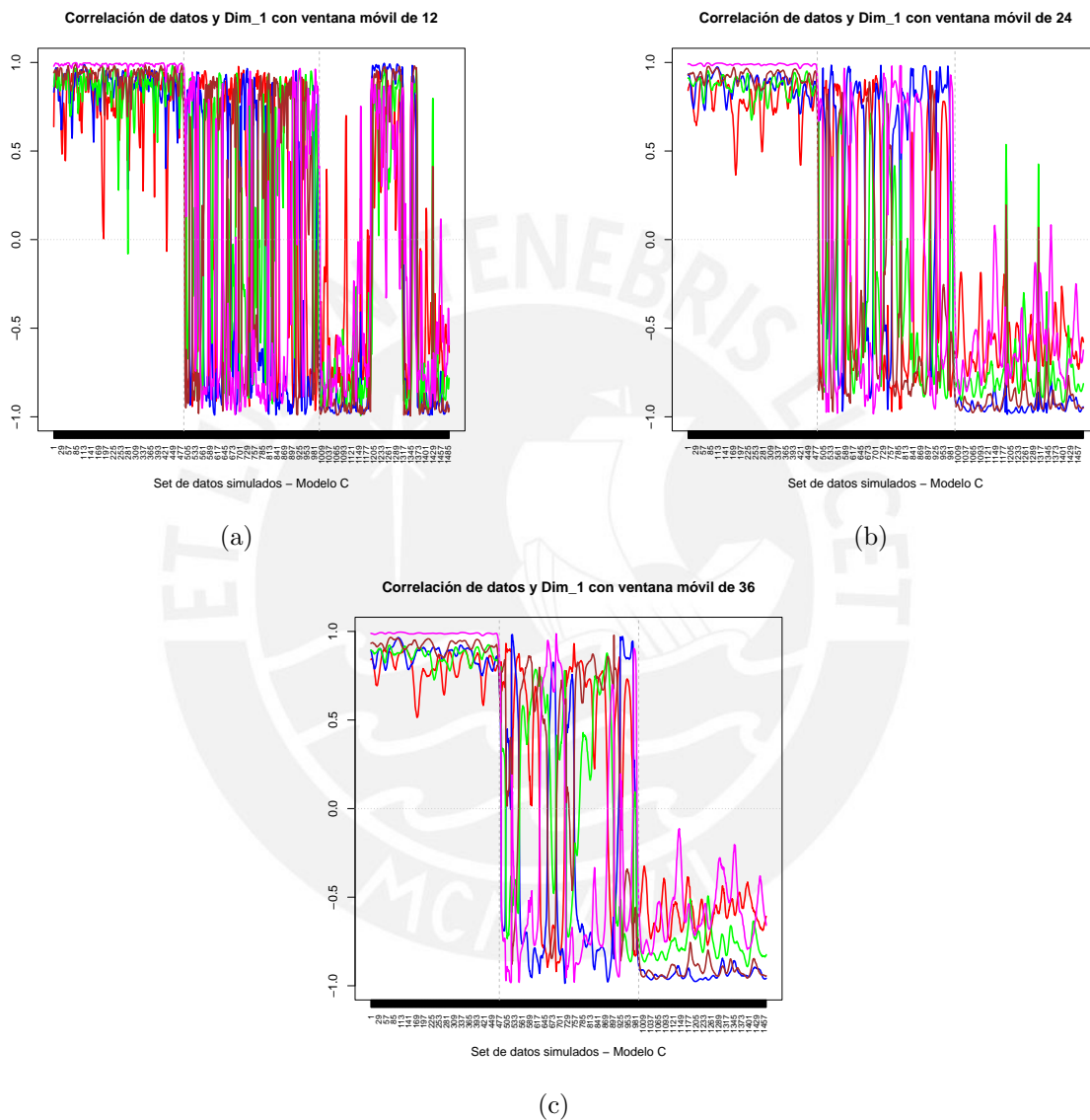


Figura A.19: Método PCA Evolutivo - Modelo C: Correlación del primer autovalor con las variables simuladas

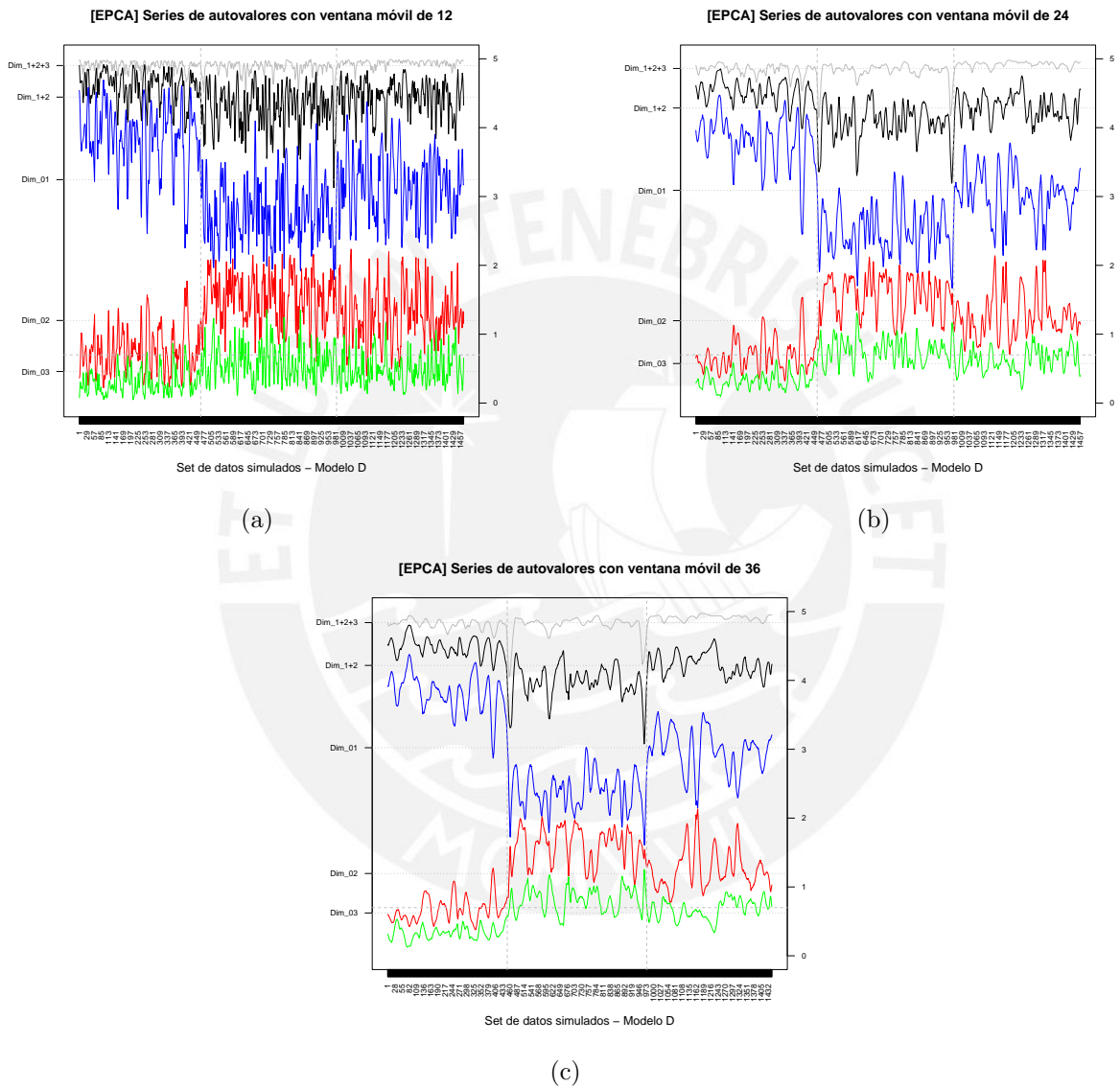


Figura A.20: Método PCA Evolutivo - Modelo D: Evolución del primer autovalor

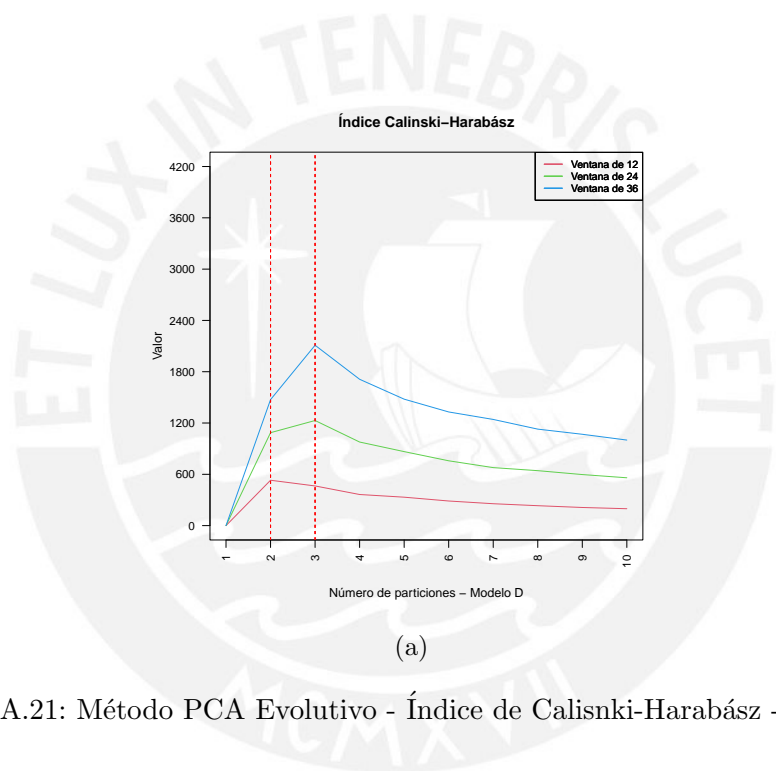


Figura A.21: Método PCA Evolutivo - Índice de Calinski-Harabász - Modelo D

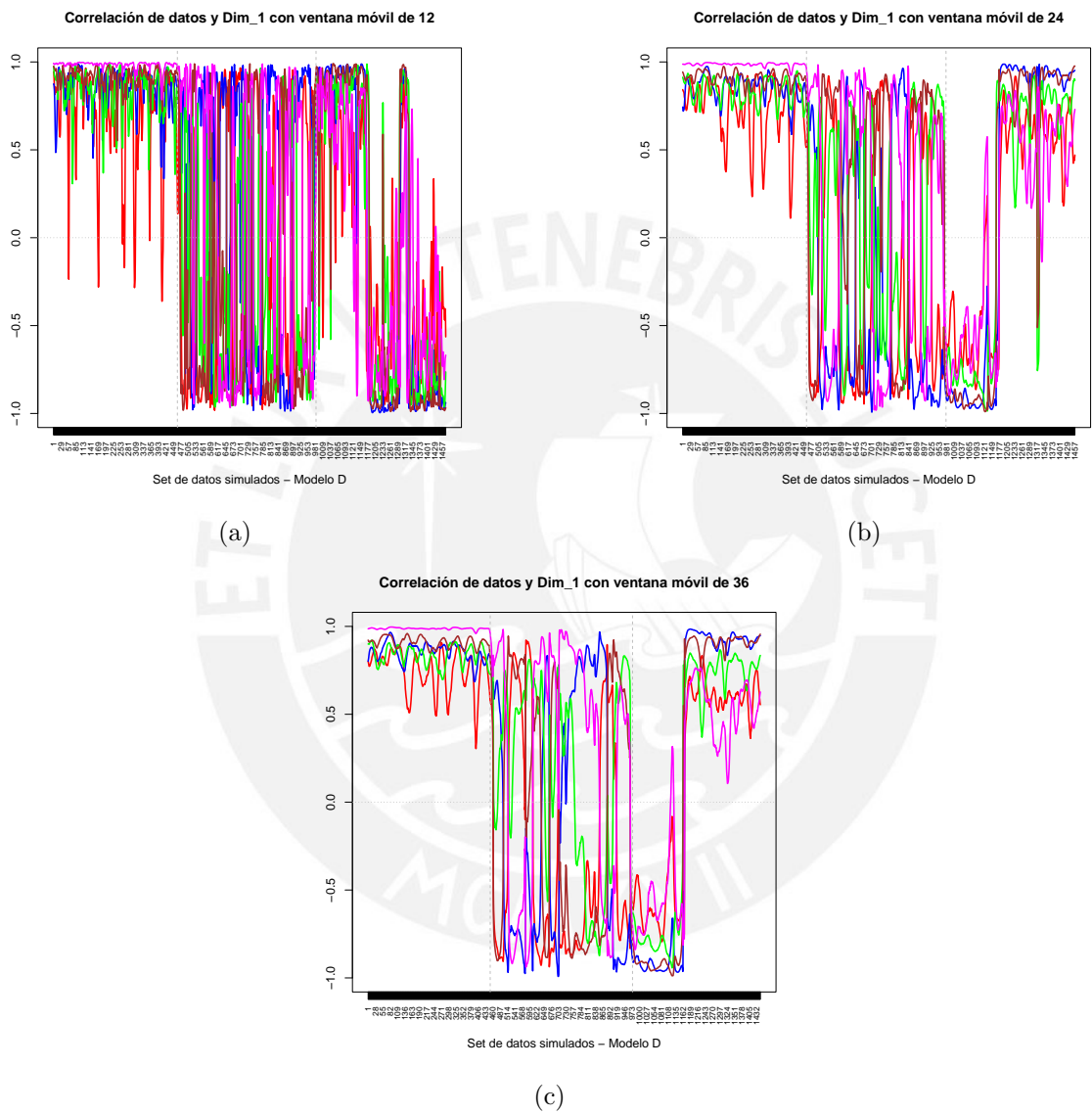


Figura A.22: Método PCA Evolutivo - Modelo D: Correlación del primer autovalor con las variables simuladas

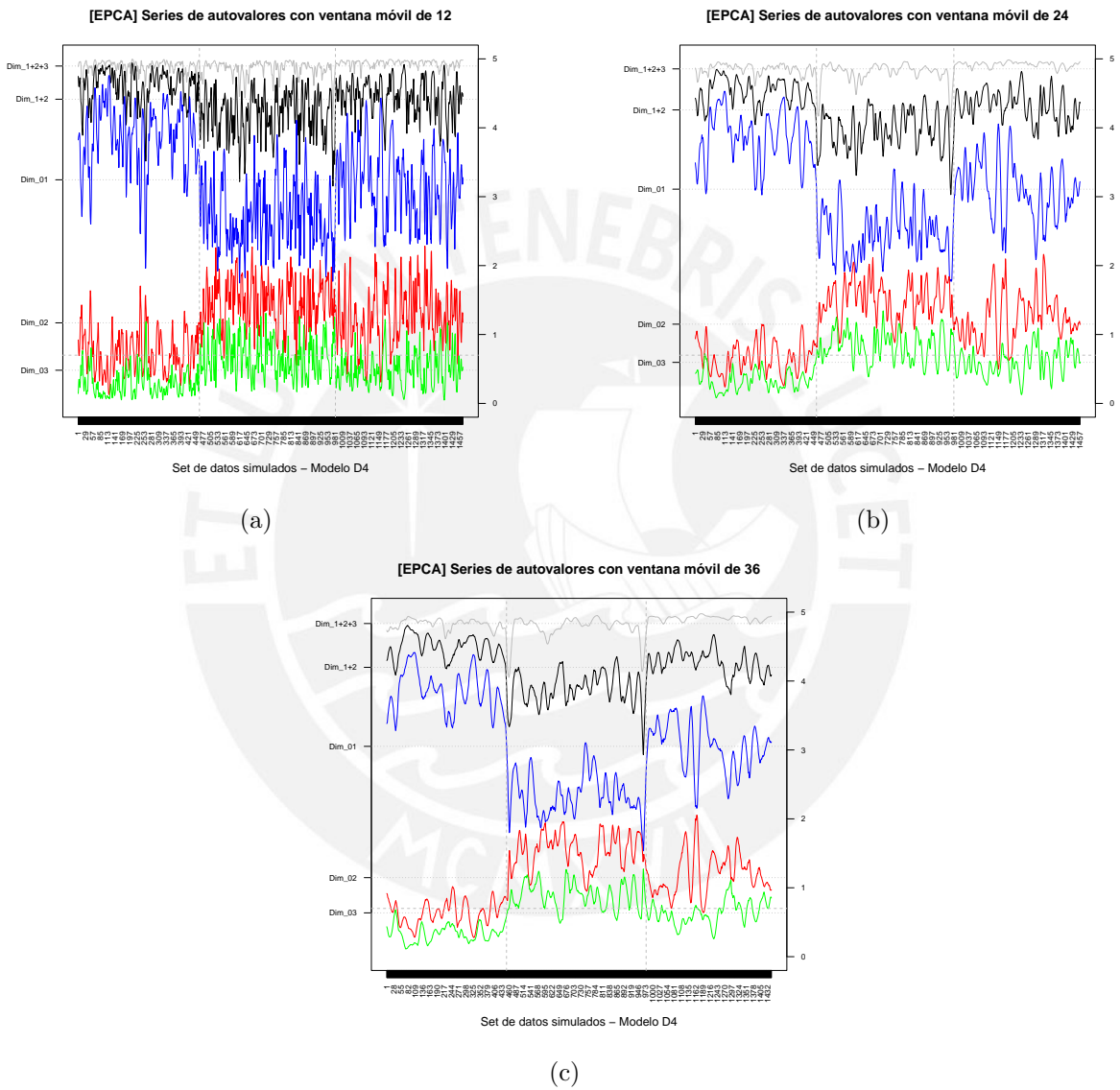


Figura A.23: Método PCA Evolutivo - Modelo D4: Evolución del primer autovalor

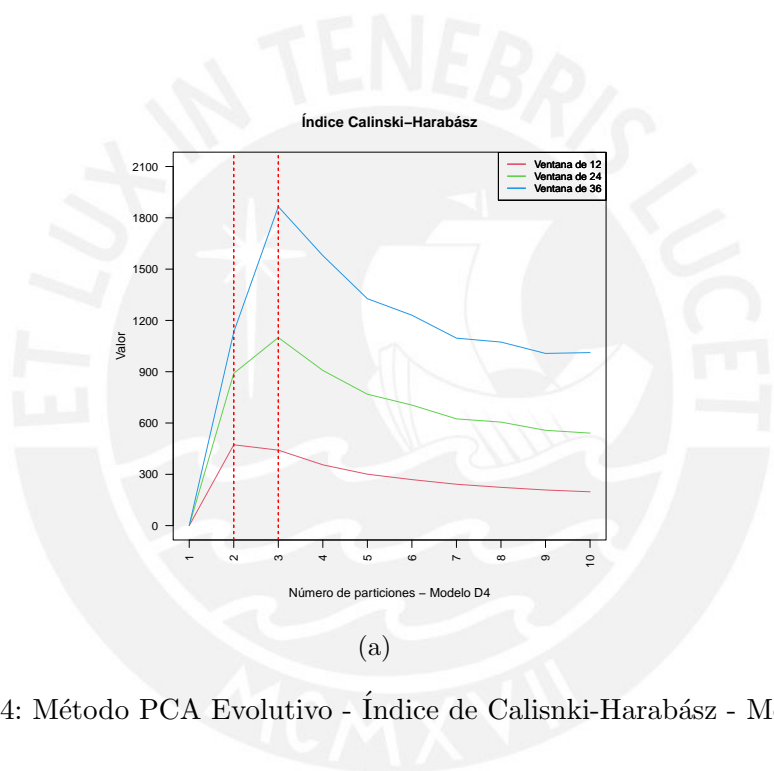


Figura A.24: Método PCA Evolutivo - Índice de Calinski-Harabász - Modelo D4

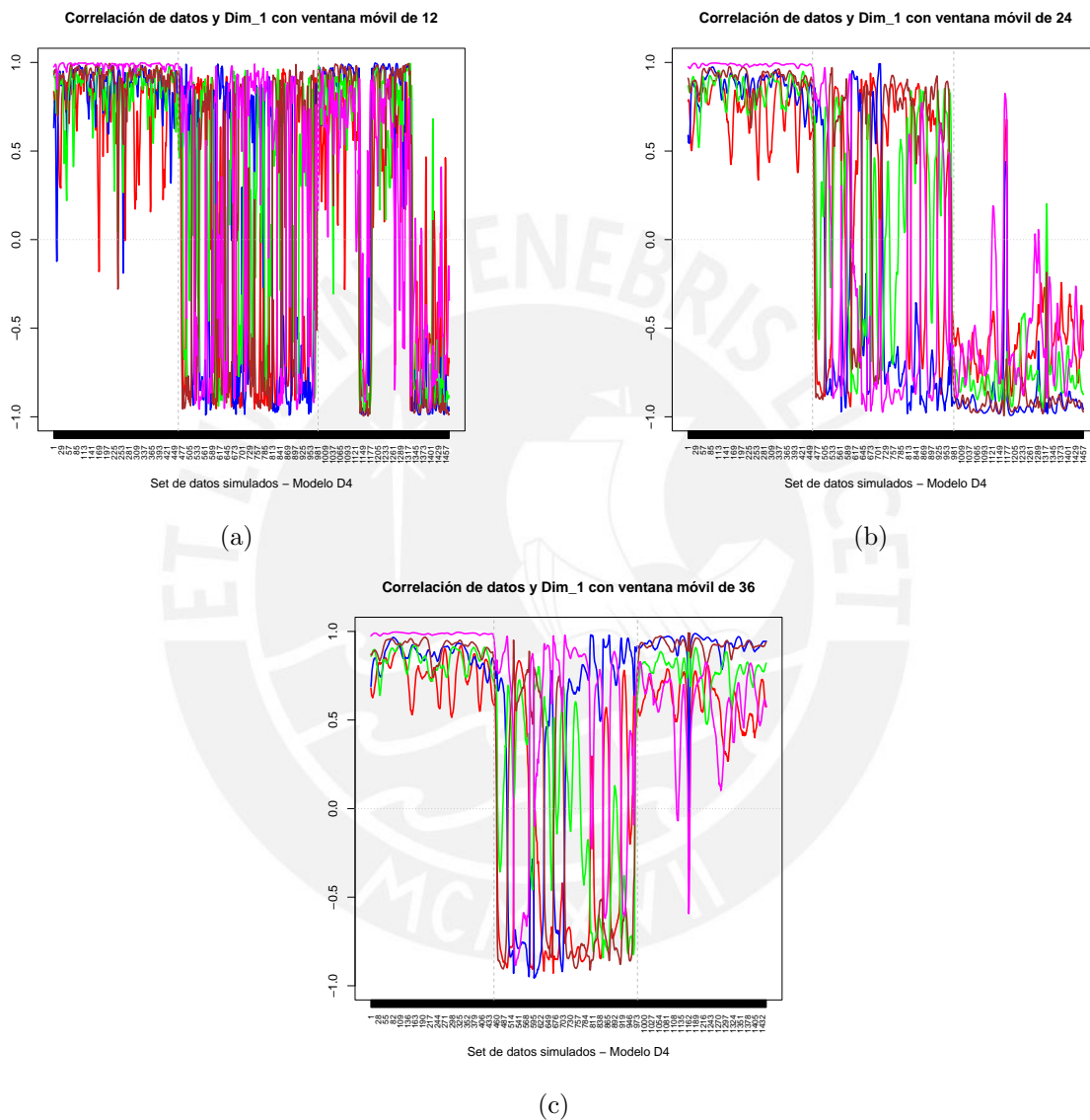


Figura A.25: Método PCA Evolutivo - Modelo D4: Correlación del primer autovalor con las variables simuladas.

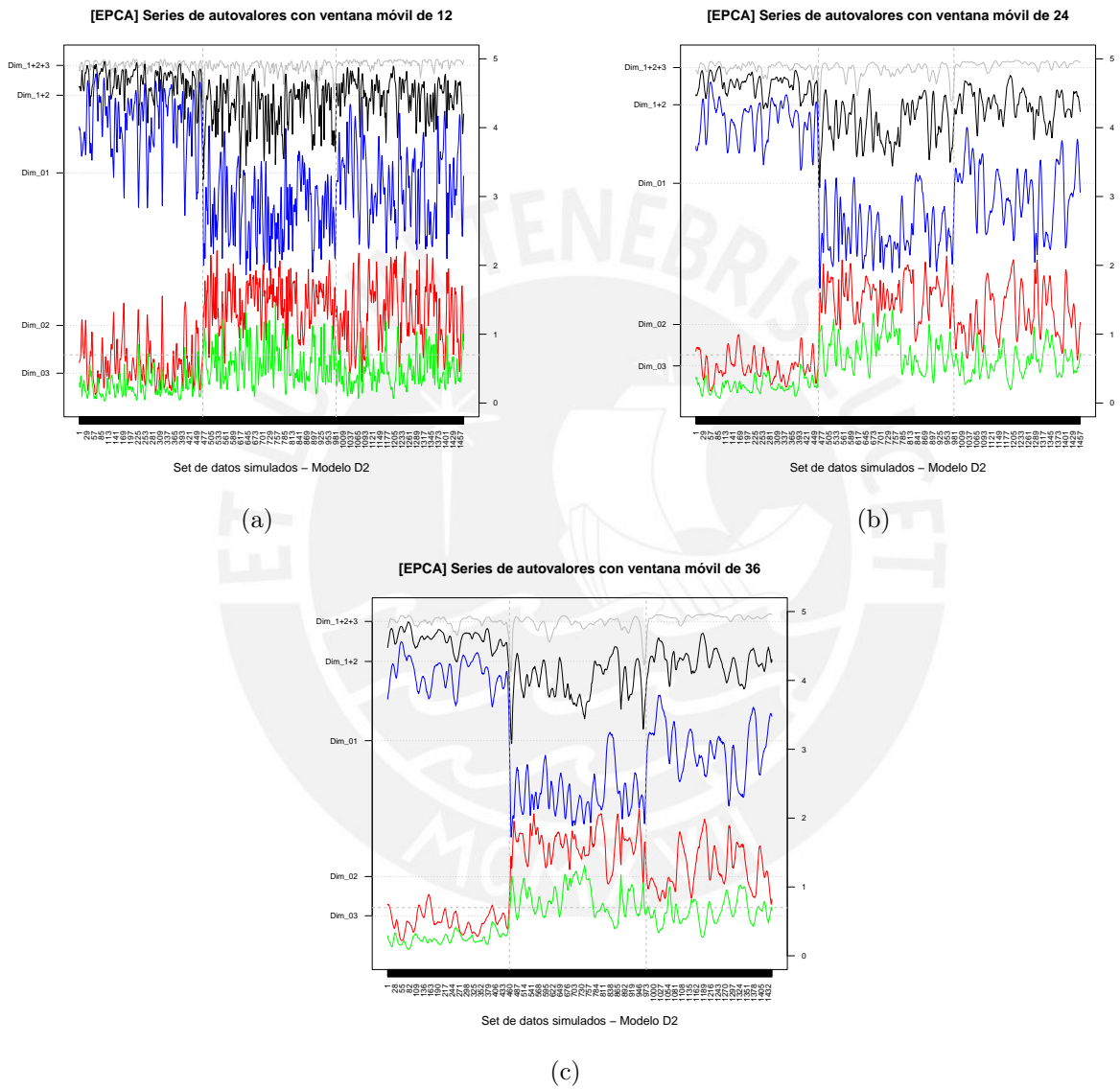


Figura A.26: Método PCA Evolutivo - Modelo D2: Evolución del primer autovalor.

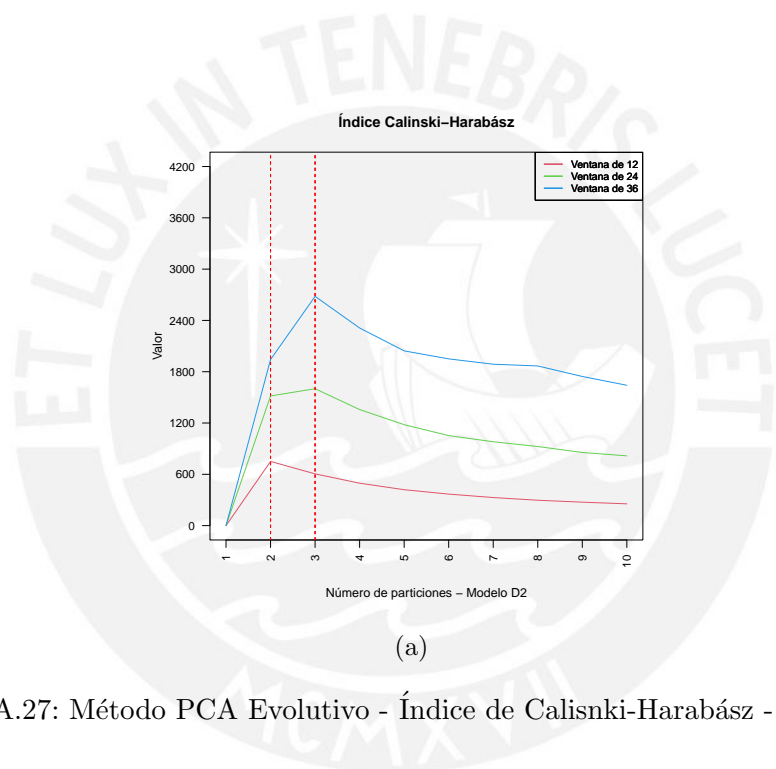


Figura A.27: Método PCA Evolutivo - Índice de Calinski-Harabász - Modelo D2

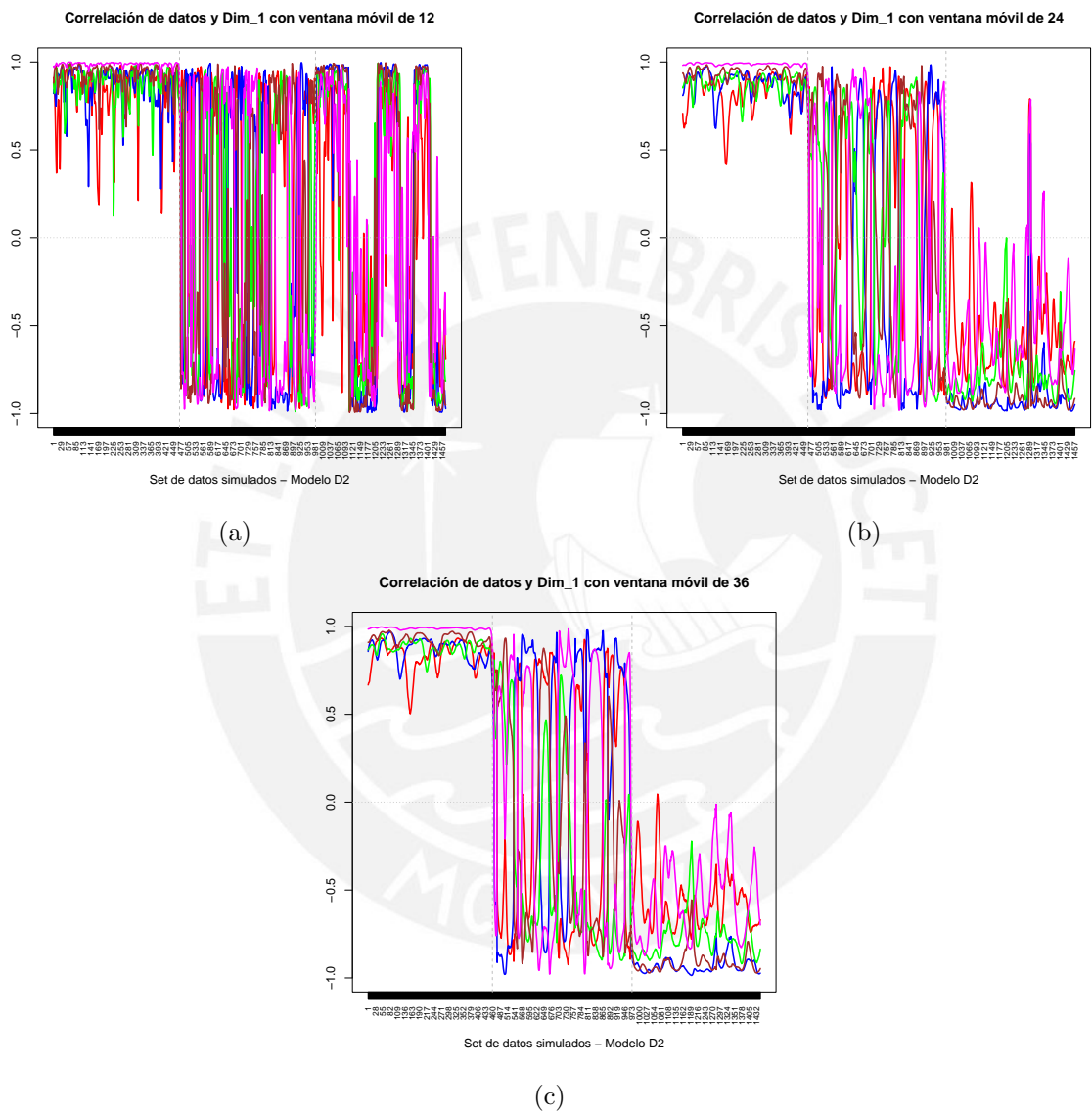


Figura A.28: Método PCA Evolutivo - Modelo D2: Correlación del primer autovalor con las variables simuladas.

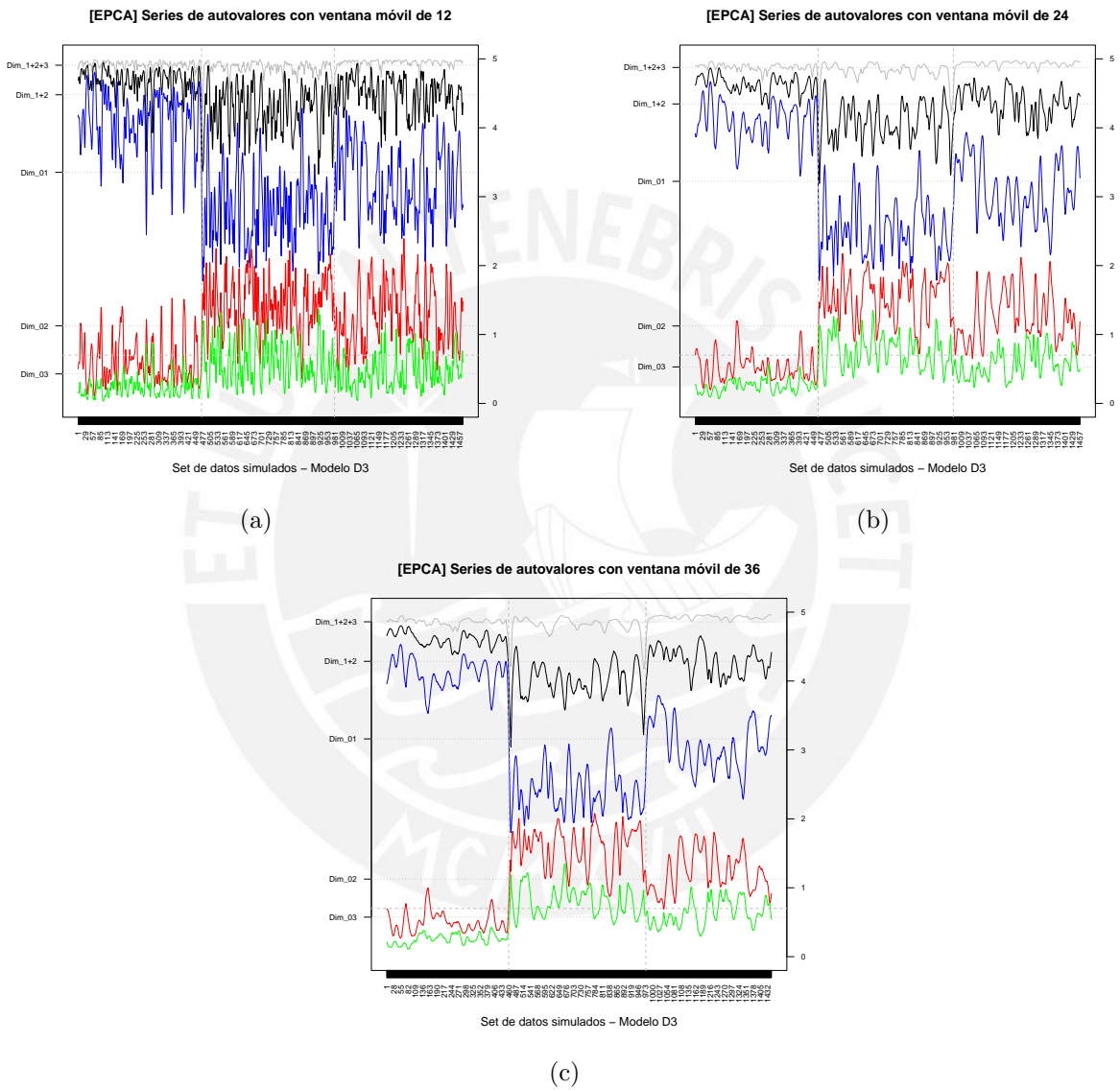


Figura A.29: Método PCA Evolutivo - Modelo D3: Evolución del primer autovalor.

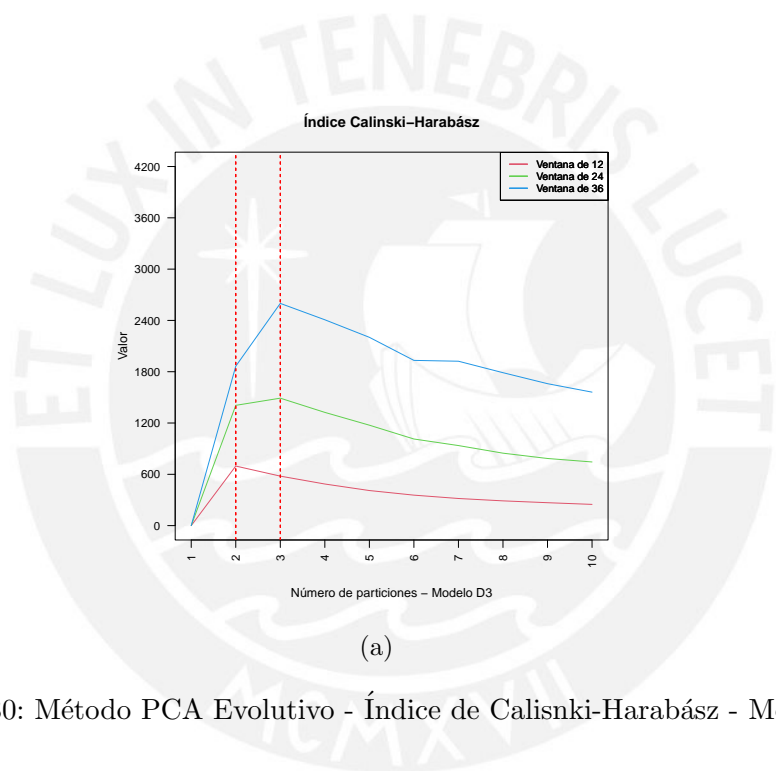


Figura A.30: Método PCA Evolutivo - Índice de Calinski-Harabász - Modelo D3

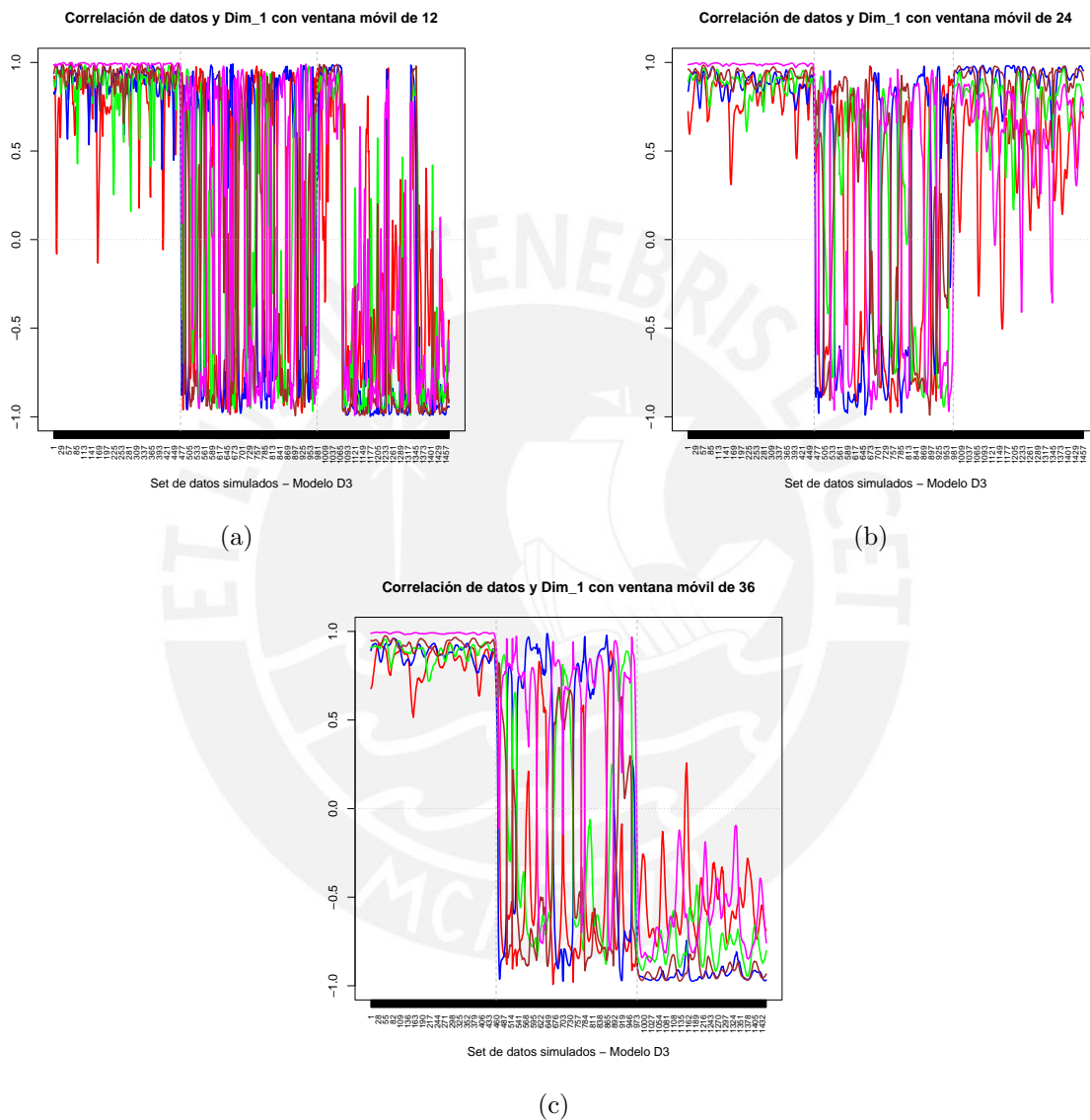


Figura A.31: Método PCA Evolutivo - Modelo D3: Correlación del primer autovalor con las variables simuladas.

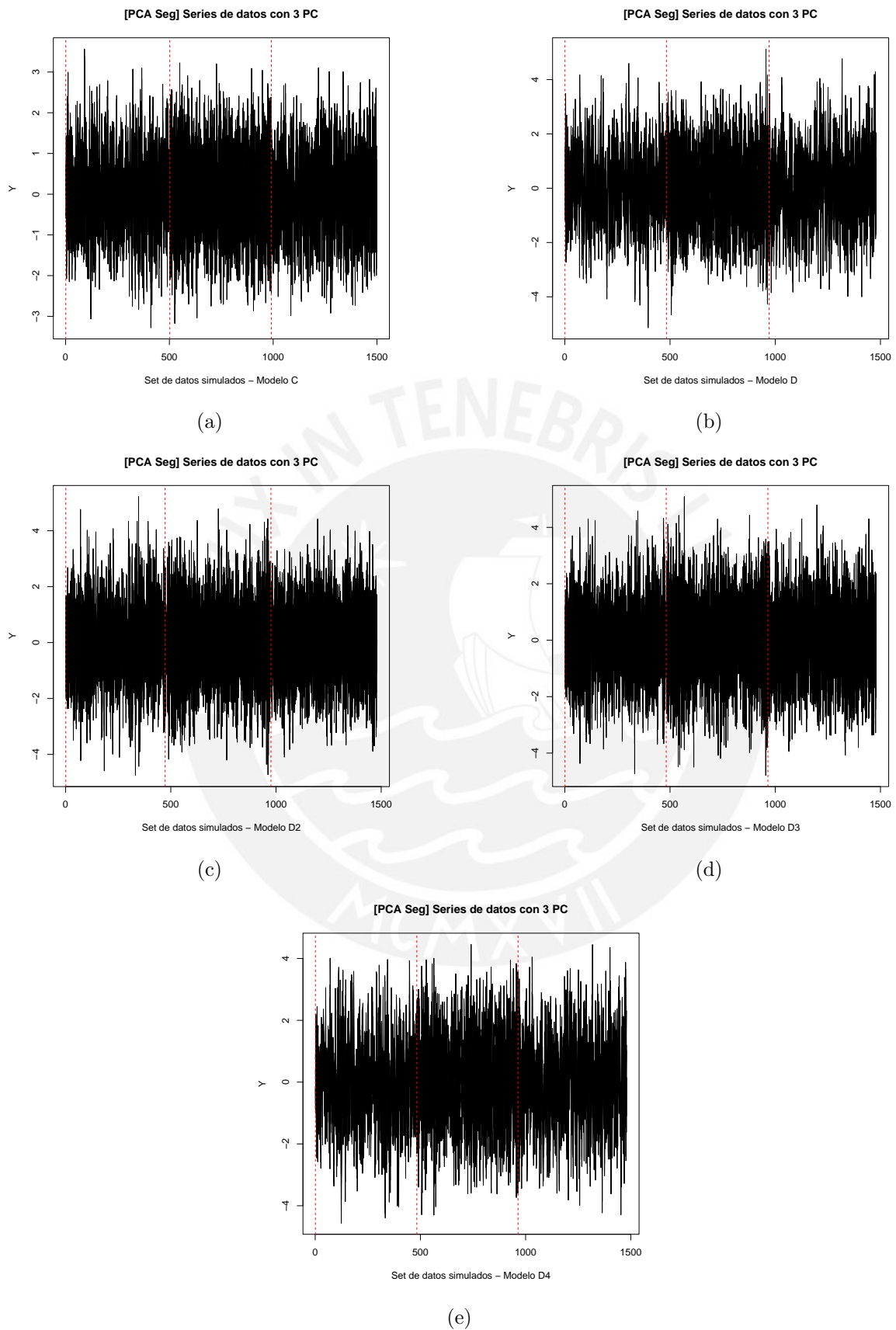


Figura A.32: Método PCA Segmentado para los modelos de la Simulación 2

Aplicación a datos de materias primas

primas

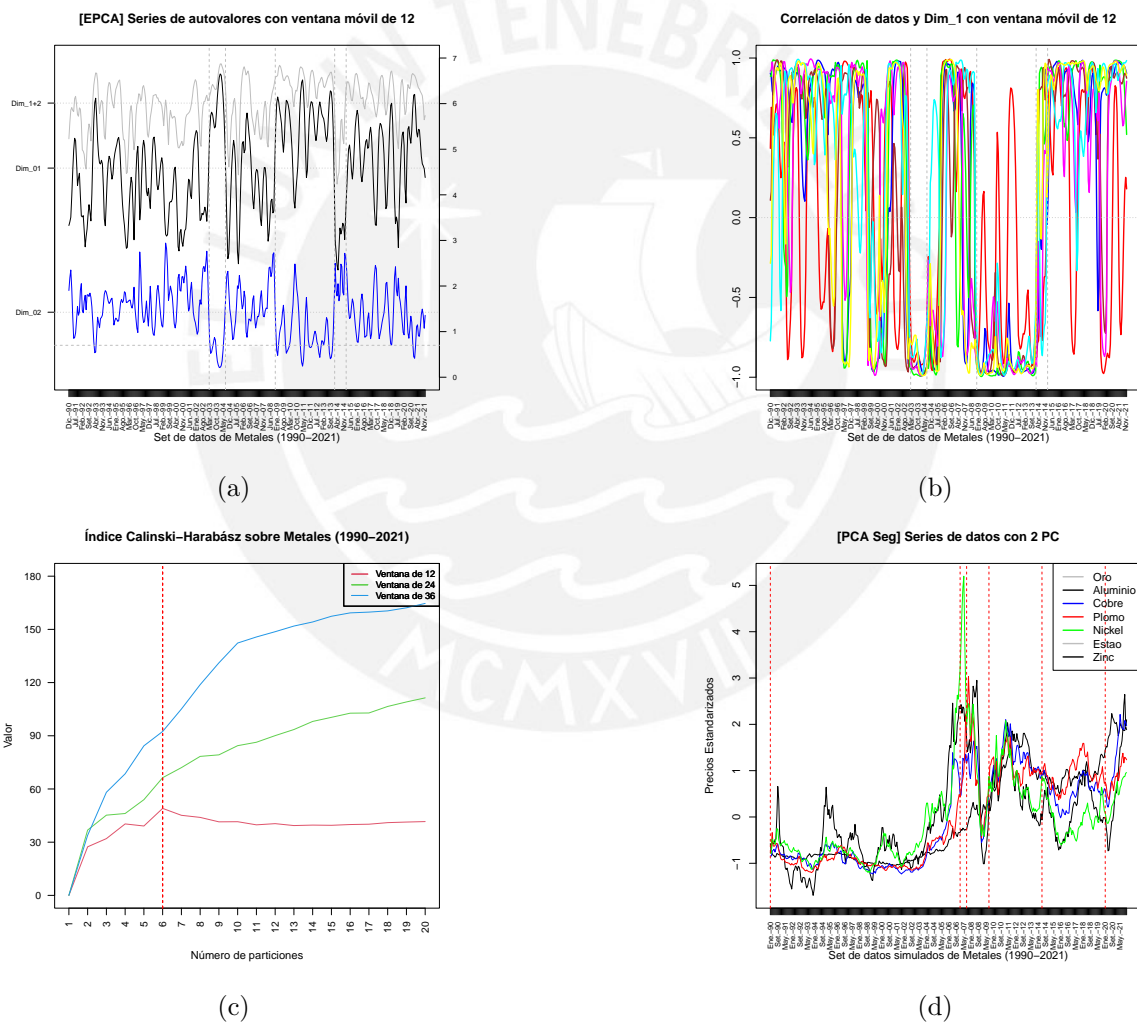


Figura A.33: Método PCA Evolutivo - Materias Primas de Metales

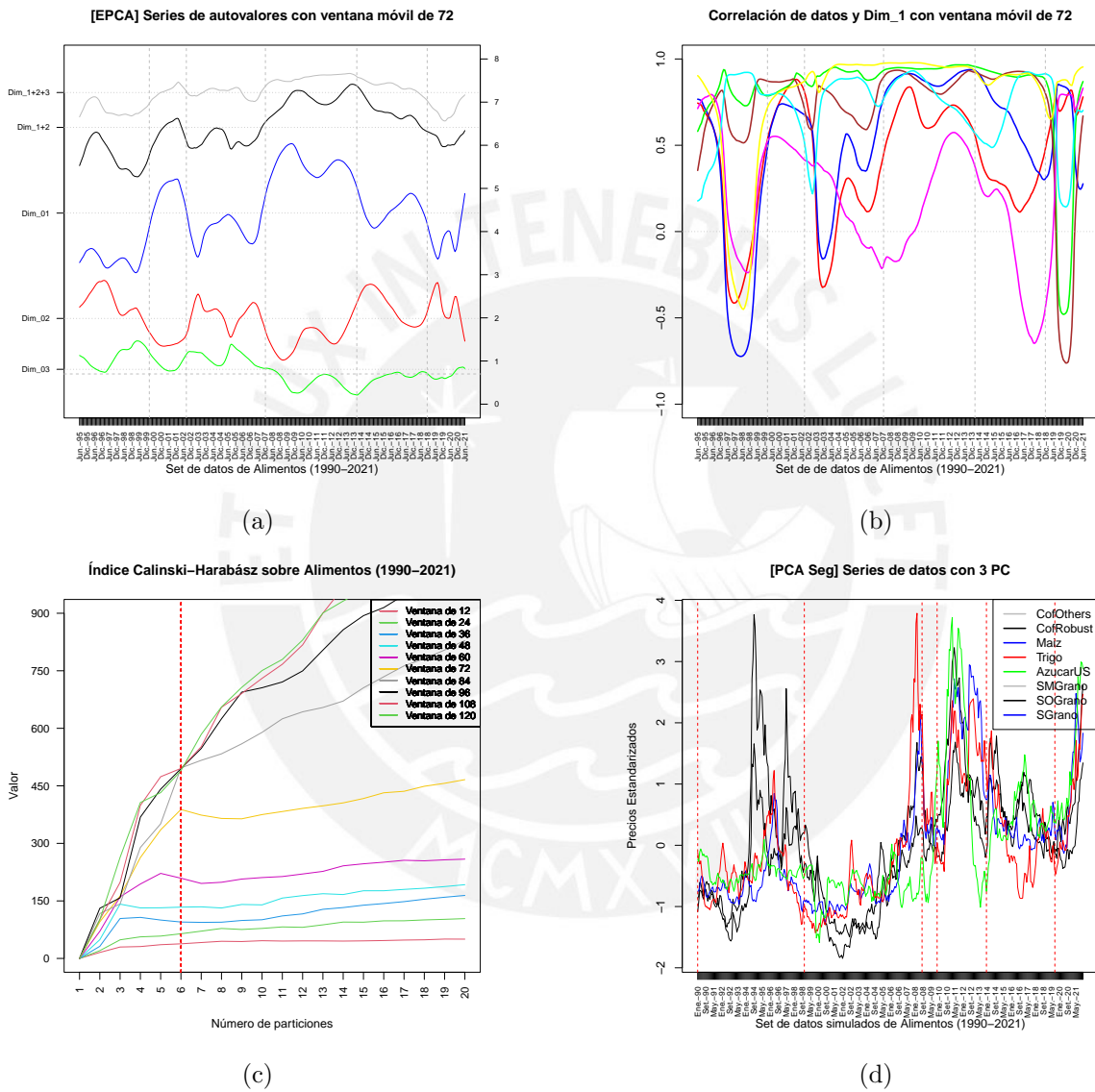


Figura A.34: Método PCA Evolutivo - Materias Primas de Alimentos

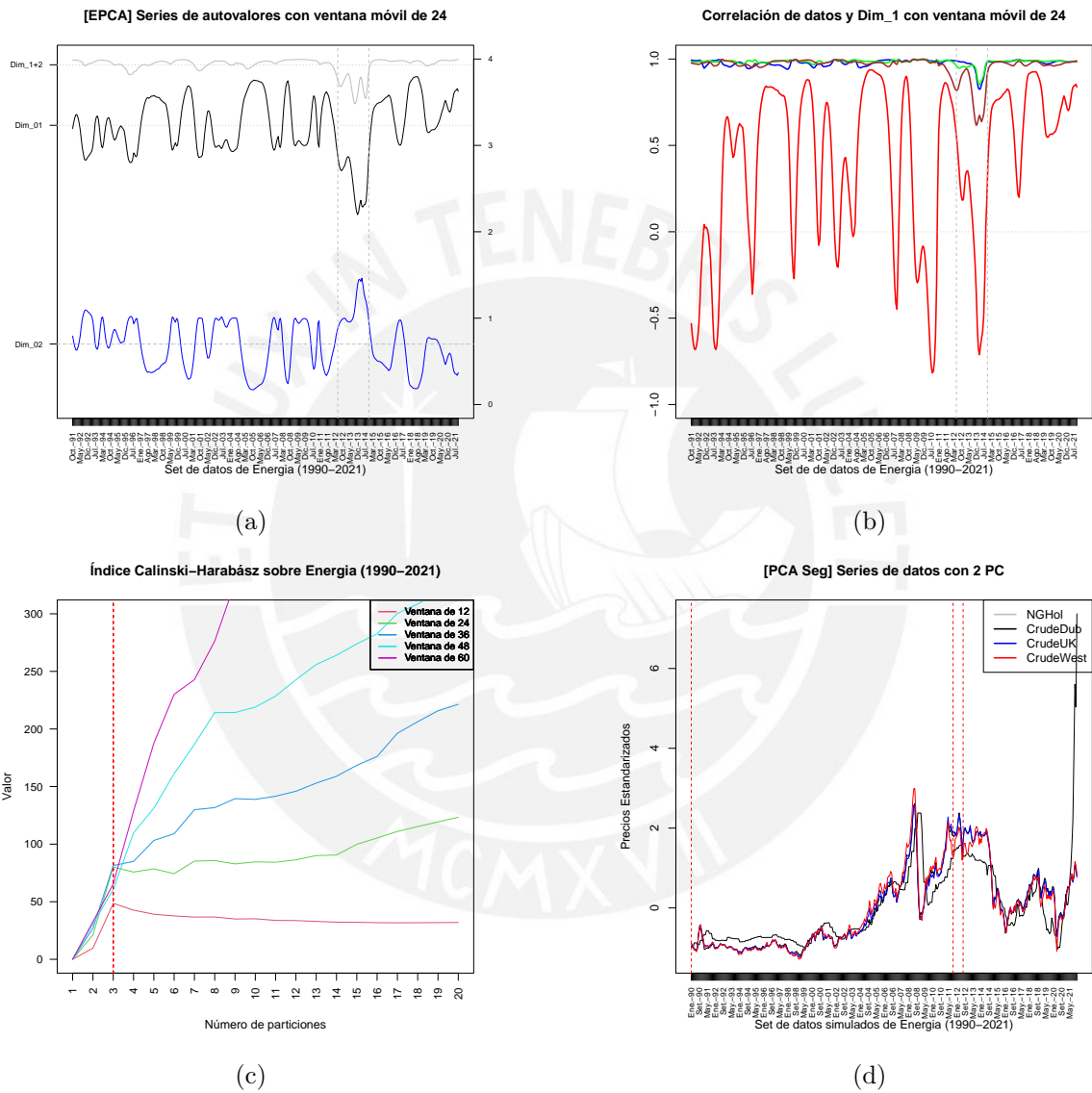


Figura A.35: Método PCA Evolutivo - Materias Primas de Energía

Bibliografía

- Atoche Murrieta, L. K. (2022). *Estimación del ciclo financiero utilizando métodos de análisis exploratorios: aplicación a Perú en el periodo 2000-2021*. tesis de maestría, Pontificia Universidad Católica del Perú, Lima, Perú.
- Bankó, Z. and Abonyi, J. (2012). Correlation based dynamic time warping of multivariate time series. *Expert Systems with Applications*, 39(17):12814–12823.
- Banko, Z., Dobos, L., and Abonyi, J. (2011). Dynamic principal component analysis in multivariate time-series segmentation. *Conservation, Information, Evolution-towards a sustainable engineering and economy*, 1(1):11–24.
- Camiz, S., Maulucci, R., and Roig, F. (2011). Exploratory analysis methods applied to dendrochronological series. In Papu, O., editor, *Actas E-ICES 5*, pages 61–67, Mendoza, Argentina. 5. E-ICES 5, Malargüe, 24-27 ooctubre 2009, Comisión Nacional de Energía Atómica.
- Camiz, S. and Pillar, V. D. (2018). Identifying the informational/signal dimension in principal component analysis. *Mathematics*, 6(11):269.
- Deaton, A. and Laroque, G. (1992). On the behaviour of commodity prices. *The review of economic studies*, 59(1):1–23.
- Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.
- Erten, B. and Ocampo, J. A. (2021). The future of commodity prices and the pandemic-driven global recession: evidence from 150 years of data. *World Development*, 137:105–164.
- Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American statistical Association*, 53(284):789–798.
- Ford, W. H. (2015). *Numerical Linear Algebra with applications: Using Matlab*. Academic Press, London, England.

- Gruss, B. and Kebhaj, S. (2019). *Commodity terms of trade: A new database*. International Monetary Fund.
- Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.
- Keogh, E., Chu, S., Hart, D., and Pazzani, M. (2001). An online algorithm for segmenting time series. In *Proceedings 2001 IEEE international conference on data mining*, pages 289–296. IEEE.
- Ku, W., Storer, R. H., and Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and intelligent laboratory systems*, 30(1):179–196.
- Liu, Y.-m., Liang, J., and Qian, J.-x. (2004). The application of dynamic principal component analysis to enhance chunk monitoring of an industrial fluidized-bed reactor. In *Fifth World Congress on Intelligent Control and Automation (IEEE Cat. No. 04EX788)*, volume 2, pages 1685–1688. IEEE.
- Lovrić, M., Milanović, M., and Stamenković, M. (2014). Algorithmic methods for segmentation of time series: An overview. *Journal of Contemporary Economic and Business Issues*, 1(1):31–53.
- Peña, D., Smucler, E., and Yohai, V. J. (2020). `gdpc`: an r package for generalized dynamic principal components. *Journal of Statistical Software*, 92:1–23.
- Peña, D. and Yohai, V. J. (2016). Generalized dynamic principal components. *Journal of the American Statistical Association*, 111(515):1121–1131.
- Vanhatalo, E. and KulaHCI, M. (2016). Impact of autocorrelation on principal components and their use in statistical process control. *Quality and Reliability Engineering International*, 32(4):1483–1500.
- Zamprogno, B., Reisen, V. A., Bondon, P., Aranda Cotta, H. H., and Reis Jr, N. C. (2020). Principal component analysis with autocorrelated data. *Journal of Statistical Computation and Simulation*, 90(12):2117–2135.