# TÉCNICA DE COMPONENTES INDEPENDIENTES: APLICACIÓN Y ANÁLISIS DE DATOS DE SERIES TEMPORALES SOBRE PUERTO RICO

Por

Ysela Ochoa Tapia

Tesis sometida en cumplimiento parcial de los requerimientos para el grado de

MAESTRÍA EN CIENCIAS

en

MATEMÁTICAS (ESTADÍSTICA)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ

2014

Aprobada por:

| | |
|---|---|
| Dámaris Santana Morant, Ph.D.<br>Miembro, Comité Graduado | Fecha |
| Olgamary Rivera Marrero, Ph.D.<br>Miembro, Comité Graduado | Fecha |
| Edgardo Lorenzo González, Ph.D.<br>Presidente, Comité Graduado | Fecha |
| Damaris Román González, M.B.A.<br>Representante de Estudios Graduados | Fecha |
| Omar Colón Reyes, Ph.D.<br>Director del Departamento | Fecha |

Abstract of Dissertation Presented to the Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master of Science

**INDEPENDENT COMPONENT ANALYSIS: APLICATION AND ANALYSIS OF TIME SERIES DATA ON PUERTO RICO**

By

Ysela Ochoa Tapia

2014

Chair: Ph.D Edgardo Lorenzo González
Major Department: Mathematical Sciences

The time series analysis is oriented to the problem of prediction. Perform a univariate analysis is simpler than a multivariate analysis because the univariate analysis is concerned only with the structure of internal dependency of a series, while the multivariate analysis also considers the dependency between series and its combinations. In this research we present a method for the prediction of multivariate time series, using its independent latent series, obtained through the independent component analysis as a blind source separation technique. The fact that the latent series are independent, allow us to reduce the multivariate analysis to a multiple univariate one.

Formally, the independent component analysis, is a mixture model of random variables $X = AS$, where the theory developed is focused on estimating the mixing matrix $A$ and the latent sources $S$, under assumptions that the matrix $A$ is full range, and the sources $S$ are independent non-Gaussian [12]. In this thesis, we have temporal data in form of time series, where the model of independent components

will be a combination of latent series as $X(t) = AS(t)$, where $t$ is time. The estimation is based on the only available information $X(t)$, that as a result of being time series data, there is no restriction on the data being Gaussian or not [11]. The model does not incorporates errors, because it assumes white noise. To perform the estimation of latent series various methods have been developed, from different points of view, based on the hypothesis that the latent series have a certain temporal structure associated with different autocorrelation functions [6]. These methods are called temporal-space decorrelation. The most commonly used algorithms are AMUSE based on whitening of data and the covariance matrix of a time delay [23] and SOBI based on second order blind identification, such that diagonalizes joint covariance matrices of a fixed number of time delays [1]. After the estimation of the latent series, the methodology proposed by Box and Jenkins is used, through SARIMA models [2], for forecast each latent series independently. Then under the model of independent components $X(t) = AS(t)$ with the predictions of the latent series $S(t + h)$, we predict the original time series $X(t + h)$. The methodology has been applied to multivariate time series of electricity consumption and the consumer index price of Puerto Rico, economic indicators that are key for decision making and economy of Puerto Rico [7]; for the development of methodology we use the AMUSE and SOBI algorithms. The results show the efficiency of the methodology and the reduction of the complexity of the prediction problem.

Resumen de Disertación Presentado a Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestría en Ciencias

# TÉCNICA DE COMPONENTES INDEPENDIENTES: APLICACIÓN Y ANÁLISIS DE DATOS DE SERIES TEMPORALES SOBRE PUERTO RICO

Por

Ysela Ochoa Tapia

2014

Consejero: Ph.D Edgardo Lorenzo González
Departamento: Ciencias Matématicas

El análisis de series de tiempo está orientado al problema de predicción. Realizar un análisis univariado es más sencillo que un análisis multivariado, porque el univariado se preocupa por la estructura de dependencia interna de una serie, en cambio el multivariado considera además la dependencia entre series y sus combinaciones. En esta investigación se presenta un método para la predicción de series de tiempo multivariadas, utilizando sus series latentes independientes, que se obtienen mediante el análisis de componentes independientes como una técnica de separación ciega de fuentes. El hecho de que las series latentes sean independientes nos permite reducir el análisis multivariado a uno univariado múltiple.

Formalmente el análisis de componentes independientes, se describe como, el modelo de mezcla de variables aleatorias $X = AS$, donde la teoría desarrollada se centra en estimar la matriz de mezcla $A$ y las fuentes latentes $S$, bajo las suposiciones de que la matriz $A$ es de rango completo, y las fuentes $S$ son independientes no gaussianas [12]. En este trabajo, tendremos datos temporales en forma de series de tiempo, donde el modelo de componentes independientes será una combinación de

series latentes de la forma $X(t) = AS(t)$, donde $t$ representa el tiempo. La estimación se hace a partir de la única información disponible $X(t)$, que por ser series de tiempo, no hay restricción de que los datos sean gaussianos o no [11]. Además el modelo no incorpora error, debido a que se asume ruido blanco. Para llevar acabo la estimación de las series latentes se han desarrollado diversos métodos, desde diferentes puntos de vista, basados en la hipótesis de que las series latentes tienen cierta estructura temporal asociada a diferentes funciones de autocorrelación [6]. Estos métodos se denominan decorrelación espacio temporal. Los algoritmos más utilizados son, AMUSE basado en el blanqueamiento de los datos y la matriz de covarianza de un tiempo de retraso [23] y SOBI basado en la identificación ciega de segundo orden, que diagonaliza de forma conjunta matrices de covarianza de un cierto número fijo de tiempos de retraso [1]. Una vez realizada la estimación de las series latentes, se utiliza la metodología propuesta por *Box and Jenkins* mediante los modelos SARIMA [2], para la predicción de cada serie latente de forma independiente. Luego bajo el modelo de componentes independientes $X(t) = AS(t)$ con la predicción de las series latentes $S(t + h)$, se predice las series de tiempo originales $X(t + h)$. La metodolgía es aplicada a series de tiempo multivariadas de consumo de energía eléctrica y el índice de precios al consumidor de Puerto Rico, indicadores económicos claves para la toma de decisiones y economía de Puerto Rico [7]; para el desarrollo de la metodología se hace uso de los algoritmos AMUSE y SOBI. Los resultados obtenidos muestran la eficiencia de la metodología y la reducción de la complejidad del problema de predicción.

A mi esposo, amigo y compañero Walter.

A mi papá Mario por sus sabios consejos.

A mi mamá Juana que Dios la tenga en su gloria.

A mis hermanos por enseñarme a vivir cada día.

## AGRADECIMIENTOS

A Dios, mi razón, mi paz y consuelo.

A mi asesor Dr. Edgardo Lorenzo por su apoyo, disposición y comprensión durante el desarrollo de la tesis.

A la Dr. Dámaris Santana por su disposición durante la tesis.

A la Dr. Olgamary Rivera por ser mi ejemplo y guía ahora y siempre.

A Madeline Ramos por su apoyo incondicional en los momentos más críticos durante la etapa de Bachillerato y Maestría.

Al Ing. Elmer Vélez incansable seguidor de Dios, sus palabras fueron de consuelo, fe y esperanza.

Al Departamento de Ciencias Matemáticas del Recinto Universitario de Mayagüez por todo el apoyo brindado.

A Velcy, Hector, John, Oscar, Frida, Moisés, Juan, Roxana, Widad, Greichaly y a todos mis amigos por su motivación, entusiasmo y confianza.

Índice general

página

Índice de cuadros

Cuadro página

# Índice de figuras

Figura página

## LISTA DE ABREVIATURAS

| | |
|---|---|
| ACF | Autocorrelation Function. |
| AIC | Akaike Information Criterion. |
| AMUSE | Algorithm for Multiple Unknown Signal Extraction. |
| BBS | Blind Source Separation |
| ICA | Independent Component Analysis. |
| ICs | Independent Components. |
| IPC | Índice de Precios al Consumidor. |
| JAD | Joint Approximate Diagonalization. |
| MAPE | Mean Absolute Percentage Error. |
| PACF | Partial Autocorrelation Function. |
| PCA | Principal component analysis. |
| RB | Ruido Blanco. |
| SOBI | Second Order Blind Identification. |
| SARIMA | Seasonal Autoregressive Integrated Moving Average. |
| VAR | Vector Autoregressive. |

LISTA DE SIMBOLOS

| | |
|---|---|
| $\rho_k$ | Función de Autocorrelación. |
| $\alpha_k$ | Función de Autocorrelación Parcial. |
| $\gamma_k$ | Función de Autocovarianza. |
| $t$ | Índice del tiempo. |
| $C_k^X$ | Matriz de Covarianza de $X$ con retraso $k$. |
| $\Delta_s^D$ | Operador de Diferencias Estacionales de orden $D$ y período $s$. |
| $\Delta^d$ | Operador de Diferencias Regulares de orden $d$. |
| $L$ | Operador de Retrasos. |
| $off\,(\cdot)$ | Operador de Suma de Cuadrados de los Elementos Fuera de la Diagonal. |
| $\mathrm{E}(\cdot)$ | Operador Valor Esperado. |
| $\Phi(L)$ | Polinomio Autorregresivo Estacional. |
| $\phi(L)$ | Polinomio Autorregresivo Regular. |
| $\Theta(L)$ | Polinomio de Media Móvil Estacional. |
| $\theta(L)$ | Polinomio de Media Móvil Regular. |
| $k$ | Tiempo de retraso. |

# Capítulo 1
# INTRODUCCIÓN

## 1.1. Justificación

Una serie de tiempo son datos ordenados cronológicamente a intervalos de tiempo constante, estos pueden ser recopilados en días, meses, años, etc. Se puede trabajar de forma univariada representada por un vector o de forma multivariada representada por una matriz. A partir de este conjunto de datos se trata de describir su comportamiento, es decir, la tendencia que estas presentan, sus variaciones estacionales y componentes aleatorios; para luego encontrar un modelo que permita hacer predicciones a corto o a largo plazo.

Los primeros en proponer una metodología de estudio para modelos lineales multivariados fueron *Box and Jenkins* (1976) [2], en el cual desarrollaron los modelos de vectores autoregresivos o VAR por sus siglas en inglés, para estudiar los componentes de las series y conseguir un modelo predictivo, pero estudios posteriores comprobaron que los modelos VAR, a pesar de trabajar con series de tiempo multivariadas, presentaban problemas de sesgo al momento de trabajar con un grupo grande de datos, por esta razón se le llamó, modelo de escala pequeña. Por el contrario un modelo que incorpora una gran número de variables es llamado modelo de escala mayor. *Watson* (2000) muestra evidencia empírica en apoyar los modelos de gran escala, comparado con los de pequeña escala [24]. Una clase particular de lo modelos de escala mayor es conocido como el modelo factorial dinámico iniciados por *Sargent and Sims* (1977). El modelo factorial dinámico plantea que las series de tiempo surgen de algunos factores ocultos o latentes, a raíz de esto se presentan métodos que

separan estos factores ocultos, uno de ellos es el análisis de componentes principales o PCA por sus siglas en inglés, (*Stock and Watson* 2002). Este método de fácil implementación obtiene los factores ocultos no correlacionados. Resultados posteriores demuestran que este procedimiento de selección del modelo factorial dinámico puede ser mejorado, por que no obtienen factores independientes [24]. Como un método alternativo al PCA, se propone el análisis de componentes independientes o ICA por sus siglas en inglés. La fortaleza de este método, es que obtiene factores ocultos que son independientes. Para la separación de factores ocultos, se han desarrollado diversos algoritmos basados en el ICA, para variables aleatorias medidas en un instante en el tiempo, es decir, datos que no dependen del tiempo; siendo el algoritmo más usado el FastICA [13]. Por otro lado, adaptaciones de algoritmos ICA para series de tiempo tales como, el AMUSE por sus siglas en inglés *Algorithm for Multiple Unknown Signal Extraction*, y el SOBI por sus siglas en inglés *Second Order Blind Identification*. La ventaja de obtener series latentes que sean independientes, es que se puede realizar un análisis a cada serie por separado. Luego utilizando la metodología de *Box and Jenkins*, para identificar un modelo que caracteriza a cada serie latente y realizar predicciones $h$ pasos adelante. Finalmente podemos reconstruir las predicciones de las series de tiempo originales bajo el modelo ICA planteado.

## 1.2. Antecedentes

El análisis de componentes independientes, es una técnica usada para revelar las fuentes ocultas de un conjunto de variables aleatorias. Dicha técnica se ha implementado como modelos ICA de mezcla instantánea, con o sin ruido, donde las variables aleatorias son medidas en un instante en el tiempo; los modelos ICA para series de tiempo donde las variables aleatorias dependen del tiempo, modelos ICA de mezclas convolutivas, etc [6].

La técnica ICA inicialmente fue propuesta para la solución del problema de separación ciega de fuentes o BSS por sus siglas en inglés *Blind Source Separation*, cuyo

planteamiento es recuperar las fuentes originales de una mezcla, sin tener conocimiento previo de como fueron mezcladas, de ahí el término ciego. *Herault and Jutten* (1986) aplicaron dicha técnica en el campo de redes neuronales. *Bell and Sejnoski* (1995) mostrarían el potencial del ICA a través de su publicación de una aproximación al problema llamado InfoMax. Actualmente existen conferencias internacionales de ICA y BSS, donde se presentan los avances teóricos y aplicativos en diferentes areas, tales como: el análisis de fuentes médicas, la separación de sonidos, el procesamiento de imágenes, la reducción de dimensión, el análisis de texto y código, por citar algunos [15]. La idea general del ICA es separar los factores ocultos de las mezclas bajo ciertas suposiciones; en el modelo ICA de mezcla instantánea para variables aleatorias, se asume que el ruido es blanco, el número de mezclas es igual al número de fuentes ocultas no gaussianas y los factores ocultos son mutuamente independientes; por otro lado en el modelo ICA de series de tiempo, se asume lo anterior sobre los factores ocultos algunas veces llamadas "series latentes", excepto que la suposición de no gaussianidad es reemplazada por otras suposiciones alternativas, por ejemplo, que las series latentes tengan diferentes autocovarianzas [11]. A consecuencia de este hecho, las investigaciones se han dividido en dos grupos, el primero trabaja con series de tiempo, basado en las estadísticas de segundo orden, que en lugar de asumir no gaussianidad, asumen que las series latentes tienen distintas varianzas; algoritmos que se basan en esta suposición son PCA, AMUSE, SOBI, TSICA. El segundo grupo trabajan con variables aleatorias, basado en las estadísticas de orden mayor, que asumen no gaussianidad y tratan de cuantificar esta medida a través de la kurtosis, negentropia, etc. Algoritmos que se basan en esta suposición son InfoMax, FastICA, JADE.

En esta investigación trabajaremos los algoritmos AMUSE y SOBI. A pesar de que la técnica ICA al igual que el BSS presentan ciertas indeterminaciones y ambiguedades, como magnitud, escalamiento y permutación, esto es, la varianza y el orden de

los componenten independientes no pueden ser determinada, ICA a reportado excelentes resultados en finanzas y economía, que fueron mostrados en [8], donde *Peña*, resalta la eficiencia del algoritmo SOBI comparado con otros, en series de tiempo simuladas, series de retornos de las acciones e índices de producción industrial. Por otro lado *Yau* en [24], resalta el uso del algoritmo TSICA para series de tiempo de diferentes indicadores macroeconómicos. Finalmente *Popescu* en [18], muestra la eficiencia del ICA en series de tiempo simuladas, donde emplea el algoritmo JADE. Estas referencias presentan una perspectiva diferente de realizar predicciones de series de tiempo multivariadas, mediante el uso de la técnica ICA.

## 1.3. Objetivos

### 1.3.1. Objetivo General

El objetivo general de esta investigación es aplicar la metodología del análisis de componentes independientes adaptada a series de tiempo para analizar la capacidad predictiva en series de tiempo multivariadas en el área de economía y finanzas de Puerto Rico.

### 1.3.2. Objetivos Específicos

- Aplicar los algoritmos *Second Order Blind Identification* (SOBI) y *Algorithm for Multiple Unknown Signal Extraction* (AMUSE), a las series de tiempo multivariadas, obtenidas en el área de economía y finanzas de Puerto Rico, para la estimación de series latentes.
- Identificar modelos adecuados de series de tiempo para cada serie latente, tanto para describir las series como para realizar predicciones de valores futuros y validar los modelos propuestos, utilizando la metodología de *Box and Jenkins*.
- Estimar las predicciones para las series de tiempo multivariadas originales a través del modelo de componentes independientes y las predicciones de cada serie latente.
- Analizar la capacidad predictiva del modelo obtenido con la metodología del análisis de Componentes Independientes.

# Capítulo 2
# REVISIÓN DE LITERATURA

## 2.1. Introducción

El análisis estadístico multivariado es un conjunto de métodos estadísticos, que analiza simultáneamente medidas múltiples de cada individuo u objeto sometido a investigación. Se puede analizar el comportamiento de una variable en función de otras, o incluso, predecir su comportamiento dadas otras variables conocidas. El análisis multivariado incluye métodos de regresión, clasificación, así como métodos de proyección sobre variables latentes o variables no observadas [17]. Las técnicas de proyección sobre variables latentes tales como el PCA o ICA, resultan interesantes, no solo por el hecho de que revelan las variables no observadas del sistema, transformando el espacio de variables originales en un nuevo espacio más representativo de la población, sino porque dicha transformación podría conllevar una reducción de la dimensionalidad del problema [11].

## 2.2. El Análisis de Componentes Principales

El PCA es una técnica del análisis multivariado que se aplica sobre variables cuantitativas *Hotelling* (1933). Dado un conjunto de $p$ variables aleatorias, el PCA se encargará de generar un nuevo conjunto de $m$ variables donde $m \leq p$, que serán representadas como una combinación lineal de las variables originales, de modo que, estas nuevas variables expliquen la mayor parte de la variabilidad que se tenía en los datos originales y además estarán no correlacionadas. Si $m < p$, hablamos de una reducción de la dimensionalidad del problema analizado. Cuando un conjunto

de variables recopiladas presentan correlación entre sí, se podría haber obtenido información que en cierto sentido fuese redundante, el PCA permite obtener un nuevo conjunto de componentes principales que están no correlacionados y libres de efectos de escala, facilitando así la interpretación de los datos. Debido a que los componentes principales son una combinación lineal de las variables observadas originalmente, se pueden obtener tantos componentes principales como variables originales, pero no todos los componentes proporcionan la misma cantidad de información o explican el mismo porcentaje de variabilidad presente en los datos, así que debemos seleccionar aquellas que aporten mayor información, desechando las restantes. Esto permite que en el PCA se utilice la gráfica de sedimentación como una herramienta para una reducción efectiva de la dimensionalidad [17].

### 2.2.1. Planteamiento Matemático del PCA

Consideremos $X_{n\times p}$ la matriz de observaciones, donde $n$ es el número de observaciones y $p$ el número de variables aleatorias, representadas por el vector aleatorio $X = (X_1, X_2, \ldots, X_p)$, inicialmente correlacionadas, los vectores aleatorios se entiende como vectores columna; se desea obtener un número $m < p$ de variables aleatorias no correlacionadas representadas por el vector aleatorio $Z = (Z_1, Z_2, \ldots, Z_m)$, que sean combinación lineal de las variables aleatorias originales y que expliquen la mayor parte de la variabilidad del problema. Desde el punto de vista geométrico el PCA se describe como una rotación ortogonal del sistema de coordenadas, tal que remueve las redundancias a causa de las correlaciones entre las $p$ variables, por lo que los $m$ componentes del nuevo vector aleatorio $Z$, son las nuevas coordenadas y están no correlacionadas.

Para el cálculo de los componentes principales, consideramos que el vector aleatorio $Z$, sea $p$-dimensional, esto es: $Z = (Z_1, Z_2, \ldots, Z_p)$. De este modo, cada variable aleatoria $Z_k$ con $k = 1, 2, \ldots, p$, puede expresarse como una combinación lineal de

las variables aleatorias originales $X_j$ con $j = 1, 2, \ldots, p$, esto es.

$$Z_k = w_{1k}X_1 + w_{2k}X_2 + \cdots + w_{pk}X_p. \tag{2.1}$$

Donde $w_{1k}, w_{2k}, \ldots, w_{pk}$ son escalares, elementos del vector columna $p$-dimensional $\mathbf{w}_k$, de la matriz de transformación lineal ortogonal $W_{p\times p}$. La forma vectorial de la ecuación (2.1) es:

$$Z_k = \mathbf{w}_k^T X \quad k = 1, 2, \ldots p \tag{2.2}$$

Donde $T$ representa la transpuesta de una matriz.

### 2.2.2. Método PCA por Maximización de la Varianza

Para encontrar las componentes principales maximizaremos las varianzas de los componentes $Z_k$, bajo la restricción de $||\mathbf{w}_k|| = 1$, de modo que el primer componente será el que explique la mayor parte de la variabilidad de los datos, el segundo componente corresponderá a la máxima variabilidad en dirección del primer componente, etc.

Consideremos la combinación lineal $Z_1 = \sum_{j=1}^{p} w_{j1}X_j = \mathbf{w}_1^T X$. El criterio de maximizar la varianza, se plantea como la maximización de:

$$J_1^{PCA}(\mathbf{w}_1) := \mathrm{E}(Z_1^2) = \mathrm{E}[(\mathbf{w}_1^T X)^2] = \mathbf{w}_1^T \mathrm{E}(XX^T)\mathbf{w}_1 = \mathbf{w}_1^T C_X \mathbf{w} \tag{2.3}$$

Bajo la restricción $||\mathbf{w}_1|| = 1$, donde $C_X$ es la matriz de covarianza con dimensión $p \times p$ del vector aleatorio $X$ de media cero; $\mathrm{E}(\cdot)$ es el operador valor esperado. Considerando la descomposición de autovalores y autovectores de $C_X$

$$C_X = UDU^T \tag{2.4}$$

Donde $U$ es la matriz de autovectores $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$ y $D$ es la matriz diagonal de autovalores $d_1, d_2, \ldots, d_p$ de $C_X$ con $d_1 \geq d_2 \geq \ldots \geq d_p$. La solución de (2.3), estará dada por $\mathbf{w}_1 = \mathbf{u}_1$, así el primer componente principal de $X$ es $Z_1 = \mathbf{u}_1^T X$. De forma general el criterio de maximizar la varianza está dado por $Z_k = \mathbf{u}_k^T X$.

Para la selección del número de componentes principales se usa el gráfico de sedimentación [17].

## 2.3. Blanqueamiento

El blanqueamiento de datos o *whitening*, se realiza a través de una matriz de transformación lineal ortogonal, multiplicada al vector de observaciones, cuyo resultado es un vector de componentes no correlacionados de varianza unidad. Esta matriz de transformación lineal ortogonal $V$ la encontramos a través de la descomposición de valores propios de la matriz de covarianza de los datos observados.

$$C_X = \mathrm{E}(XX^T) = UDU^T \tag{2.5}$$

Finalmente $V = D^{-1/2}U^T$. PCA puede hacer uso del blanqueamiento de datos para conseguir los componentes no correlacionados, pero el objetivo principal de discutir este tema es que blanqueamiento será empleado por el ICA como un paso de pre procesamiento, que reducirá la búsqueda de la matriz de mezcla al espacio de matrices ortogonales [12].

## 2.4. El Análisis de Componentes Independentes para Variables Aleatorias

El ICA es una técnica originalmente utilizada para la solución de un problema particular de la separación ciega de fuentes, conocida como el *Cocktail Party Problem*, este problema ilustra la idea del ICA y plantea lo siguiente: "Imaginemos a dos personas hablando simultaneamente en una sala de espera, si dos micrófonos son colocados en diferentes lugares de la sala, en donde cada uno graba la combinación de voces. Usando solamente las grabaciones, podía ser posible recuperar las voces originales". [5].

### 2.4.1. Modelo de Variables Latentes

El ICA básico es un modelo que ignora el ruido y la dependencia en el tiempo. Consideremos el vector aleatorio $X = (X_1, X_2, \ldots, X_p)$, donde cada variable aleatoria $X_i$ es una combinación lineal de $p$ variables aleatorias $S_i$ miembros del vector aleatorio $S = (S_1, S_2, \ldots, S_p)$, el modelo básico de ICA es.

$$X = AS \tag{2.6}$$

Donde $A_{p\times p}$ representa la matriz de mezcla. El objetivo de ICA es recuperar los factores latentes $S_i$, a partir de las variables observadas $X_i$. La Figura 2–1 muestra este proceso.

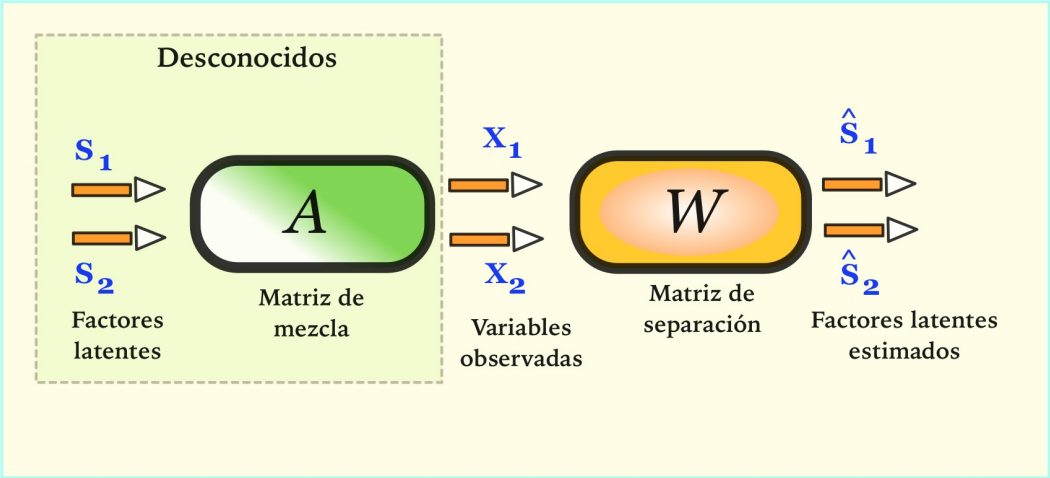


Figura 2–1: Modelo Conceptual de ICA

Para obtener los factores latentes estimados, primero se debe obtener la matriz de separación $W$ donde $W = A^{-1}$, así obtenemos $Y$ que será la mejor aproximación de $S$.

$$Y = WX \approx S \quad \text{ó} \quad \hat{S} = WX \tag{2.7}$$

Algunas veces necesitaremos la notación de (2.6) usando las columnas de $A$, esto es.

$$X = \sum_{j=1}^{p} \mathbf{a}_j S_j \tag{2.8}$$

Donde $\mathbf{a}_j$ es la $j$-ésima columna de $A$.

**Suposiciones de ICA** Se deben tener en cuenta ciertas consideraciones para la estimación del modelo [21].

1. Las $S_j$ son estadísticamente independientes entre ellas (el concepto de independencia es el principio fundamental del ICA).
2. La matriz $A$ se asume cuadrada de rango completo (el número de variables aleatorias $X_j$ debe ser igual al número de componentes independientes $S_j$).
3. No hay ruido externo en el modelo (todos los factores aleatorios considerados ruido están en las componentes independientes).
4. Las $S_j$ no deben tener distribución gaussiana, excepto a lo más una.

**Ambigüedades del ICA** Existen dos ambigüedades en ICA [15].

1. Ambigüedad de magnitud y escalamiento; no es posible determinar las varianzas de las $S_j$. Veamos, considere la ecuación (2.8), para cualquier escalar $\alpha_j \neq 0$, la nueva mezcla es, $X = \sum_{j=1}^{p} (\frac{1}{\alpha_j}\mathbf{a}_j)(S_j\alpha_j)$, entonces $S_j$ y $S_j\alpha_j$ podrían ser las componentes independientes de $X$, pero sus varianzas toman diferentes valores. Para evitar tal indeterminación se fija $\mathrm{E}(S_j^2) = 1$. Sin embargo, los componentes independientes aún quedan con la indeterminación del signo, afortunadamente en la mayoría de aplicaciones esta ambigüedad es insignificante.
2. No es posible determinar el orden las componentes independientes. La explicación formal a esto es que, si insertamos una matriz de permutación $P$ y su respectiva inversa al modelo básico ICA (2.6), tendríamos $X = AP^{-1}PS$, donde los elementos de $PS$ serían los componentes independientes $S_j$ con el orden cambiado, $AP^{-1}$ es simplemente una nueva matriz de mezcla desconocida.

### 2.4.2. Estimación del Modelo Básico ICA

La mayoría de algoritmos propuestos para la estimación del modelo ICA, usualmente incorporan suposiciones adicionales a las antes mencionadas. La más popular es asumir que la matriz de mezcla $A$ es ortogonal. Esta suposición es naturalmente incluida en el modelo, a través del pre procesamiento (*whitening*) del vector aleatorio $X$, tal que $Z = VX$, cuyo modelo ICA es, $Z = AS$. Esta suposición trae como ventaja reducir el número de parámetros a ser estimados, de $p^2$ a $\frac{p(p+1)}{2}$ en la nueva matriz de mezcla ortogonal [8]. Dado el modelo ICA para los datos blanqueados $Z$.

$$Z = AS \tag{2.9}$$

Se desea encontrar, componentes independientes dados por la combinación lineal de datos blanqueados.

$$Y = WZ \tag{2.10}$$

Donde $W_{p\times p}$ es la matriz ortogonal de separación ($W = A^{-1}$), $Y = (Y_1, Y_2, \ldots, Y_p)$, tal que los $Y_i$ sean lo más independientes posible. Existen muchas formas de medir la independencia, la principal se basa en la condición de no gaussianidad de los componentes independientes [13].

**Cálculo de los Componentes Independientes:** Considerando una combinación lineal del vector $X$, denotada por.

$$Y_i = \mathbf{w}_i^T X \tag{2.11}$$

Si $\mathbf{w}_i$ es una de las filas de $A^{-1}$, entonces la combinación lineal $Y_i$ corresponde a un componente independiente. Para que la condición se cumpla se hace uso del teorema del límite central que implica: "Bajo ciertas condiciones, la distribución de la suma de dos variables, tiende hacia una distribución que será más gaussiana que las variables originales" [12].

Dicho esto; como $Y$ es tambien combinación lineal de $S$, esto es:

$$Y = \mathbf{w}^T X = \mathbf{w}^T AS = \mathbf{q}^T S,$$

entonces $Y$ será menos gaussiana cuanto más se aproxime a $S$. La independencia de los componentes $Y$ es a través de la maximización de la no gaussianidad, las medidas usadas son la kurtosis y la entropía.

- La kurtosis se usa como medida de no gaussianidad; para datos blanqueados la kurtosis es igual al cuarto momento, dado que los datos tienen varianza unidad, entonces la kurtosis del $i$-ésimo componente independiente $Y_i = \mathbf{w}_i^T Z$, está dado por:

  $$\text{Kurt}(Y_i) = \text{E}(Y_i^4) - 3$$

  La kurtosis puede ser positiva o negativa y es igual a cero cuando $Y_i$ es gaussiano. Aunque sus cálculos son sencillos, no es considerado una medida robusta para ICA puesto que es sensible a los *outliers* [20].

- La Entropía: Dada una matriz de covarianza de datos blanqueados, la distribución que tiene más entropía es la distribución gaussiana [20] , luego el principio de maximizar la no gaussianidad equivale a minimizar la entropía, sin embargo la entropía no es invariante a las transformaciones lineales, como una alternativa a esto, ICA usa la negentropía como una medida de no gaussianidad. La negentropía para los componentes independientes es:

  $$\text{N}(Y_i) = \text{H}(Y_{gauss}) - \text{H}(Y_i)$$

  La negentropía es siempre positiva y será cero cuando el componente tiene distribución gaussiana. Existen algoritmos que implementan dicha medida, entre los cuales estan el FastICA, JADE e InfoMax [11].

## 2.5. Series de Tiempo

Una serie de tiempo son datos ordenados cronológicamente a intervalos de tiempo constante. Una característica de una serie de tiempo es que sus observaciones están correlacionadas, por tanto, el orden de las observaciones es importante. Una series de tiempo, se denota como un modelo aditivo:

$$X_t = T_t + E_t + I_t$$

Donde, $T_t$ es el componente de tendencia y se define como el cambio a largo plazo que representa el crecimiento o disminución de la serie, $E_t$ es el componente estacional y se define como la variación periódica, inferior o igual a un año e $I_t$ es el componente aleatorio que se define como la variabilidad ocasionada por factores imprevistos. La Figura 2–2, muestra estos componentes de una serie de tiempo.

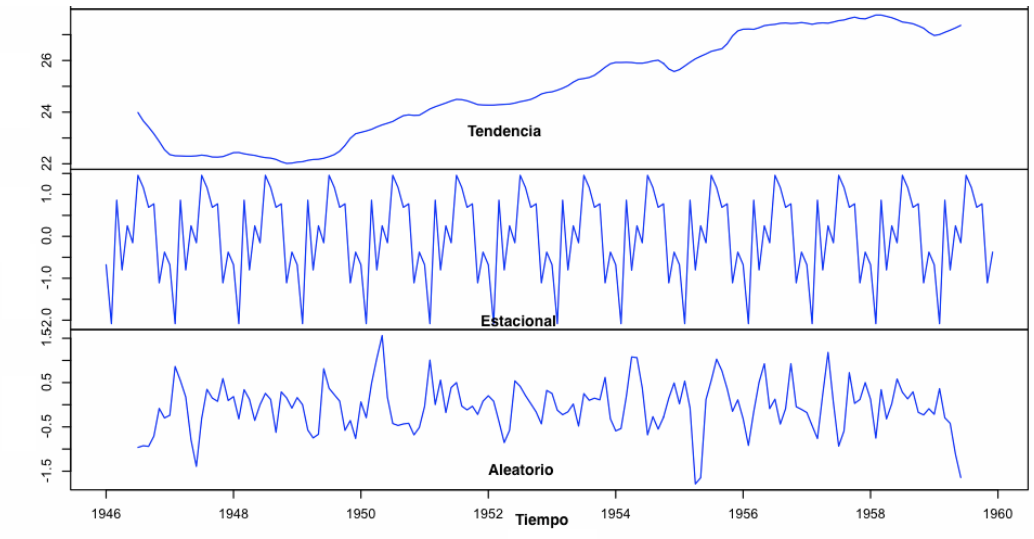


Figura 2–2: Componentes de una Serie de Tiempo

El enfoque para analizar una serie de tiempo es el propuesto por *Box and Jenkins* [2], el cual, ajusta un modelo a los datos, seleccionado de la familia SARIMA por sus siglas en inglés *Seasonal Autoregressive Integrated Moving Average*, cuyos objetivos son.

1. Identificar un modelo que describa el comportamiento de la serie de datos.

2. Predecir valores futuros de la serie a partir de la información disponible.

La predicción se realiza suponiendo que el modelo es adecuado a lo largo del tiempo [3]. La metodología propuesta por *Box and Jenkins* está basada en modelos lineales de procesos estocásticos estacionarios, ARMA por sus siglas en inglés *Autoregressive Moving Average* y de procesos que se puedan transformar en estacionarios, SARIMA.

### 2.5.1. Procesos Estocásticos

Un proceso estocástico es una familia de variables aleatorias $\{X_t, t \in T\}$ donde $T$ es el conjunto de índices. Una serie de tiempo es una realización de un proceso estocástico, denotado por $\{x_t, t \in T_0\}$ , donde $T_0$ representa el tiempo. Un proceso estocástico es estacionario o estacionario de segundo orden si su media y varianza son constantes en el tiempo y las autocovarianzas sólo dependen del número de periodos de separación entre las variables y no del tiempo [10]. El ruido blanco es el caso más simple de los procesos estocásticos estacionarios, donde las variables son independientes e idénticamente distribuidas a lo largo del tiempo, con media cero y varianza constante. Un proceso estocástico estacionario está caracterizado por sus funciones de autocovarianza, autocorrelación y autocorrelación parcial.

- La función de autocovarianza es denotada por $\gamma_k$, dónde:

$$\gamma_k = \text{cov}(X_{t-k}, X_t) = \text{E}\{(X_{t-k} - \text{E}(X_{t-k}))(X_t - \text{E}(X_t))\} \quad \forall k = 1, 2, \ldots \quad (2.12)$$

  Ésta recoge toda la información sobre la estructura de variabilidad del proceso, pero depende de las unidades de medida de la variable.

- La función de autocorrelación o ACF por por sus siglas en inglés *Autocorrelation Function*, es denotada por $\rho_k$, dónde:

$$\rho_k = \text{corr}(X_{t-k}, X_t) = \frac{\text{cov}(X_{t-k}, X_t)}{\sqrt{\text{var}(X_{t-k})\text{var}(X_t)}} = \frac{\gamma_k}{\gamma_0} \quad \forall k = 1, 2, \ldots \quad (2.13)$$

Ésta mide la correlación lineal entre dos variables separadas por $k$ periodos. La ACF se representa gráficamente mediante el correlograma. Las características de la ACF de un proceso estocástico estacionario son:

- $\rho_0 = 1$.
- $-1 \leq \rho_k \leq 1$.
- $\rho_k = \rho_{-k}$.
- $\rho_k \to 0$ cuando $k \to \infty$.

- La función de autocorrelación parcial o PACF por por sus siglas en inglés *Parcial Autocorrelation Function*, denotada por $\alpha_k$, se define por:

$$\alpha_k = \text{corr}(X_{t-k}, X_t | X_{t-k+1}, \ldots, X_{t-1}) \quad \forall k = 1, 2, \ldots \tag{2.14}$$

Ésta mide la correlación lineal entre las dos variables después de que las dependencias lineales sobre las variables que intervienen $X_{t-k+1}, \ldots, X_{t-1}$ han sido removidas. Las propiedades de la PACF, son similares a la ACF [14].

### 2.5.2. Modelos Lineales Estacionarios

Los modelos lineales estacionarios paramétricos, en el estudio de las series de tiempo, son los modelos autorregresivos (AR) por sus siglas en inglés, modelos de medias móviles (MA) por sus siglas en inglés, y la combinación de estos, ARMA. Estos modelos se deben representar de forma finita.

**El Modelo Autorregresivo: AR($p$)** es de la forma:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \tag{2.15}$$

Donde $\phi_1, \phi_2, \ldots, \phi_p$ son los parámetros del modelo, $\varepsilon_t$ es un ruido blanco con media cero y varianza $\sigma^2$, denotado por $\varepsilon_t \sim RB(0, \sigma^2)$. En el modelo AR($p$), la información del pasado se transmite mediante las observaciones anteriores hasta un retraso $p$. El modelo AR($p$), se puede escribir de forma:

$$\phi(L) X_t = \varepsilon_t \tag{2.16}$$

Donde $\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$, representa el polinomio autorregresivo de orden $p$ y $L$ es el operador de retrasos definido como: $L_t^X = X_{t-k} \quad \forall k = 1, 2, \ldots$. El siguiente teorema proporciona condiciones necesarias y suficientes para que el modelo AR($p$) sea estacionario [3].

**Teorema:** Un proceso autorregresivo finito AR($p$) es estacionario si y sólo si el módulo de las raíces del polinomio autorregresivo $\phi(L)$ están fuera del círculo unitario.

En un proceso AR($p$) la función de autocorrelación $\rho_k$, decrece lentamente hacia cero según una exponencial o sinusiode. La función de autocorrelación parcial $\alpha_k$, es cero para $k > p$

**El Modelo de Medias Móviles: MA($q$)** es de la forma:

$$X_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{2.17}$$

Donde $\theta_1, \theta_2, \ldots, \theta_q$ son los parámetros del modelo y $\varepsilon_t \sim RB(0, \sigma^2)$ En el modelo MA($q$), la información del pasado se transmite mediante los errores o innovaciones anteriores hasta un retraso $q$. El modelo MA($q$), se puede escribir de forma:

$$X_t = \theta(L)\varepsilon_t \tag{2.18}$$

Donde $\theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q$, representa el polinomio de medias móvilles de orden $q$. El modelo MA($q$) debe ser invertible, esto es, cuanto más nos alejamos en el pasado, la influencia de éste disminuye en el presente. El siguiente teorema proporciona condiciones necesarias y suficientes para que el modelo MA($q$) sea invertible [3].

**Teorema:** Un proceso de medias móviles finito MA($q$) es invertible si y sólo si el módulo de las raíces del polinomio de medias móviles $\theta(L)$ están fuera del circulo unitario.

En un proceso MA($q$) la función de autocorrelación $\rho_k$, es cero para $k > q$. La función de autocorrelación parcial $\alpha_k$, decrece lentamente hacia cero según una exponencial o sinusiode.

Los procesos de medias móviles son procesos de memoria corta, esto debido a que la influencia del pasado en el presente, termina en pocos retrasos; mientras que los autorregresivos son procesos de memoria larga, ya que la influencia del pasado en el presente, se prolonga en muchos retrasos. La Figura 2–3 muestra la ACF de una MA(1) y un AR(1), donde el ACF del MA(1), termina en el retraso 1, mientras que el ACF del AR(1), se prolonga en muchos retrasos.

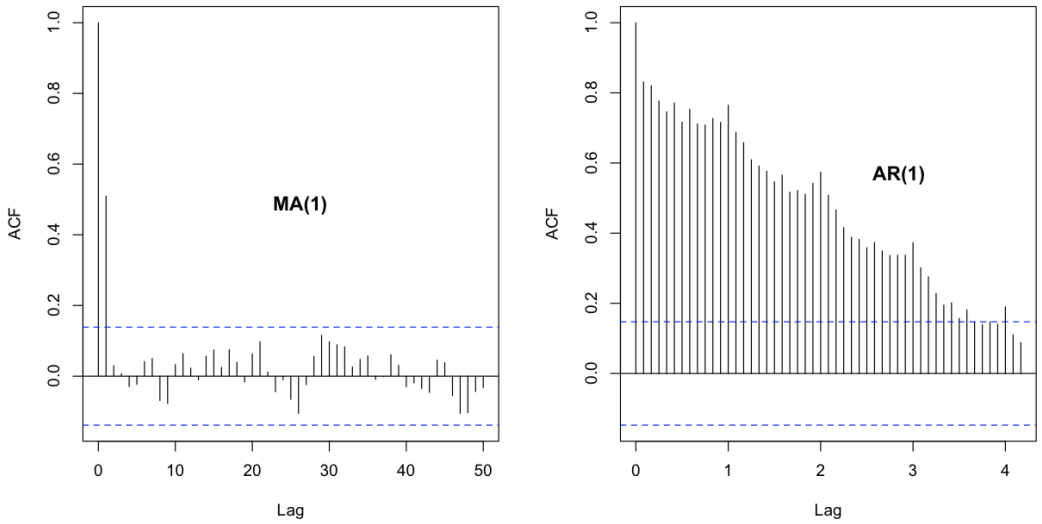


Figura 2–3: Memorias de los procesos MA(1) y AR(1)

**El Modelo Autorregresivo de Medias Móviles: ARMA$(p, q)$** es de la forma:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (2.19)$$

Donde $\varepsilon_t \sim RB(0, \sigma^2)$; $\phi_1, \phi_2, \ldots, \phi_p$; $\theta_1, \theta_2, \ldots, \theta_q$ son los parámetros del modelo. En el modelo la información se transmite mediente la combinación del AR($p$) y

MA($q$); se puede escribir de forma:

$$\phi(L)X_t = \theta(L)\varepsilon_t \tag{2.20}$$

Las condiciones de estacionariedad e invertibilidad del modelo ARMA$(p, q)$ vienen impuestas por la parte autorregresiva y de medias móviles respectivamente [3], las características también son combinaciones de ambas partes.

### 2.5.3. Modelos Lineales no Estacionarios

En esta sección se describen los modelos lineales de procesos estocásticos no estacionarios que se pueden transformar en estacionarios, mediante diferencias regulares o estacionales. Las series de tiempo no estacionarias en media, es decir, tienen tendencia en su estructura; son diferenciadas de forma regular, para transformarlas en estacionarias. Si una serie tiene el componente estacional, se diferencia de forma estacional, para transformarlas en estacionarias. De esta manera se generan los procesos ARIMA y SARIMA [3].

**Modelo Autorregresivo Integrado de Media Móvil: ARIMA$(p, d, q)$** Si $X_t$ es no estacionario en media, usamos el operador de diferencia regular, definido por $\Delta^d X_t = (1 - L)^d X_t$, luego el proceso transformado es estacionario y puede ser modelado por un modelo ARMA$(p, q)$. En consecuencia, el modelo autorregresivo integrado de medias móviles de orden $(p, d, q)$, denotado por ARIMA$(p, d, q)$ es:

$$\phi(L)\Delta^d X_t = \theta(L)\varepsilon_t \tag{2.21}$$

donde $\phi(L)$ es el polinomio autorregresivo y $\theta(L)$ es el polinomio de media móvil, definidos en (2.16) y (2.18). Notar que, si el proceso $\Delta^d X_t$ es estacionario de orden $d$, entonces $X_t$ es integrado de orden $d$.

**Modelo Estacional Autorregresivo Integrado de Media Móvil:** Si $X_t$ presenta el componente estacional, usamos el operador de diferencia estacional, definodo por $\Delta_s^D X_t = (1 - L^s)^D X_t$ , donde $s$ es el período estacional. Luego el modelo

estacional autorregresivo integrado de medias móviles de orden $(p, d, q)(P, D, Q)_s$, denotado por SARIMA$(p, d, q)(P, D, Q)_s$ es:

$$\phi(L)\Phi(L^s)\Delta^d\Delta_s^D X_t = \theta(L)\Theta(L^s)\varepsilon_t \tag{2.22}$$

donde $\phi(L)$ es el polinomio autorregresivo y $\theta(L)$ es el polinomio de media móvil, definidos en (2.16) y (2.18), $\Phi(L^s) = 1-\Phi_1 L^s-\Phi_2 L^{2s}-\cdots-\Phi_P L^{sP}$ es el polinomio autorregresivo estacional de orden $P$, $\Theta(L^s) = 1-\Theta_1 L^s-\Theta_2 L^{2s}-\cdots-\Theta_Q L^{sQ}$ es el polinomio de media móvil estacional de orden $Q$

### 2.5.4. Protocolo para la Identificación de los modelos SARIMA

Los modelos SARIMA$(p, d, q)(P, D, Q)_s$ pertenecen a un familia amplia de procesos lineales, se hace necesaria el uso de una estrategia de construcción de modelos. La estrategia involucra las siguientes etapas:

- Identificación de la estructura.
- Estimación de los parámetros.
- Diagnóstico del modelo.
- Pronóstico

Este ciclo iterativo fue descrito y popularizado por *Box and Jenkins* (1976), ver Figura 2–4. Para evaluar la calidad del ajuste teniendo en cuenta el número de parámetros estimados en el modelo y la verosimilitud, existe el criterio de información de *Akaike* o AIC por sus siglas del inglés *Akaike Information Criterion*: Cuanto más pequeño sea el valor del criterio de información, mejor será el modelo [10].

- **Identificación del Modelo SARIMA$(p, d, q)(P, D, Q)_s$:** Para el modelamiento de una serie de tiempo conocemos el período $s$, luego identificar el modelo es la fase crítica del proceso iterativo descrito en la Figura 2–4, lo primero será poder identificar los valores $d$ y $D$; de esa forma logramos que la serie de tiempo sea estacionaria.

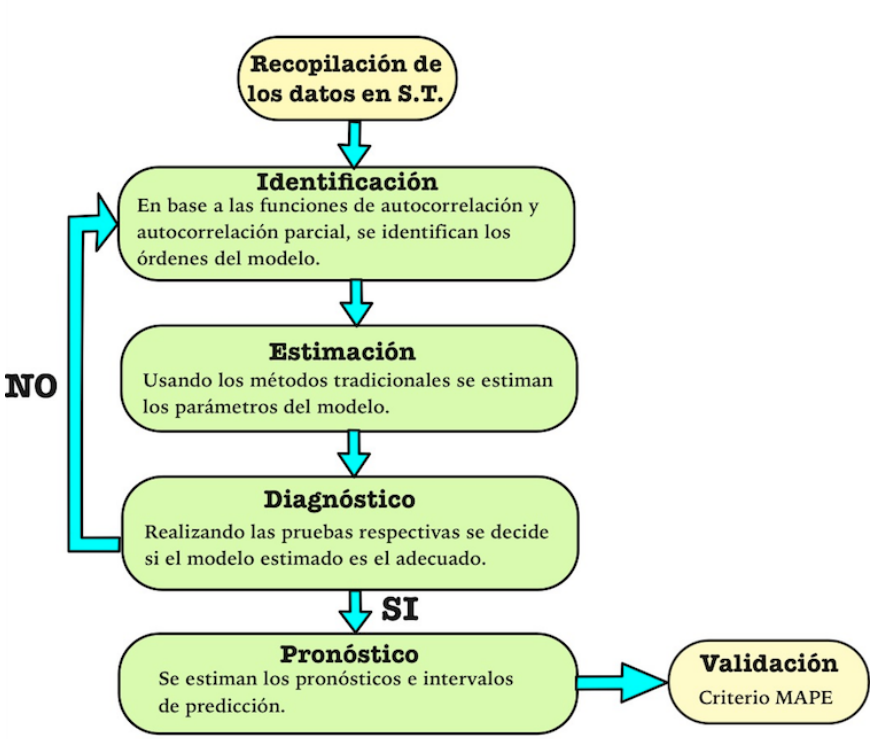


Figura 2–4: Ciclo Iterativo de *Box and Jenkins*

| | | | |
|---|---|---|---|
| AR(p) | | ACF | PACF |
| | 1) Mirar ACF<br>2) Mirar PACF para el orden p | Decae de acuerdo a una exponencial, sinusoidal. | Corte en el orden p. Esto es: toma valores cero k>p |
| MA(q) | | PACF | ACF |
| | 1) Mirar PACF<br>2) Mirar ACF para el orden q | Decae de acuerdo a una exponencial, sinusoidal. | Corte en el orden q. Esto es: toma valores cero k>q |
| SAR(P) | | ACF | PACF |
| | 1) Mirar ACF<br>2) Mirar PACF para el orden P | Decae en los múltiplos del periódo, exponencial o sinusoidal. | Corte en el orden P. Esto es: toma valores cero k>P |
| SMA(Q) | | PACF | ACF |
| | 1) Mirar PACF<br>2) Mirar ACF para el orden Q | Decae en los múltiplos del periódo, exponencial o sinusoidal. | Corte en el orden Q. Esto es: toma valores cero k>Q |

Figura 2–5: Identificación de los ordenes del modelo SARIMA

El número de diferencias necesarias para que el proceso se transforme estacionario, es alcanzado cuando el correlograma de la ACF y PACF decrecen rápidamente a cero, de forma heurística se recomienda hacer una o dos diferencias. Luego se debe proponer valores para los ordenes $p, q, P, Q$ descritos en (2.22). La Figura 2–5, nos proporciona una alternativa resumida para escoger valores de los ordenes $p, q, P, Q$, de acuerdo al proceso autorregresivo o media móvil, regular o estacional; la selección depende del comportamiento del ACF y PACF descritos en sus correlogramas.

- **Estimación del Modelo SARIMA$(p, d, q)(P, D, Q)_s$:** Teniendo un modelo tentativo para nuestra serie de tiempo, el siguiente paso es estimar sus parámetros, el método de máxima verosimilitud será usado para este propósito.
- **Diagnóstico del Modelo SARIMA$(p, d, q)(P, D, Q)_s$:** Luego de la estimación de los parámetros la tarea se centra en verificar si el modelo representa o no adecuadamente los datos, esto se consigue analizando los residuales del modelo estimado de (2.22), denotados por $r_t = \widehat{\varepsilon}_t$. Si el modelo es adecuado, los residuos $r_t$ deben estar próximos a $\varepsilon_t$ , y por tanto deben ser ruido blanco y no autocorrelacionados. Una prueba de hipótesis estadística para la significancia de las autocorrelaciones de los errores, es la prueba de *Ljung-Box* basado en el estadístico Chi-cuadrado.

$$Q(t) = n(n+2)\sum_{t=1}^{m}\frac{\hat{\rho}_t^2}{n-t} \sim \chi_m^2 \tag{2.23}$$

Donde $\hat{\rho}_t$ es el coeficiente de autocorrelación de los residuos, $m$ es el número de retrasos a ser probados. La hipótesis nula es que los errores son ruido blanco. Si se comprueba que el modelo es adecuado, se puede continuar con el procedimiento y calcular las predicciones. Caso contrario, se repite el proceso iterativo descrito por *Box and Jenkins*.

- **Pronóstico del Modelo SARIMA$(p,d,q)(P,D,Q)_s$:** Después de obtener el modelo y comprobar su validez, se puede proceder a predecir. La predicción óptima de $X_{n+h}$ mediante $\hat{X}_{n+h}$ es el valor esperado condicionado a que se conoce $X_1, X_2, \ldots, X_n$, esto es:

$$\hat{X}_{n+h} = \mathrm{E}[X_{n+h}|X_1, X_2, \ldots, X_n] \tag{2.24}$$

donde $h$ es el ńumero de periodos hacia adelante. El cálculo de las esperanzas condicionales se realizará usando la estimación de la ecuación (2.22), que es:

$$\hat{\phi}(L)\hat{\Phi}(L^s)\Delta^d\Delta_s^D X_t = \hat{\theta}(L)\hat{\Theta}(L^s)\varepsilon_t \tag{2.25}$$

Dado que los errores son ruido blanco, la distribución condicional de $X_{n+h}$ es normal con media $\hat{X}_{n+h}$ y varianza $\hat{\sigma}^2(1+\sum_{j=1}^{h-1}\hat{\psi}_j^2)$ donde $\hat{\psi}_j^2$, son los coeficientes que provienen de la multiplicación de los polinomios autorregresivos y medias móviles, regulares y estacionales [2]. Finalmente, conociendo la distribución de las predicciones podemos construir intérvalos de confianza.

- **Validación de un Modelo SARIMA$(p,d,q)(P,D,Q)_s$:** La validación es una técnica estadística para evaluar los resultados obtenidos del análisis de un modelo estadístico, para lograr esto se requiere dividir la muestra, en una muestra de entrenamiento donde se realiza el análisis estadístico y una muestra de prueba donde se evalúa la eficiencia del modelo. Para el caso del análisis de series de tiempo las predicciones realizadas por el modelo estimado SARIMA$(p,d,q)(P,D,Q)_s$ son la principal fuente de información para evaluar la eficiencia del modelo propuesto. La evaluación de la eficiencia de la predicción, se realiza mediante algunos criterios conocidos, que se describen a continuación:

$$\hat{X}_t \quad \text{con } t = n+1, n+2, \ldots, n+h$$

Las predicciónes del modelo estimado SARIMA$(p,d,q)(P,D,Q)_s$, así la muestra de entrenamiento es de tamaño $n$ y la muestra de prueba es de tamaño $h$. Luego $\varepsilon_t = X_t - \hat{X}_t$ es el error de predicción, bajo el modelo estimado SARIMA$(p,d,q)(P,D,Q)_s$. Luego se define el criterio para evaluar la eficiencia de la predicción como el porcentaje de error absoluto medio MAPE por sus siglas del inglés *Mean Absolute Percentage Error* [9], dado por:

$$MAPE = \frac{1}{h}\sum_{t=n+1}^{n+h}\frac{|\varepsilon_t|}{X_t} \tag{2.26}$$

Un MAPE relativamente pequeño, muestra la eficiencia de la predicción de un modelo.

## 2.6. Modelo ICA para Series de Tiempo

Cuando los datos observados son series de tiempo, estos poseen una estructura de autocovarianzas que las variables aleatorias no tienen, esta información adicional permite la estimación aún en el caso de que haya normalidad.

El modelo será expresado como:

$$X(t) = AS(t) \tag{2.27}$$

Donde $X(t) = (X_1(t), X_2(t), \ldots, X_p(t))$, es la serie de tiempo multivariada, con matrices de covarianza $C_k^X = \mathrm{E}[X(t)X(t-k)^T]$, donde cada $X_i(t)$ la $i$-ésima serie de tiempo. $S(t) = (S_1(t), S_2(t), \ldots, S_p(t))$ es la serie latente multivariada, con matrices de covarinza $C_k^S = \mathrm{E}[S(t)S(t-k)^T]$, donde cada $S_i(t)$ la $i$-ésima serie latente independiente, $i = 1, 2, \ldots, p$ y $A_{p\times p}$ matriz de mezcla.

**Suposiciones de ICA para series de tiempo:** Tenemos las siguientes suposiciones [11]:

1. Las series latentes $S_i(t)$ son estadísticamente independientes, luego las matrices $C_k^S$ son diagonales.
2. La matriz $A$, se asume de rango completo.

3. No hay ruido externo en el modelo
4. La suposición de no gaussianidad es reemplazada por cualquiera de las siguientes:
   *a*) Las series latentes $S_i(t)$ tiene distintas autocovarianzas (en particular todas distintas de cero).
   *b*) Las varianzas de las series latentes $S_i(t)$ son no estacionarias.

### 2.6.1. Estimación del Modelo ICA en Series de Tiempo

Para la estimación del modelo usaremos la suposición (4.a). Considerando el modelo ICA en series de tiempo, para series blanqueadas $Z(t)$.

$$Z(t) = AS(t) \tag{2.28}$$

Las series latentes independientes, están dadas por las combinaciones lineales de $Z(t)$.

$$S(t-k) = WZ(t-k), \quad \forall\ k \geq 0 \tag{2.29}$$

Donde $W_{p\times p}$ es una matriz ortogonal, tal que $W$ maximice la independencia estadística de $S(t)$.

La teoría ICA para series de tiempo, propone hacer que las correlaciones cruzadas en un tiempo de retraso fijo de las series latentes sean iguales a cero [19], esto equivale a que la matriz $W$ cancele los elementos fuera de la diagonal de las matrices $C_k^S$

- **Cálculo de Las Series Latentes Independientes:** Hay dos enfoques de aproximación de acuerdo al criterio de optimizacíon, que propone diagonalizar una o varias matrices de covarianza $C_k^S$.

  La primera aproximación fue introducida por *Tong* (1990), donde propone estimar $S(t)$ a través de la diagonalización de una de sus matrices de covarianza en un tiempo de retraso fijo.

**Planteamiento:** Sea la matriz de covarianza de las series blanqueadas $Z(t)$

$$C_k^Z = \mathrm{E}[Z(t)Z(t-k)^T]$$

Para un $k$ fijo, se debe descomponer esta matriz en sus valores propios; sin embargo debido a que $C_k^Z$ en general no es simétrica [23], se hace una modificacíon para conseguir que sea simétrica, en consecuencia la descomposición de valores propios estará bien definida. *Tong* propone la siguiente transformación:

$$\bar{C}_k^Z = \frac{1}{2}[C_k^Z + (C_k^Z)^T]$$

Luego como $W$ es ortogonal,

$$\begin{aligned}
\bar{C}_k^Z &= \frac{1}{2}[C_k^Z + (C_k^Z)^T] \\
&= \frac{1}{2}\left[\mathrm{E}\{Z(t)Z(t-k)^T\} + (\mathrm{E}\{Z(t)Z(t-k)^T\})^T\right] \\
&= \frac{1}{2}\left[\mathrm{E}\{Z(t)Z(t-k)^T\} + \mathrm{E}\{Z(t-k)Z(t)^T\}\right] \\
&= \frac{1}{2}\left[\mathrm{E}\{W^T S(t)S(t-k)^T W\} + \mathrm{E}\{W^T S(t-k)S(t)^T W\}\right] \\
&= W^T\left[\frac{1}{2}\left(\mathrm{E}\{S(t)S(t-k)^T\} + \mathrm{E}\{S(t-k)S(t)^T\}\right)\right] W \\
\bar{C}_k^Z &= W^T \bar{C}_k^S W
\end{aligned}$$

Donde $\bar{C}_k^S$ es diagonal $\forall\ k \geq 1$, debido a la suposición que $C_k^S$ son diagonales. Así:

$$\bar{C}_k^Z = W^T D W \quad \forall k \geq 1 \tag{2.30}$$

Luego de obtener (2.30) y una vez fijado $k$, se aplica la descomposición de valores propios para $\bar{C}_k^Z$, y los vectores propios de $\bar{C}_k^Z$, son los estimados de las filas de $W$. De esta forma las series latentes independientes son estimadas por:

$$\hat{S}(t) = \widehat{W} Z(t)$$

El algoritmo AMUSE que implementa esta aproximación fue propuesto por *Tong* (1990).

**El Algoritmo AMUSE**

1. Blanquear los datos observados de media cero, $Z(t) = VX(t)$.
2. Calcular la descomposición de los valores y vectores propios de la matriz de covarianza $\bar{C}_k^Z = \frac{1}{2}[C_k^Z + (C_k^Z)^T]$, fijando $k$.
3. La columnas de la matriz de separación $W$ están dadas por los vectores propios de $\bar{C}_k^Z$.

AMUSE encuentra $W$ de forma rápida y es eficiente cuando los vectores propios de $\bar{C}_k^Z$ son diferentes. Si los vectores propios no son diferentes, AMUSE no podrá estimar las series latentes; de esta manera AMUSE se considera sensible a la elección de $k$ [8].

La segunda aproximación intenta evitar el problema de la elección de $k$ en el algoritmo AMUSE. El algoritmo SOBI desarrollado por *Belouchrani* (1997) [1] quien declara "la robustez se incrementa significativamente mediante el procesamiento de un conjunto de matrices en lugar de una sola", propone considerar más retrasos y hacer una diagonalización conjunta de todas las matrices de covarianza $C_k^S,\ \forall\ k = 1, 2, \ldots, K$. Entonces la matriz ortogonal $\widehat{W}$ será el diagonalizador conjunto de $K$ matrices de covarianza de $\hat{S}(t)$, denotada por:

$C_k^{\hat{S}} = \mathrm{E}[\hat{S}(t)\hat{S}(t-k)^T], \qquad k = 1, 2, \ldots, K$. El algoritmo SOBI asume que, "las series latentes independientes $S_i(t)$ esta mutuamente no correlacionadas".

**Planteamiento:** Considere un conjunto de $K$ matrices de covarianza de las series blanqueadas $Z(t)$.

$$C_k^Z = \mathrm{E}[Z(t)Z(t-k)^T], \qquad k = 1, 2, \ldots, K$$

Realizamos la misma transformación de AMUSE.

$$\bar{C}_k^Z = W^T \bar{C}_k^S W \Longrightarrow \bar{C}_k^S = W \bar{C}_k^Z W^T$$

Diagonalizar $\bar{C}_k^S$ equivale a minimizar, el criterio de diagonalización conjunta, dado por:

$$\mathscr{J}(W) = \sum_{k=1}^{K} off\,(W\bar{C}_k^Z W^T)$$

Donde $off\,(M) = \sum_{i\neq j} m_{ij}^2$. Luego se dice que: "Una matriz ortogonal $W$ es un diagonalizador conjunto de $\{\bar{C}_1^S, \bar{C}_2^S, \ldots, \bar{C}_K^S\}$, si éste minimiza el criterio de diagonalización conjunta $\mathscr{J}$" [4].

Así estimamos $\widehat{W}$ que minimiza $\mathscr{J}$. De esta forma las series latentes independientes son estimadas por:

$$\hat{S}(t) = \widehat{W}Z(t)$$

*Belouchrani* (1997), muestra que: "Si existen dos componentes diferentes que tiene distintas covarianzas para al menos un tiempo de retraso $k$, el diagonalizador existe y es único" [1].

SOBI utiliza el diagonalizador conjunto aproximado o JAD por sus siglas en inglés *Joint Approximate Diagonalization* propuesto por *Souloumiac and Cardoso* (1993), para encontrar $W$, usando la técnica de *Jacibi* generalizada JAD usa rotaciones de *Givens* en el proceso de optimización [1].

**El Algoritmo SOBI**

1. Blanquear los datos observados de media cero, $Z(t) = VX(t)$.
2. Fijar el conjunto de tiempos de retraso $k \in \{1, 2, \ldots, K\}$ y estimar el conjunto de matrices de covarianzas de datos blanqueados para sus respectivos tiempos de retraso, $\left\{C_k^Z | k = 1, 2, \ldots, K\right\}$.
3. Una matriz ortogonal $\widehat{W}$ es obtenida como la solución de problema de la diagonalización conjunta $\mathscr{J}$, mediante rotaciones de *Givens*.
4. Estimar las componentes de acuerdo a, $\hat{S}(t) = \widehat{W}Z(t)$

# Capítulo 3
# METODOLOGÍA

El objetivo de esta tésis es aplicar la técnica del ICA adaptada a series de tiempo, para analizar la capacidad predictiva en el área de economía y finanzas de Puerto Rico. Para esto se seleccionó dos series de tiempo multivariadas.

## 3.1. Conjunto de Datos Analizados

Se seleccionaron dos conjuntos de datos. El primer conjunto de datos está relacionado al problema de electricidad y el segundo relacionado con indicadores económicos, ambos de Puerto Rico.

- **Series de Tiempo de Electricidad:** La autoridad de energía eléctrica de Puerto Rico (AEE), mediante el departamento de proyecciones y estadísticas almacenan datos longitudinales o series de tiempo, disponibles en su página *web*.

  Se consideró el consumo mensual de energía eléctrica, medido en millones kilowatts hora (mkWh), para los sectores residencial, comercial e industrial. Al momento de realizar el análisis de los datos, el historial de las series de tiempo mensuales comprendendían desde el periódo del 1 julio de 1999 hasta 1 marzo del 2014, para un total de 177 datos.

- **Series de Tiempo de Indicadores Económicos:** El banco gubernamental de fomento para Puerto Rico, presentan datos mensuales de diferentes indicadores económicos, estos datos se encuentran disponibles en su pagina *web*.

  Se consideró el índice de precios al consumidor (IPC), que es un índice económico en el cuál se valoran los precios de un conjunto de productos y servicios conocido

como canasta familiar, agrupados en ocho grupos: Alimentos y bebidas, Alojamiento, Vestido, Transportación, Cuidado médico, Entretenimiento, Educación y comunicación, Otros artículos y servicios. La importancia de estudiar el IPC conlleva medir la inflación en el costo de vida y estimar el poder adquisitivo del dólar [7]. Al momento de realizar el análisis estadístico, el historial de los datos fue de julio de 2004 a abril del 2014, para un total de 118 datos.

## 3.2. Procedimiento

Para el análisis de los datos, mediante la técnica del ICA en series de tiempo, tenemos las siguientes etapas generales:

1. Dada una serie de tiempo multivariada, se estiman las series latentes mediante el modelo ICA para series de tiempo, usando los algoritmos AMUSE y SOBI.
2. Se hacen predicciones para cada serie latente, aquí se procede con el protocolo de indentificación de modelos SARIMA para cada una de las series latentes.
3. Se predicen las series de tiempo multivariadas, con las predicciones de las series latentes, y con la estimación de la matriz de mezcla $A$ del modelo ICA para series de tiempo.

A continuación se describen estas etapas generales de forma explicita para la series de tiempo multivariada de consumo de energía eléctrica.

- Inicialmente la serie de tiempo multivariada de consumo de energía eléctrica está representada en el vector aleatorio observado $X = (X_1(t), X_2(t), X_3(t))$, donde cada componente representa a cada serie de tiempo de consumo de energía eléctrica en los sectores residencial, comercial e industrial respectivamente. Se realiza, el pre procesamiento, centrando y reescalando los datos originales; la verificación de no nulidad de la correlación cruzada entre las series de tiempo mediante la significancia de la función de autocorrelación cruzada utilizando los correlogramas.

- Mediante los algoritmos AMUSE y SOBI, se estiman las series latentes que serán los componentes del vector aleatorio $S = (S_1(t), S_2(t), S_3(t))$; también se estima la matriz de separación $W_{3\times3}$ y la matriz de mezcla $W_{3\times3}^{-1}$.
- Para cada serie latente $S_1(t)$, $S_2(t)$ y $S_3(t)$ se procede a realizar el ciclo iterativo de *Box and Jenkins*, esto es, identificación, estimación, diagnóstico, predicción y validación.
    - En la fase de identificación, debido a la presencia de los componentes de tendencia y estacional de cada serie latente, tenemos identificados posibles valores de $s, p, d, q, P, D, Q$, órdenes del modelo SARIMA. Esta fase es importante debido a la forma heurística de identificacíon y al hecho de seleccionar un modelo parsimonioso. Las herramientas que se usan para identificar los órdenes, son los correlogramas de las funciones de autocorrelación y autocorrelación parcial. Criterios como el $AIC$ y $log-likelihood$ también son usados para la selección de los modelos.
    - En la fase de estimación de los parámetros del modelo SARIMA, el método de estimación es máxima verosimilitud.
    - La fase de diagnóstico del modelo estimado SARIMA se realiza mediante el análisis de los residuales o innovaciones del modelo. Este análisis comprende que los errores se distribuyan como un ruido blanco, es decir tengan media cero, varianza constante y estén no correlacionados. Por lo general es siempre cierto que los residuos tengan media cero; para las series analizadas la varianza es constante; en cambio para verificar la no correlación, debemos realizar la prueba de hipótesis estadística de *Ljung-Box*. Si los residuos se comportan como un ruido blanco, procedemos a la fase de predicción, de lo contrario debemos buscar otros parámetros para el modelo SARIMA, esto significa volver a la fase de indentificacíon, estimación y diagnóstico para el nuevo modelo propuesto.

    - La fase de predicción se realiza luego de haber encontrado el modelo adecuado SARIMA. Se predice un número fijo de pasos en el futuro, usualmente 12 predicciones y se encuentran los intervalos de predicción [3].

- Las predicciones de cada serie latente $S_1(t)$, $S_2(t)$ y $S_3(t)$ son combinadas, ponderando cada una con las entradas de la matriz de mezcla $W_{3\times3}^{-1}$, esto es equivalente a multiplicar $(W_{3\times3}^{-1})(S_{3\times1})$. Obteniendo así las predicciones de las series de tiempo de consumo de energía eléctrica $X$.
- En la fase de validación, la muestra de entrenamiento consta de 165 datos, es decir, dejamos las 12 últimas observaciones correspondientes al período abril del 2013 a marzo del 2014, esta elección de los tamaños de muestra de entrenamiento y prueba son debido a que los modelos SARIMA se consideran eficientes para períodos de predicción a corto plazo [14]. La validación es útil para encontrar los errores de predicción del modelo, el criterio de eficiencia de las predicciones es MAPE. La validación se realiza para las series originales.

Las etapas para la series de tiempo multivariada de índice de precios al consumidor, son similares a las de la serie consumo de energía eléctrica. El vector aleatorio observado es $X = (X_1(t), X_2(t), \ldots, X_8(t))$, donde cada componente representa a cada una de las ocho series de tiempo del IPC. Mediante el algoritmo SOBI se estima las series latentes, componentes del vector $S = (S_1(t), S_2(t), \ldots, S_8(t))$, además se estima la matriz de separación $W_{8\times8}$. Luego se identifican modelos SARIMA para predecir cada una de las series latentes; para finalmente encontrar las predicciones de la series de tiempo índice de precios al consumidor $X$.

La validación del modelo, tiene una muestra de entrenamiento de 106 datos, es decir, dejamos las 12 últimas observaciones correspondientes al período mayo del 2013 a abril del 2014; la validación se considera para poder calcular el MAPE y ver la eficiencia del modelo.

Debido a que ésta serie, tiene ocho columnas, se considera el problema de ordenamiento de las series latentes, para una reducción en el número de estas.

**El problema de Ordenar:** Para poder ordenar las series latentes (ICs) en términos de la importancia para explicar los datos, *Back and Weigen* (1997), proponen ordenar los ICs en términos de sus varianzas explicadas [8], así los ICs más importantes serán aquellos que expliquen la máxima varianza de los datos.

Bajo esta reducción en el número de series latentes, se reconstruyen las series originales, es decir $\hat{X} = A_{8\times 8}S$, donde $S$ contiene las series latentes más importantes, y ceros en lugar de las series latentes menos importantes. Luego se calculan predicciones de las series originales del IPC bajo el modelo SARIMA-SOBI reducido.

La implementación computacional de la metodología, se realizó en el *software* R, usando la librería JADE [16], mostrada en el Apéndice B.

# Capítulo 4
# RESULTADOS

En este capítulo se muestran los resultados obtenidos durante esta investigación en relación a las series de tiempo multivariadas del consumo de energía eléctrica e índice de precios al consumidor de Puerto Rico.

## 4.1. Series de Tiempo de Electricidad

La serie de tiempo multivariada mensual, consumo de energía eléctrica (nKwh), en el área residencial, comercial e industrial, durante el periódo del 1 julio de 1999 hasta 1 marzo del 2014, comprende 177 datos. La Figura 4–1, muestra la evolución de las tres series de tiempo. La serie residencial es estacional, la serie comercial es estacional con una ligera tendencia creciente, la serie industrial tiene tendencia lineal decreciente en la mitad superior del período de observación.

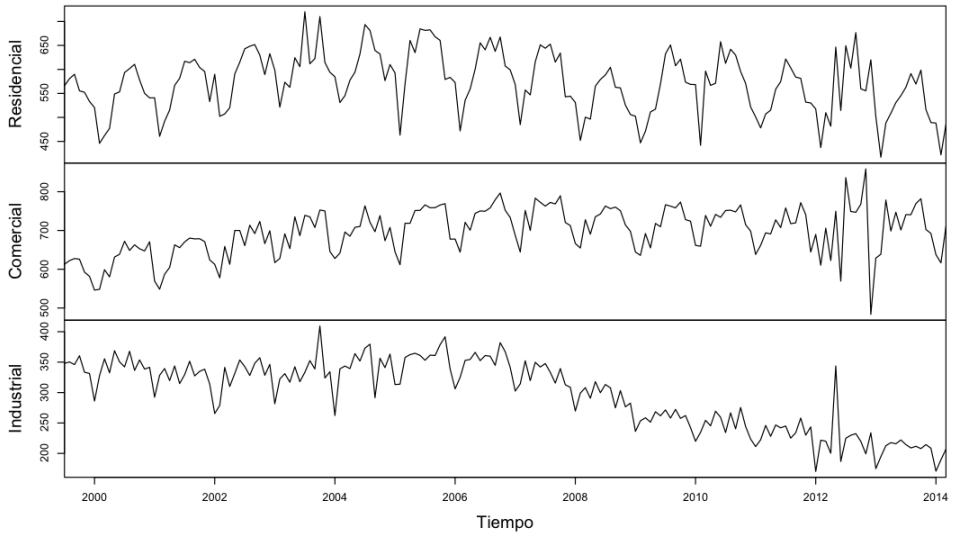


Figura 4–1: Series de Tiempo de Consumo de Energía Eléctrica (mkwh)

La Figura 4–2, muestra las autocorrelaciones y correlaciones cruzadas de las series de consumo de energía eléctrica; en la diagonal están las funciones de autocorrelación de las tres series, de las cuales, la residencial y comercial muestran decaimiento amortiguado lento y la serie industrial decaimiento lento hacia cero, lo que es evidencia de no estacionariedad. Por otro lado las correlaciones cruzadas fuera de la diagonal, son estadísticamente diferentes de cero, por tanto las series de tiempo son dependientes entre sí.

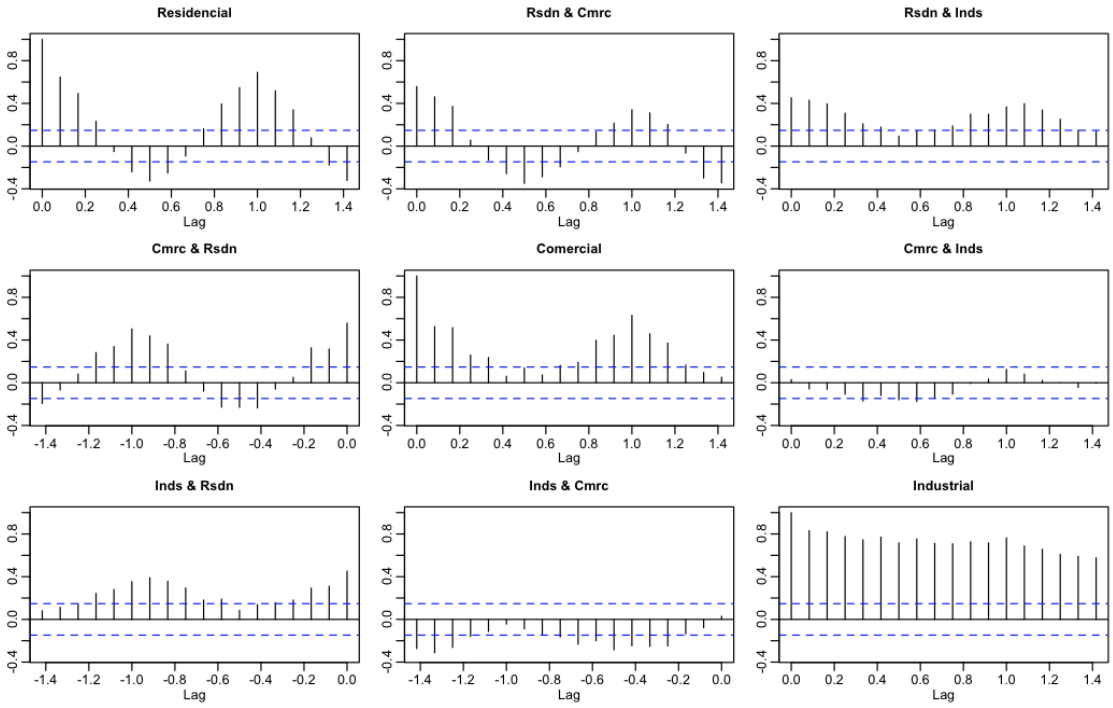


Figura 4–2: Autocorrelaciones y Correlaciones Cruzadas de las Series de Tiempo de Consumo de Energía Eléctrica (mkwh)

La estimación de las series latentes se realizó considerando el algoritmo SOBI con $k = 100$ matrices de covarianza para la diagonalización conjunta ([5],[22]).

La Figura 4–3, muestra las tres series latentes estimadas, la serie latente $S_1$ presenta el componente de tendencia decreciente, la serie latente $S_2$ presenta el componente estacional y la serie latente $S_3$ presenta el componente estacional.

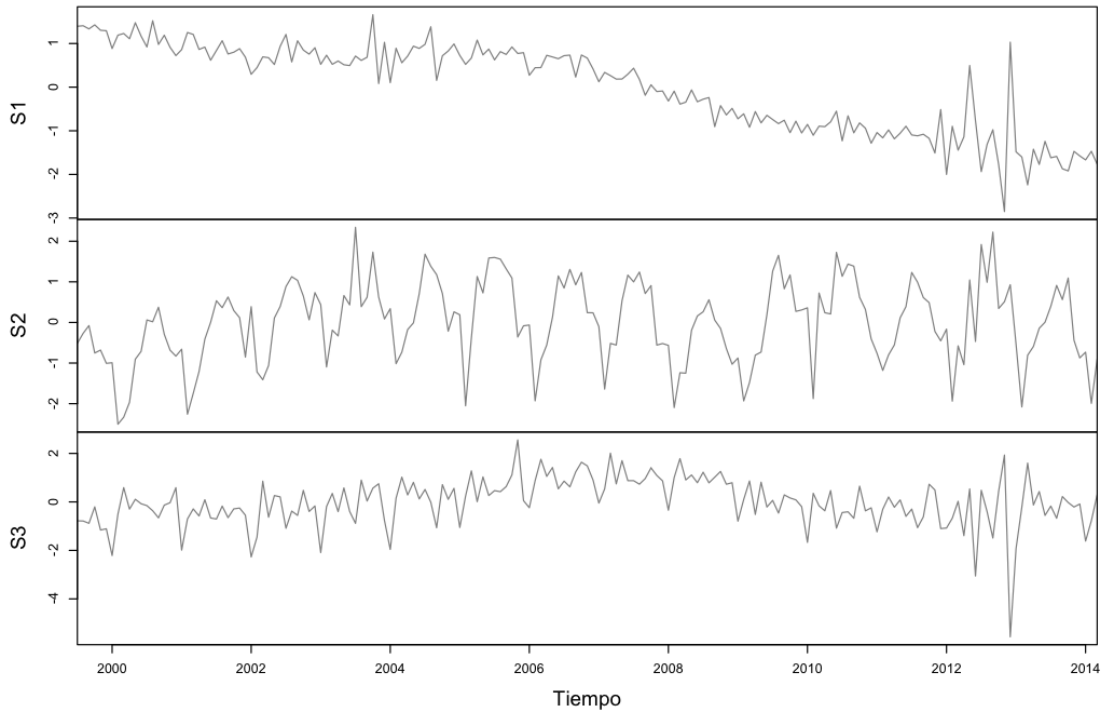


Figura 4–3: Series Latentes de Consumo de Energía Eléctrica, Estimadas mediante SOBI $k = 100$

La estimación de la matriz de separación $W$ se muestra a continuación (4.1). Esta matriz sirve para calcular la matriz de mezcla $W^{-1}$.

$$W = \begin{pmatrix} 0.2358493 & -0.51473337 & 0.8158944 \\ 1.0558134 & 0.07671076 & -0.3396182 \\ -0.9167757 & 1.15268738 & 0.7768874 \end{pmatrix} \tag{4.1}$$

En la Figura 4–4, se muestran las autocorrelaciones y correlaciones cruzadas de las series latentes $S_1, S_2$ y $S_3$. Las autocorrelaciones de $S_1$ y $S_3$ presentan un decaimiento lento hacia cero, la autocorrelación de $S_2$ decae de forma amortiguada; estas características son usadas para la indentificación de modelos SARIMA para cada serie latente. La aproximación a la independencia de las series latentes, se visualiza en las correlaciones cruzadas, en donde la mayoría de las correlaciones se aproximan a cero; esto permite hacer un análisis univariado a cada serie latente.

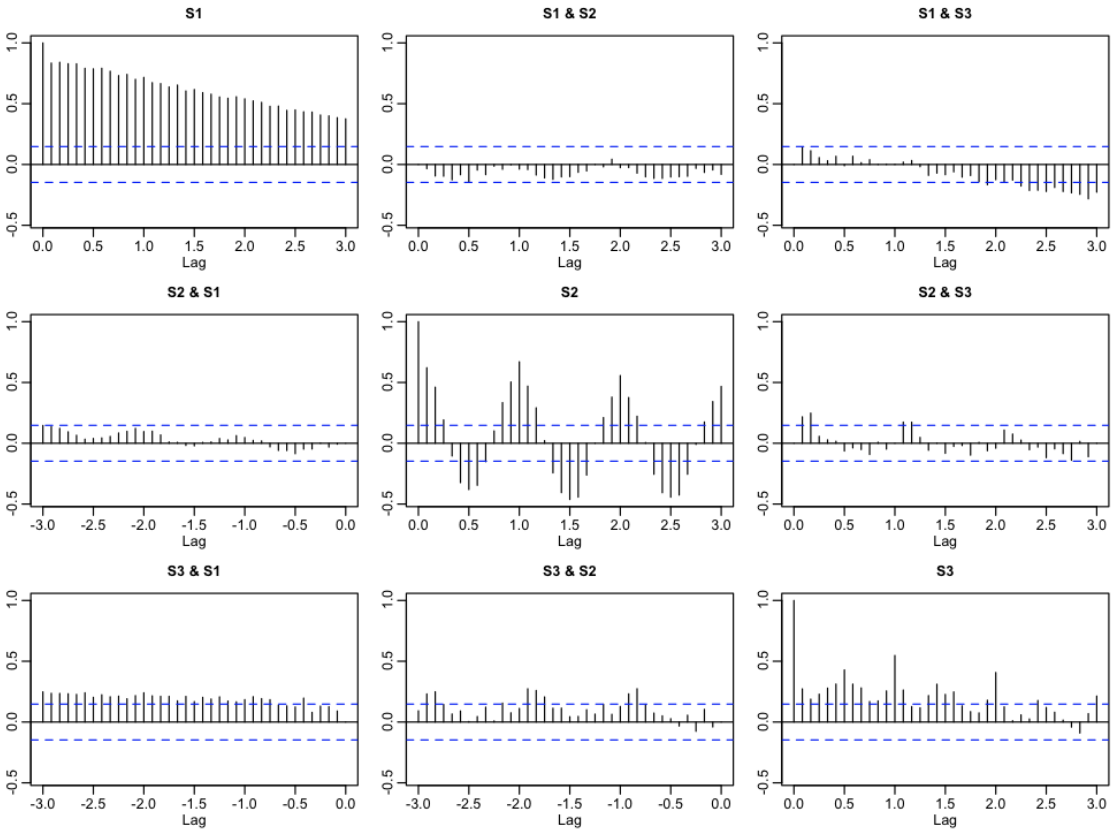


Figura 4–4: Autocorrelaciones y Correlaciones Cruzadas Estimadas de las Series Latentes de Consumo de Energía Eléctrica

Para cada una de las series latentes, se realizó, el procedimiento de identificación, estimación y diagnóstico del modelo SARIMA, los detalles de este proceso se muestran en el Apéndice A.

El Cuadro 4–1, muestra los modelos óptimos SARIMA seleccionados y ajustados para cada serie latente, así como sus correspondientes, estimación de la varianza del modelo, el *log-likelihood*, y el AIC respectivamente, medidas que se consideran para seleccionar modelos adecuados.

Cuadro 4–1: Selección de Modelos SARIMA para las Series Latentes de Consumo de Energía Eléctrica

| Series | SARIMA$(p,d,q)(P,D,Q)_s$ | $\hat{\sigma}^2$ | log-likelihood | AIC |
|---|---|---|---|---|
| $S_1$ | SARIMA$(6,1,1)(0,1,1)_{12}$ | 0.1332 | -74.29 | 166.57 |
| $S_2$ | SARIMA$(0,1,2)(0,1,1)_{12}$ | 0.2256 | -117.82 | 243.64 |
| $S_3$ | SARIMA$(5,1,1)(0,1,1)_{12}$ | 0.4304 | -173.36 | 362.71 |

La Figura 4–5, muestra las predicciónes y los intervalos de predicción al 95 % de confianza de las tres series latentes. Las predicciones siguen el comportamiento de tendencia y estacional de las series latentes.

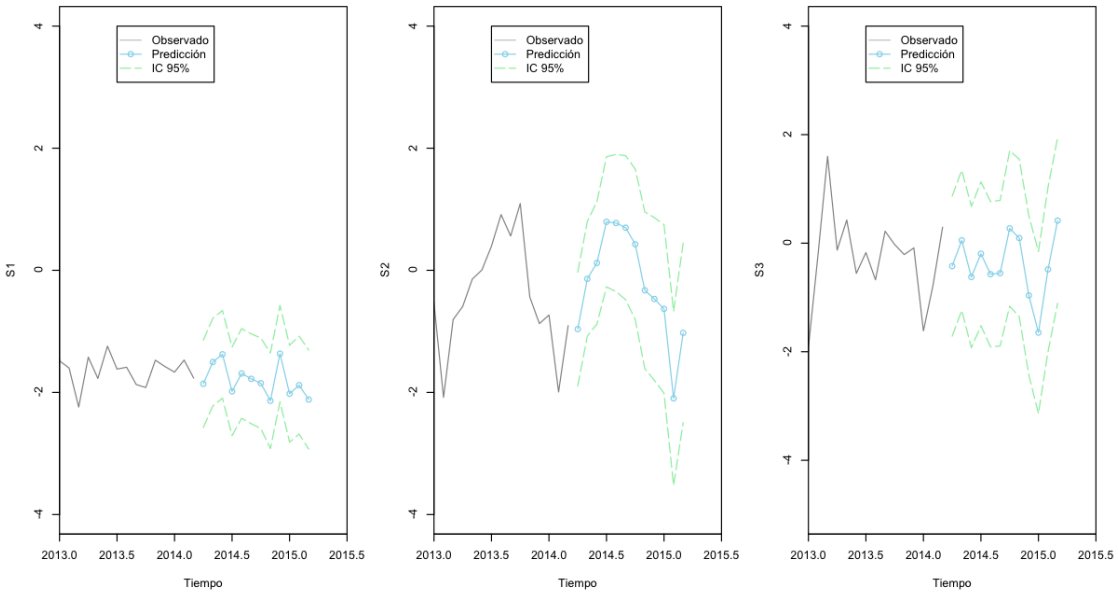


Figura 4–5: Predicciones e Intervalos de Predicción del 95 % de Confianza de las Series Latentes de Consumo de Energía Eléctrica, para el período abril del 2014 a marzo del 2015

Puesto que el objetivo es realizar predicciones para las series de tiempo originales, realizamos ahora la fase de retornar. Dado que tenemos las predicciones

de las series latentes $h = 12$ pasos adelante, esto es el vector de predicciones $S(t) = (S_1(t), S_2(t), S_3(t))$, con $t = 1, 2, \ldots, 12$, y tenemos la matriz de mezcla $W^{-1}$, obtenida de (4.1), obtenemos las predicciones de las series de tiempo originales $X(t) = (X_1(t), X_2(t), X_3(t))$, de la siguiente forma:

$$X(t) = W^{-1}S(t), \text{ con } t = 1, 2, \ldots, 12$$

El Cuadro 4–2, muestra los valores numéricos de las predicciones de las series de tiempo de consumo de energía eléctrica en los sectores residencial, comercial e industrial, para el período abril del 2014 a marzo del 2015, en Puerto Rico.

Cuadro 4–2: Predicciones del Consumo de Energía Eléctrica, según los Sectores, para el período abril del 2014 a marzo del 2015, en Puerto Rico, mediante los Modelos SARIMA-SOBI

| Período | Residencial | Comercial | Industrial |
|---|---|---|---|
| Apr 2014 | 476.72 | 681.22 | 185.82 |
| May 2014 | 534.66 | 726.02 | 221.26 |
| Jun 2014 | 549.16 | 706.14 | 214.67 |
| Jul 2014 | 578.87 | 764.24 | 198.54 |
| Aug 2014 | 581.70 | 741.58 | 204.97 |
| Sep 2014 | 575.48 | 741.13 | 200.26 |
| Oct 2014 | 562.11 | 765.53 | 212.87 |
| Nov 2014 | 511.16 | 734.12 | 188.17 |
| Dec 2014 | 512.88 | 668.11 | 202.95 |
| Jan 2015 | 486.94 | 648.04 | 152.78 |
| Feb 2015 | 409.14 | 633.34 | 174.31 |
| Mar 2015 | 472.07 | 718.66 | 190.80 |

La Figura 4–6 muestra las predicciónes y el intervalo de predicción al 95 % de confianza, para la serie de tiempo consumo de energía eléctrica en el sector residencial para el período abril del 2014 a marzo del 2015 en Puerto Rico, estas predicciones muestran la estacionalidad del consumo, así como una ligera baja.

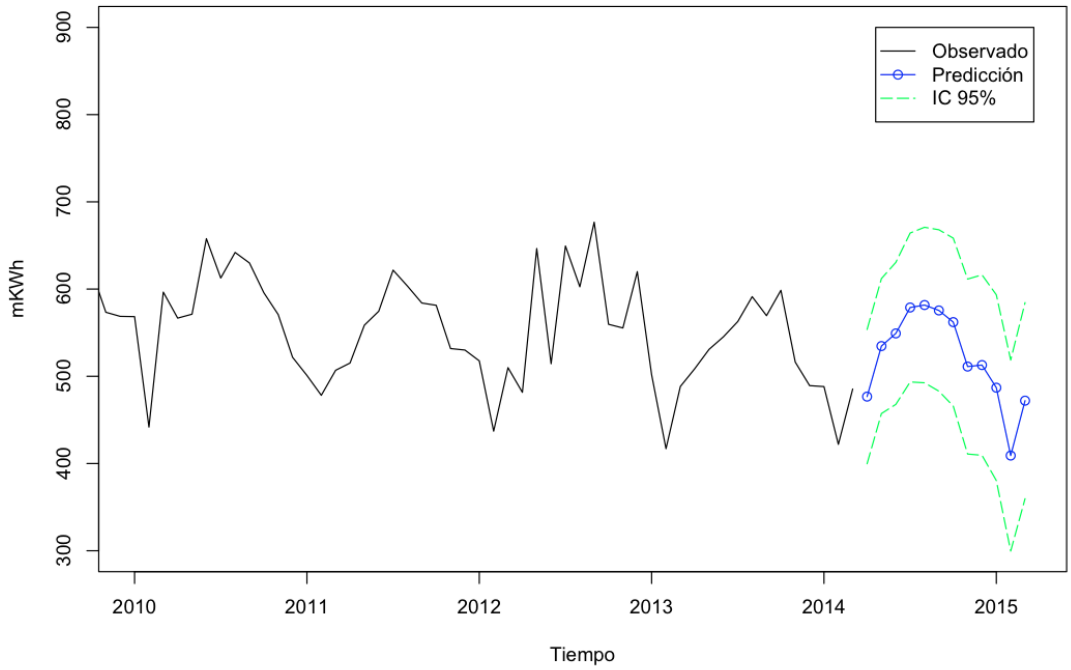


Figura 4–6: Predicciones e Intervalos de Predicción del 95 % de Confianza de la Serie de Tiempo de Consumo de Energía Eléctrica en el sector Residencial, para el período abril del 2014 a marzo del 2015

La Figura 4–7 muestra las predicciónes y el intervalo de predicción al 95 % de confianza, para la serie de tiempo consumo de energía eléctrica en el sector comercial para el período abril del 2014 a marzo del 2015 en Puerto Rico, estas predicciones muestran la estacionalidad del consumo sin tendencia.

La Figura 4–8 muestra las predicciónes y el intervalo de predicción al 95 % de confianza, para la serie de tiempo consumo de energía eléctrica en el sector industrial para el período abril del 2014 a marzo del 2015 en Puerto Rico, estas predicciones muestran una ligera tendencia a la baja y un componente estacional imperceptible.

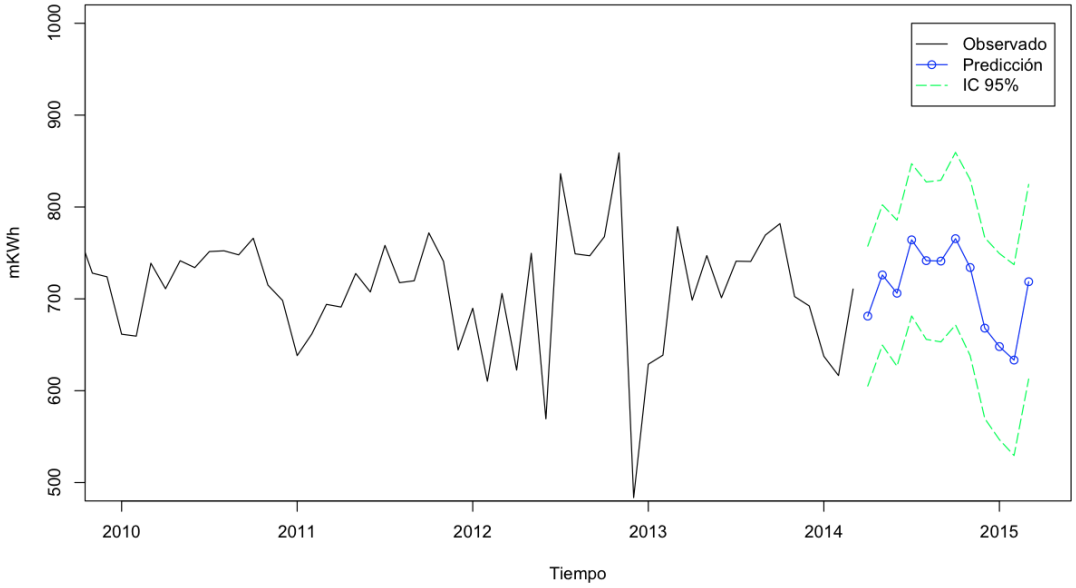


Figura 4–7: Predicciones e Intervalos de Predicción del 95 % de Confianza de la Serie de Tiempo de Consumo de Energía Eléctrica en el sector Comercial, para el período abril del 2014 a marzo del 2015

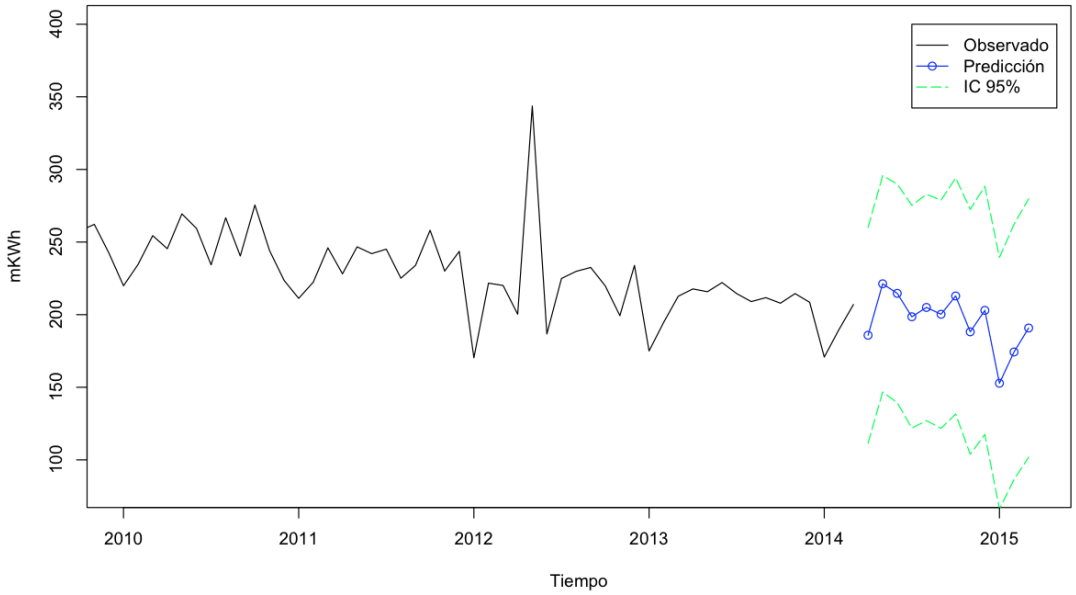


Figura 4–8: Predicciones e Intervalos de Predicción del 95 % de Confianza de la Serie de Tiempo de Consumo de Energía Eléctrica en el sector Industrial, para el período abril del 2014 a marzo del 2015

Hay que destacar que las predicciones para el sector residencial, presentan el componente estacional, en agosto del 2014 estará llegando al pico más alto con un consumo de 581.7 mKwh, mientras que en febrero del 2015 se llega al pico más bajo con un consumo de 409.14 mKwh. El sector comercial, tendrá valores de consumo altos entre julio del 2014 y noviembre del 2014, y tendrá su pico más bajo en febrero del 2015 con 633 mKwh. El sector industrial tiene tendencia decreciente, siendo el pico más bajo en enero del 2015 con 152.78 mKwh.

Para la fase de validación del modelo SARIMA-SOBI, consideramos una muestra de entrenamiento desde el período de julio 1999 a marzo del 2013, que hacen un total de 165 observaciones, luego la muestra de prueba comprende desde abril del 2013 a marzo del 2014, que representan 12 observaciones.

El Cuadro 4–3 muestra la medida de validación MAPE para el modelo SARIMA-SOBI para la serie de tiempo multivariada de consumo de energía eléctrica en los tres sectores. El modelo presenta un buen comportamiento en los sectores residencial y comercial con porcentajes de error absoluto medio (MAPE) bastante pequeños de 4.3 % y 6.7 %, comparados con el industrial que presenta un 14.6 %, esto debido a que la serie industrial presenta tendencia, complicando el modelamiento.

Cuadro 4–3: Validación de los Modelos SARIMA-SOBI de Consumo de Energía Eléctrica, para el período abril del 2013 a marzo del 2014

| Criterio | Residencial | Comercial | Industrial |
|---|---|---|---|
| MAPE | 0.04315983 | 0.06654076 | 0.14639497 |

Para comparar el modelo SARIMA-SOBI con SARIMA-AMUSE, mediante la validación de modelos; se consideran los modelos con los mismos ordenes $p, d, q, P, D, Q, s$. Las estimaciones de las series latentes, la matriz de separación, el ajuste SARIMA y predicciones de la serie de tiempo consumo de energía eléctrica, bajo el modelo SARIMA-AMUSE, con $k = 1, 2$ [5], se muestra en el Apéndice A.

El Cuadro 4–4, muestra la medida de validación MAPE para el modelo SARIMA-AMUSE con $k = 1, 2$, para la serie de tiempo multivariada de consumo de energía eléctrica en los tres sectores. En el sector residencial los MAPE son bastante pequeños de 4.3 % y 5.6 %, en el sector comercial 5.2 % y 4.1 %, en el sector industrial 8.2 % y 7.6 %, respectivamente para $k = 1$ y $k = 2$.

Cuadro 4–4: Validación de los Modelos SARIMA-AMUSE de Consumo de Energía Eléctrica , para el período abril del 2013 a marzo del 2014

| Retraso | Criterio | Residencial | Comercial | Industrial |
|---|---|---|---|---|
| $k = 1$ | MAPE | 0.04357919 | 0.05199041 | 0.08191397 |
| $k = 2$ | MAPE | 0.05633164 | 0.04095660 | 0.07600404 |

Por otro lado al contrastar los modelos SARIMA-SOBI y SARIMA-AMUSE, se describe un mejor ajuste y predicción bajo el modelo SARIMA-SOBI, para el sector residencial con un error porcentual promedio de 4.32 % ligeramente menor que el de los modelos SARIMA-AMUSE con 4.38 % y 5.63 %; en el sector comercial, el modelo SARIMA-AMUSE con $k = 2$ tiene una ligera ventaja sobre el SARIMA-SOBI de 4.1 % a 6.7 %; en el sector industrial el modelo SARIMA-SOBI muestra una eficiencia pobre comparado con los otros, de un 14.6 % a un 8.2 % y 7.6 %.

## 4.2. Series de Tiempo de Índice de Precios

La serie de tiempo multivariada mensual, de índices de precio al consumidor en los grupos de alimentos y bebidas, alojamiento, vestido, transportación, cuidado médico, entretenimiento, educación y comunicación, otros artículos y servicios; durante el periódo de julio de 2004 hasta abril del 2014, comprende 118 datos.

La Figura 4–9 muestra la evolución de las series de índices de precio al consumidor. Todas tienen tendencia creciente, excepto la serie del IPC del grupo vestido que tiene tendencia decreciente; no se observa componente estacional.

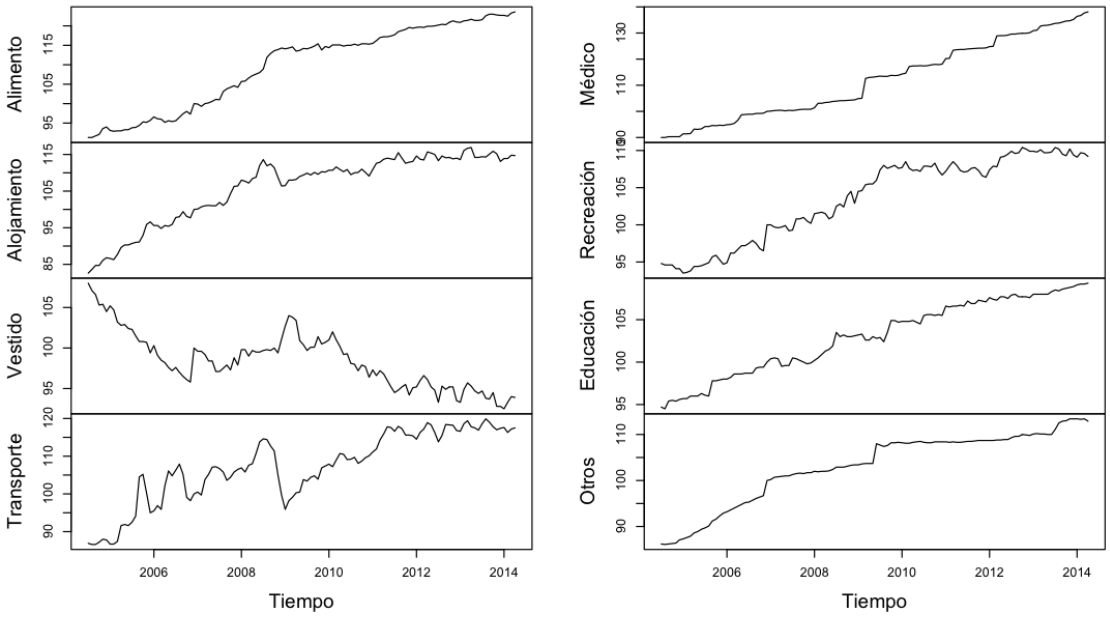


Figura 4–9: Series de Tiempo de Índices de Precios al Consumidor (%)

Esta serie de tiempo multivariada tiene una estructura de autocorrelación y correlación cruzada significativa. En el Apéndice A (Figuras A–10 y A–11) se muestran las ACF de la serie, las cuales verifican este hecho .

La Figura 4–10, muestra la estimación de las series latentes, mediante el SOBI con $k = 100$ matrices de retraso para la diagonalización conjunta; la serie latente $S_1$ presenta tendencia creciente, $S_2$ tendencia creciente y luego decreciente, el resto de las series latentes se comportan aleatoriamente. Las ACF de las series latentes estimadas se muestran en el Apéndice A (Figuras A–12 y A–13).

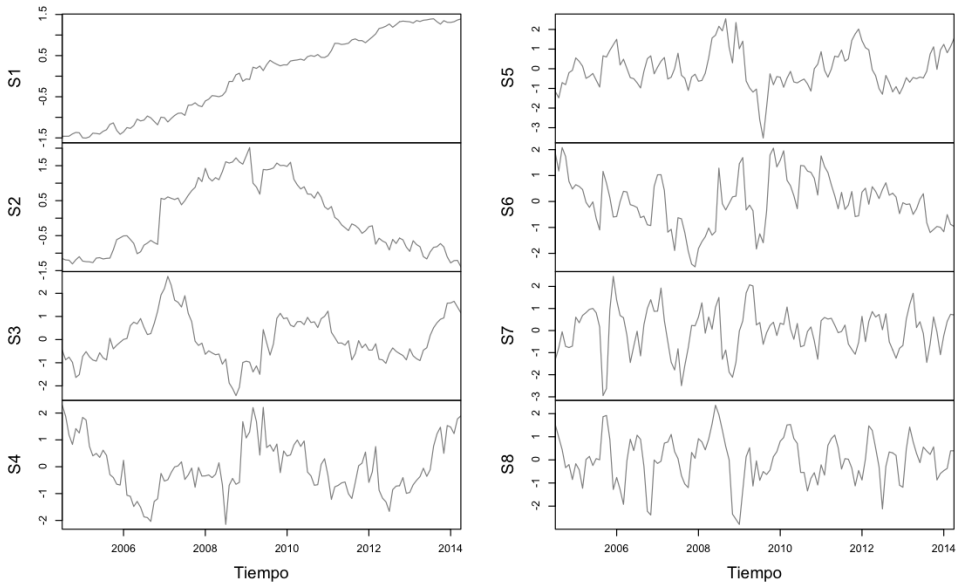


Figura 4–10: Series Latentes del Índices de Precios al Consumidor, Estimadas mediante SOBI con $k = 100$

La estimación de la matriz de separación, está dada por (4.2).

$$W = \begin{pmatrix} 0.66 & -0.11 & -0.08 & 0.10 & 0.35 & 0.50 & -0.12 & -0.44 \\ 0.24 & 0.66 & 0.33 & -0.08 & -2.10 & 0.41 & -0.09 & 1.33 \\ -3.68 & -1.66 & 0.48 & -0.15 & 0.37 & -1.20 & 2.20 & 4.56 \\ 3.24 & -2.32 & 1.28 & 0.24 & 1.91 & -3.63 & -2.18 & 3.69 \\ 1.62 & -0.59 & 0.82 & 0.43 & -1.15 & -5.49 & 4.40 & 1.42 \\ -3.25 & -0.92 & 0.89 & -0.08 & -2.23 & 2.39 & 6.45 & -1.85 \\ -5.73 & 5.93 & 0.75 & -2.78 & 4.20 & -1.42 & 1.34 & -0.97 \\ -2.48 & 0.39 & 2.21 & 2.55 & 2.67 & -1.18 & -1.12 & 1.15 \end{pmatrix} \tag{4.2}$$

El procedimiento para la identificación, estimación y diagnóstico del modelo SARIMA se realizó iterativamente hasta lograr que los modelos ajusten adecuadamente a las series latentes, los detalles se muestran en el Apéndice A.

El Cuadro 4–5 muestra los modelos SARIMA seleccionados para las series latentes del IPC, bajo los criterios de la varianza estimada del modelo, *log-likelihood* y AIC.

Cuadro 4–5: Selección de Modelos SARIMA para las Series Latentes de Índices de Precios al Consumidor

| Series | SARIMA$(p,d,q)(P,D,Q)_s$ | $\hat{\sigma}^2$ | log-likelihood | AIC |
|---|---|---|---|---|
| $S_1$ | SARIMA$(5,1,2)(0,0,0)_{12}$ | 0.005574 | 137 | -258 |
| $S_2$ | SARIMA$(1,0,0)(1,0,0)_{12}$ | 0.04884 | 8.87 | -9.73 |
| $S_3$ | SARIMA$(1,0,0)(0,0,0)_{12}$ | 0.1754 | -65.6 | 137.2 |
| $S_4$ | SARIMA$(4,1,2)(0,0,0)_{12}$ | 0.3215 | -100.33 | 214.67 |
| $S_5$ | SARIMA$(1,0,0)(0,0,0)_{12}$ | 0.4004 | -113.9 | 233.8 |
| $S_6$ | SARIMA$(5,1,0)(0,0,0)_{12}$ | 0.4419 | -118.47 | 248.93 |
| $S_7$ | SARIMA$(4,1,2)(0,0,0)_{12}$ | 0.548 | -134.3 | 282.59 |
| $S_8$ | SARIMA$(4,1,2)(0,0,0)_{12}$ | 0.4661 | -126.27 | 266.54 |

Las Figuras 4–11 y 4–12, muestra las predicciones de las series latentes del IPC, separadas en grupos de cuatro, para el período de mayo del 2014 a abril del 2015, mediante SOBI con $k = 100$.

Puesto que el objetivo es realizar predicciones para las series de tiempo originales, realizamos ahora la fase de retornar. Dado que tenemos las predicciones de las series latentes $h = 12$ pasos adelante, esto es el vector de predicciones $S(t) = (S_1(t), S_2(t), \ldots, S_8(t))$, con $t = 1, 2, \ldots, 12$, y tenemos la matriz de mezcla $W^{-1}$, obtenida de (4.2), obtenemos las predicciones de las series de tiempo originales $X(t) = (X_1(t), X_2(t), \ldots, X_8(t))$, de la siguiente forma:

$$X(t) = W^{-1}S(t), \text{ con } t = 1, 2, \ldots, 12$$

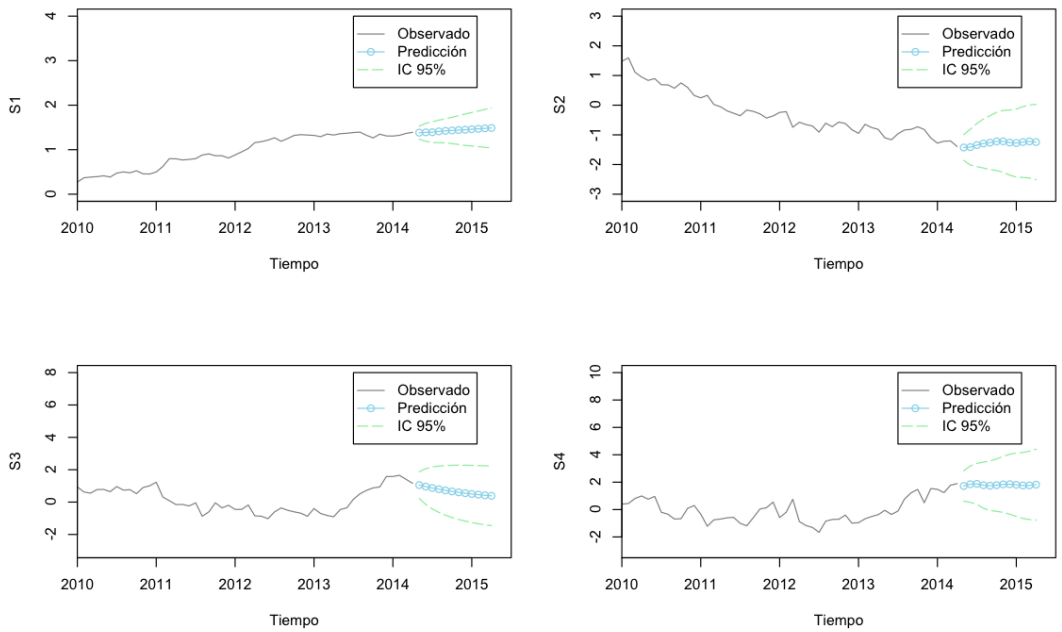


Figura 4–11: Predicciones e Intervalos de Predicción del 95 % de Confianza de las cuatro primeras Series Latentes de Índices de Precios al Consumidor, para el período mayo del 2014 a abril del 2015 en Puerto Rico, mediante SOBI $k = 100$

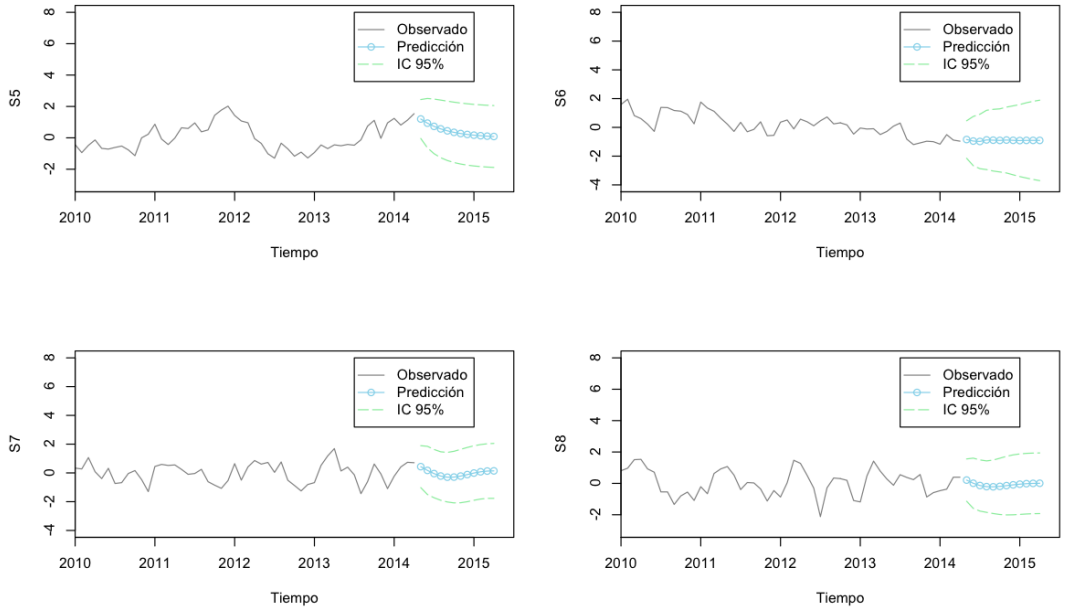


Figura 4–12: Predicciones e Intervalos de Predicción del 95 % de Confianza de las cuatro últimas Series Latentes de Índices de Precios al Consumidor, para el período mayo del 2014 a abril del 2015 en Puerto Rico, mediante SOBI $k = 100$

El Cuadro 4–6, muestra los valores numéricos de las predicciones del IPC para los primeros sectores: alimentos y bebidas, alojamiento, vestido, transportación, para el período mayo del 2014 a abril del 2015 en Puerto Rico. La Figura 4–13, muestra la representación gráfica de estas predicciones con sus respectivos intervalos de confianza del 95 %.

Cuadro 4–6: Predicciones de Índices de Precios al Consumidor según los cuatro primeros Sectores, para el período mayo del 2014 a abril del 2015, en Puerto Rico, mediante los Modelos SARIMA-SOBI

| Período | Alimento | Alojamiento | Vestido | Transporte |
|---|---|---|---|---|
| May 2014 | 123.13 | 113.94 | 93.93 | 116.97 |
| Jun 2014 | 123.13 | 113.34 | 94.08 | 116.12 |
| Jul 2014 | 123.16 | 113.07 | 94.23 | 115.62 |
| Aug 2014 | 123.27 | 113.05 | 94.29 | 115.61 |
| Sep 2014 | 123.33 | 113.08 | 94.33 | 115.64 |
| Oct 2014 | 123.45 | 113.14 | 94.51 | 115.47 |
| Nov 2014 | 123.50 | 113.07 | 94.74 | 115.16 |
| Dec 2014 | 123.48 | 113.03 | 94.82 | 115.04 |
| Jan 2015 | 123.52 | 113.21 | 94.81 | 115.15 |
| Feb 2015 | 123.64 | 113.48 | 94.85 | 115.24 |
| Mar 2015 | 123.77 | 113.62 | 94.93 | 115.21 |
| Apr 2015 | 123.83 | 113.53 | 95.02 | 115.04 |

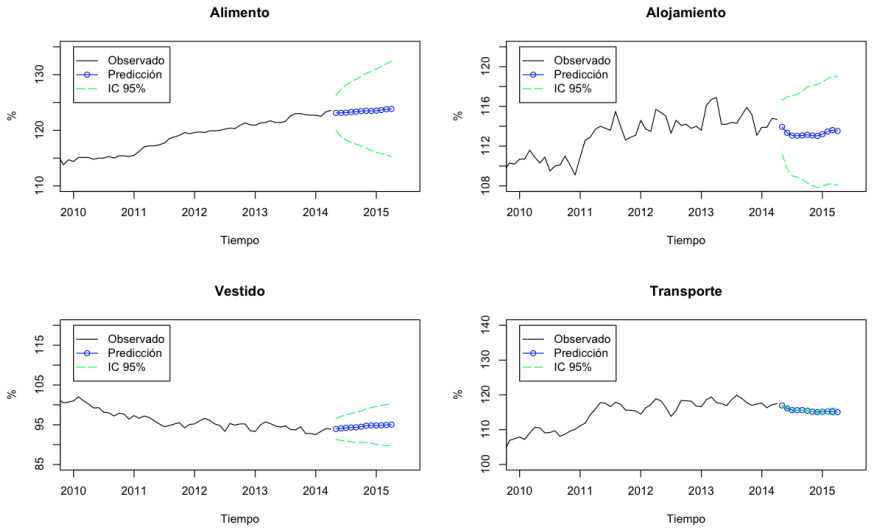


Figura 4–13: Predicciones e Intervalos de Predicción del 95 % de Confianza de las Series de Tiempo de los Índices de Precios al Consumidor de los Sectores: Alimento, Alojamiento, Vestido y Transporte; para el período mayo del 2014 a abril del 2015

El Cuadro 4–7, muestra los valores numéricos de las predicciones del IPC para los primeros sectores: cuidado médico, entretenimiento, educación y comunicación, otros artículos y servicios, para el período mayo del 2014 a abril del 2015 en Puerto Rico. La Figura 4–14, muestra la representación gráfica de estas predicciones con sus respectivos intervalos de confianza del 95 %.

Cuadro 4–7: Predicciones de Índices de Precios al Consumidor según los cuatro últimos Sectores, para el período mayo del 2014 a abril del 2015, en Puerto Rico, mediante los Modelos SARIMA-SOBI

| Período | Médico | Recreacióon | Educación | Otros |
|---|---|---|---|---|
| May 2014 | 137.43 | 109.23 | 109.07 | 112.43 |
| Jun 2014 | 137.08 | 109.30 | 108.84 | 112.26 |
| Jul 2014 | 136.60 | 109.42 | 108.69 | 112.19 |
| Aug 2014 | 136.38 | 109.63 | 108.69 | 112.19 |
| Sep 2014 | 136.29 | 109.74 | 108.65 | 112.19 |
| Oct 2014 | 136.25 | 109.84 | 108.60 | 112.21 |
| Nov 2014 | 136.31 | 109.90 | 108.55 | 112.11 |
| Dec 2014 | 136.51 | 109.90 | 108.52 | 111.99 |
| Jan 2015 | 136.72 | 109.95 | 108.54 | 111.99 |
| Feb 2015 | 136.78 | 110.05 | 108.57 | 112.06 |
| Mar 2015 | 136.87 | 110.12 | 108.59 | 112.11 |
| Apr 2015 | 136.99 | 110.14 | 108.57 | 112.04 |

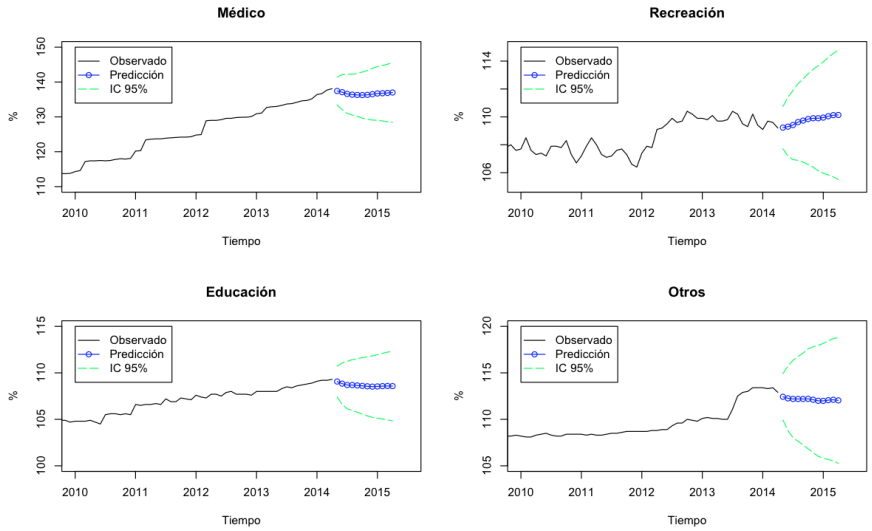


Figura 4–14: Predicciones e Intervalos de Predicción del 95 % de Confianza de las Series de Tiempo de los Índices de Precios al Consumidor de los Sectores: Médico, Recreación, Educación y Otros; para el período mayo del 2014 a abril del 2015

Hay que destacar que las predicciones para el grupo alimento del IPC, son casi constantes, unicamente incrementando en décimas el índice; el grupo alojamiento del IPC presenta una ligera baja y luego se estabiliza, el cambio es en décimas del índice; en vestido el IPC se incrementa de 93.9 % a 95 %; en transporte el IPC, decrece de 116.97 % a 115.04 %; en médico el IPC se mantiene constante; en recreación sube ligeramente; y en educación y otros se mantiene constante. Por otro lado el ancho de los intervalos de predicción es influenciado por los modelos SARIMA seleccionados y la matriz de mezcla; si seleccionamos un modelo AR, estos siguen la trayectoria de la serie, por tanto la variabilidad del modelo crece, así el intervalo de predicción se amplía; este comportamiento no sucede en los modelos MA donde la variabilidad del modelo se mantiene constante, por lo que el ancho del intervalo de predicción no se amplía.

Para la fase de validación del modelo SARIMA-SOBI, consideramos una muestra de entrenamiento desde el período de julio del 2004 a abril del 2013, que hacen un total de 106 observaciones, luego la muestra de prueba comprende desde mayo del 2013 a abril del 2014, que representan 12 observaciones.

El Cuadro 4–8 muestra la medida de validación MAPE para el modelo SARIMA-SOBI para la serie de tiempo multivariada de IPC en los ocho grupos. El modelo presenta un buen comportamiento en todos los grupos con porcentajes de error absoluto medio (MAPE) bastante pequeños, donde el más alto es en la serie de Alojamiento con un error porcentual promedio de 2.1 %.

Bajo el mismo procedimiento se realiza la validación del modelo SARIMA-AMUSE con $k = 1, 2$. El Cuadro 4–9, muestra la medida de validación MAPE para el modelo SARIMA-AMUSE con $k = 1, 2$, para la serie de tiempo multivariada de índices de precios al consumidor en los ocho grupos. El comportamiento del MAPE del modelo SARIMA-AMUSE con $k = 1$ y $k = 2$ para los ocho grupos, es bastante pequeño, llegando a alcanzar un 4.2 % de porcentaje de error absoluto medio como máximo.

Cuadro 4–8: Validación de los Modelos SARIMA-SOBI de Índices de Precios al Consumidor, para el período mayo del 2013 a abril del 2014

| Grupo | MAPE |
|---|---|
| Alimento | 0.0042 |
| Alojamiento | 0.0214 |
| Vestido | 0.0127 |
| Transporte | 0.017 |
| Médico | 0.0194 |
| Recreación | 0.0066 |
| Educación | 0.0033 |
| Otros | 0.0177 |

Cuadro 4–9: Validación de los Modelos SARIMA-AMUSE de Índices de Precios al Consumidor, para el período mayo del 2013 a abril del 2014

| Grupo | $k = 1$ MAPE | $k = 2$ MAPE |
|---|---|---|
| Alimento | 0.0160 | 0.0127 |
| Alojamiento | 0.0075 | 0.0126 |
| Vestido | 0.0171 | 0.0217 |
| Transporte | 0.0065 | 0.0129 |
| Médico | 0.0421 | 0.0221 |
| Recreación | 0.0037 | 0.0033 |
| Educación | 0.0079 | 0.0053 |
| Otros | 0.0276 | 0.0254 |

Por otro lado al contrastar los modelos SARIMA-SOBI y SARIMA-AMUSE, los errores porcentuales promedios son bastante semejantes, por lo que se consideran adecuados los dos modelos.

**El problema de Ordenar:** Siguiendo el criterio de ordenamiento de las series latentes propuesto por *Back and Weigen*, tenemos los siguientes resultados para las series latentes de índice de precios al consumidor.

El Cuadro 4–10, muestra el ordenamiento de las series latentes estimadas bajo el algoritmo SOBI. La serie latente $S_1$ explica un 84.7 % de la variabilidad de los datos. Las cuatro primeras series latentes acumulan un porcentaje de variabilidad de 95.16 %.

Cuadro 4–10: Series Latentes Ordenadas según el Porcentaje Variabilidad Explicada de los datos IPC, SOBI

| $S_1$ | $S_4$ | $S_2$ | $S_3$ | $S_6$ | $S_5$ | $S_8$ | $S_7$ |
|---|---|---|---|---|---|---|---|
| 84.70 | 4.16 | 3.93 | 2.37 | 2.36 | 1.19 | 0.91 | 0.38 |

La Figura 4–15, muestra las series latentes ordenadas según la variabilidad explicada de los datos IPC.

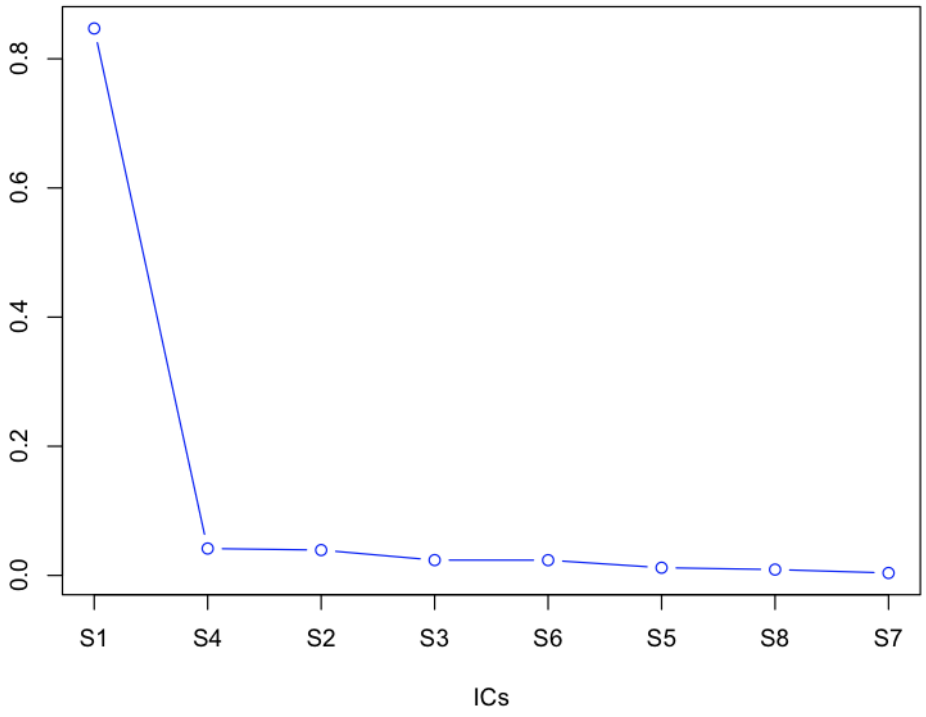


Figura 4–15: Series Latentes Ordenadas según su Variabilidad Explicada de los datos, SOBI

Bajo el algoritmo AMUSE $k = 1$, el Cuadro 4–11 muestra que la serie latente $S_1$ explica un 68.9 % de la variabilidad de los datos, $S_3$ explica 14.45 %, $S_4$ explica 10.97 % y $S_2$ explica 3.77 % de la variabilidad de los datos, para un total acumulado de 98.1 %.

Cuadro 4–11: Series Latentes Ordenadas según el Porcentaje Variabilidad Explicada de los datos IPC, AMUSE $k = 1$

| $S_1$ | $S_3$ | $S_4$ | $S_2$ | $S_5$ | $S_7$ | $S_8$ | $S_6$ |
|---|---|---|---|---|---|---|---|
| 68.91 | 14.45 | 10.97 | 3.77 | 0.65 | 0.44 | 0.42 | 0.39 |

La Figura 4–16, muestra las series latentes ordenadas según la variabilidad explicada de los datos.

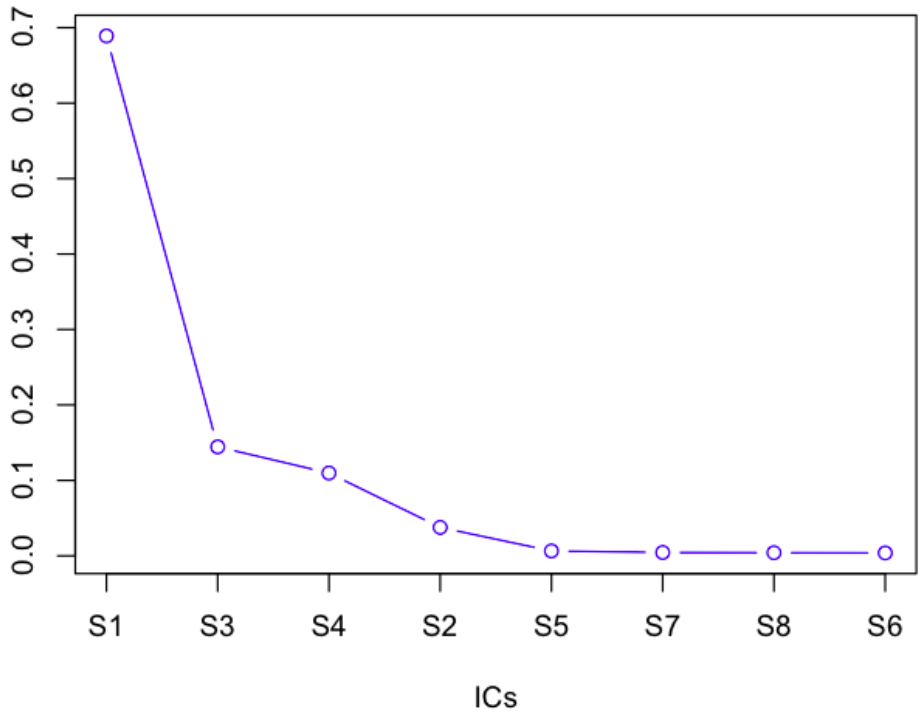


Figura 4–16: Series Latentes Ordenadas según su Variabilidad Explicada de los datos, AMUSE $k = 1$

Consideremos el modelo SARIMA-SOBI reducido con las cuatro series latentes más importantes $(S_1, S_4, S_2, S_3)$, que explican un 95.16 % de variabilidad de los

datos, es decir, recrear y realizar predicciones de las 8 series de tiempo IPC, con 4 series latentes.

La Figura 4–17, nos ilustra las 8 series de tiempo de índice de precios al consumidor, reconstruidas con las cuatro series latentes más importantes $S_1, S_4, S_2, S_3$. La recreación de las series IPC son muy parecidas a las verdaderas. Mediante el modelo SARIMA-SOBI reducido podemos realizar predicciónes.

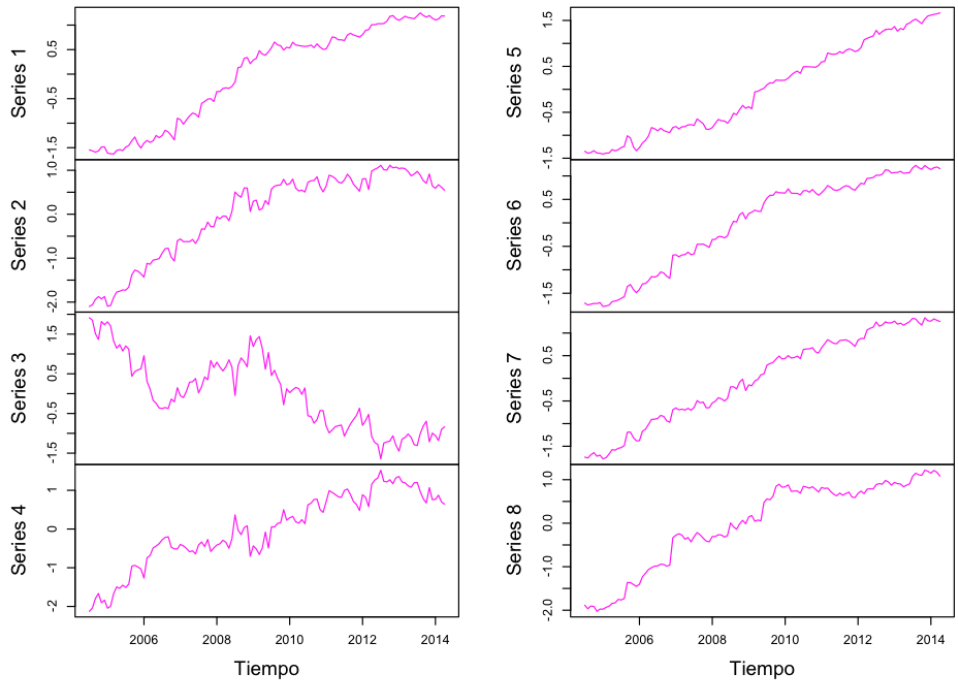


Figura 4–17: Series de Tiempo de IPC Reconstruidas, mediante cuatro Series Latentes $S_1, S_4, S_2, S_3$, con SOBI

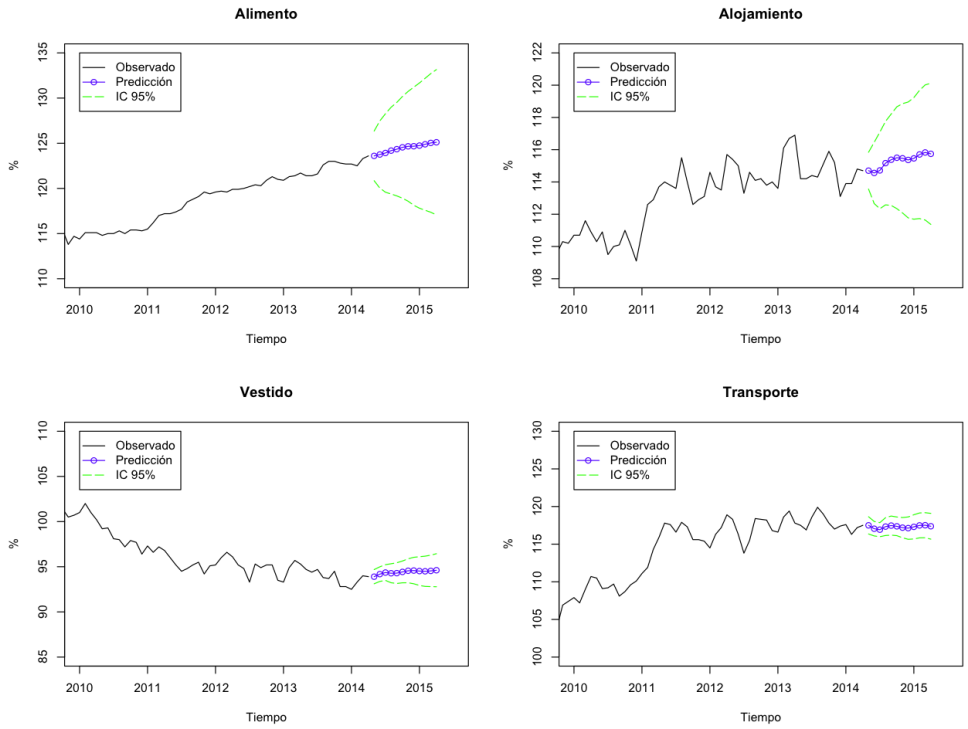


Figura 4–18: Predicciones e Intervalos de Predicción del 95 % de Confianza, de las cuatro primeras Series de Tiempo de índice de précios al consumidor de Puerto Rico para el periodo mayo 2014 a abril 2015, mediante el modelo SARIMA-SOBI reducido con cuatro Series Latentes $S_1, S_4, S_2, S_3$

Las Figuras 4–18 y 4–19, muestran las predicciones de las series de tiempo del índice de précios al consumidor de Puerto Rico para el periodo mayo 2014 a abril 2015, mediante el modelo SARIMA-SOBI reducido con cuatro Series Latentes $S_1, S_4, S_2, S_3$, hay que mencionar que dichas predicciones consideran únicamente las series latentes más importantes como generadoras.

El Cuadro 4–12 muestra los porcentajes de error absoluto medio, de los modelos SARIMA-SOBI reducido a cuatro series latentes $S_1, S_4, S_2, S_3$, de los índices de precios al consumidor en Puerto Rico, para el período mayo del 2013 a abril del 2014.

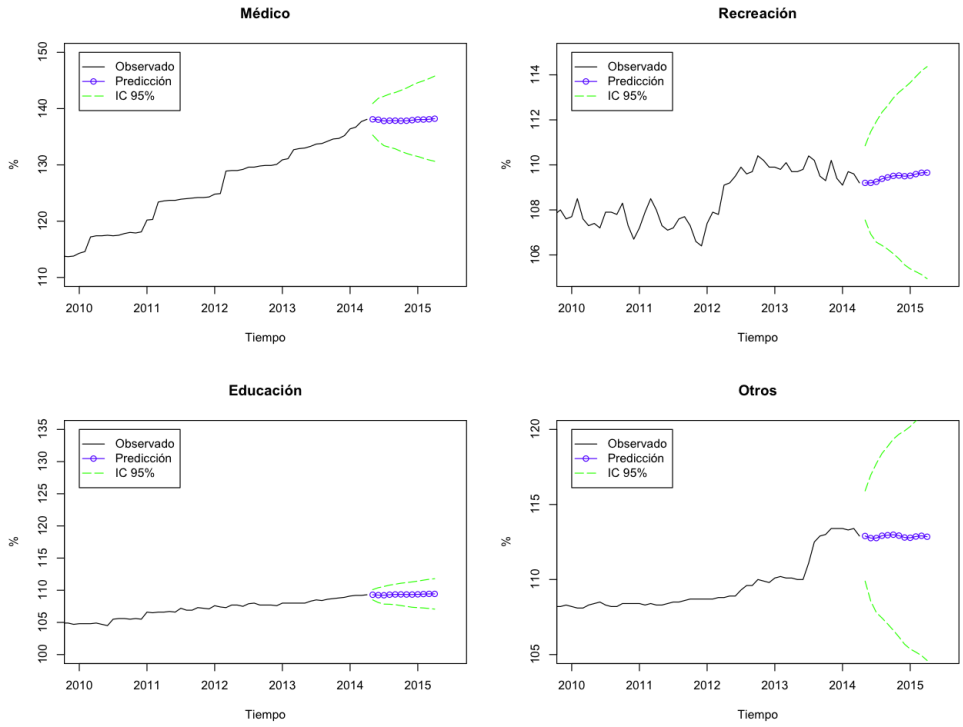


Figura 4–19: Predicciones e Intervalos de Predicción del 95 % de Confianza, de las cuatro últimas Series de Tiempo de índice de précios al consumidor de Puerto Rico para el periodo mayo 2014 a abril 2015, mediante el modelo SARIMA-SOBI reducido con cuatro Series Latentes $S_1, S_4, S_2, S_3$

Cuadro 4–12: Validación de los Modelos SARIMA-SOBI reducido con cuatro Series Latentes $S_1, S_4, S_2, S_3$, de Índices de Precios al Consumidor, para el período mayo del 2013 a abril del 2014

| Grupo | MAPE |
|---|---|
| Alimento | 0.0062 |
| Alojamiento | 0.0187 |
| Vestido | 0.0106 |
| Transporte | 0.0113 |
| Médico | 0.0197 |
| Recreación | 0.0073 |
| Educación | 0.0028 |
| Otros | 0.0168 |

# Capítulo 5
# CONCLUSIONES Y TRABAJOS FUTUROS

El análisis de componentes independientes adaptada a series de tiempo, permitió un análisis univariado a cada una de las series latentes, evitando la complejidad matemática de un análisis multivariado de series de tiempo. En esta investigación hemos explorado como se realiza el ICA, para la extracción de series latentes y predicción de las series de tiempo multivariadas no estacionarias de consumo de energía eléctrica e índice de precios al consumidor de Puerto Rico; utilizando los métodos SARIMA. Para la extracción de las series latentes, ICA asume que las series de tiempo multivariadas son linealmente generadas por un conjunto de series latentes, que son estadísticamente independientes; esta extracción se realizó mediante los algoritmos AMUSE y SOBI.

La series latentes independientes que se obtienen, se asocian a las distintas componentes de una serie de tiempo, esto es: Tendencia, Estacional y Aleatorio. Particularmente en la serie de tiempo multivariada de consumo de energía eléctrica de Puerto Rico, se consideró la extracción de tres series latentes. La serie latente $S_1$, se asocia con el componente de tendecia, $S_2$ con el componente estacional y $S_3$ con el componente aleatorio. Por otro lado en la serie de tiempo multivariada de índices de precios al consumidor de Puerto Rico, las series latentes $S_1$, $S_2$, $S_3$ y $S_1$ se asocian con el componente de tendencia creciente o decreciente, las series latentes $S_5$, $S_6$, $S_7$ y $S_8$ se asocian con el componente aleatorio. Los comportamientos de tendencia, estación y aleatorio estaban interrelacionados en las series de tiempo originales, esta metodología nos permite analizar cada uno de estos comportamientos por separado.

La capacidad predictiva de los modelos se cuantifica mediante el porcentaje de error absoluto medio (MAPE) en donde se requiere valores relativamente pequeños. En la serie de consumo de energía eléctrica, bajo el modelo SARIMA-SOBI se obtivieron 4.3 % (residencial), 6.6 % (comercial) y 14.6 % (industrial). Bajo el modelo SARIMA-AMUSE se obtuvieron 4.3 % (residencial), 5.2 % (comercial) y 8.2 % (industrial). En la serie de índice de precios al cosumidor, bajo el modelo SARIMA-SOBI se obtivieron 0.4 % (alimento), 2.1 % (alojamiento), 1.3 % (vestido), 1.7 % (transporte), 1.9 % (médico) y 0.7 % (recreación), 0.3 % (educación) y 1.8 % (otros). Bajo el modelo SARIMA-SOBI reducido a cuatro series latentes, se obtivieron 0.6 % (alimento), 1.9 % (alojamiento), 1.1 % (vestido), 1.1 % (transporte), 2.0 % (médico) y 0.7 % (recreación), 0.3 % (educación) y 1.7 % (otros). Finalmente, bajo el modelo SARIMA-AMUSE se obtivieron 1.6 % (alimento), 0.8 % (alojamiento), 1.7 % (vestido), 0.7 % (transporte), 4.2 % (médico), 0.4 % (recreación), 0.8 % (educación) y 2.8 % (otros). Claramente los errores son pequeños por lo que se garantiza una buena predicción de los modelos.

La ejecución de la metodología propuesta permite una reducción de la dimensionalidad, simplificando el modelamiento a un grupo reducido de series latentes. A pesar de la reducción se logra recrear e manera adecuada las series originales. Particularmente en la serie de tiempo multivariada de índice de precios al consumidor, bajo el algoritmo SOBI, se ha considerado la reducción de ocho a cuatro series latentes ( $S_1$, $S_2$, $S_3$ y $S_4$), que representan un 95.16 % de variabilidad de los datos.

Usando los modelos SARIMA-SOBI y SARIMA-AMUSE, se obtuvieron predicciones de la serie de tiempo multivariadas de consumo de energía eléctrica en los sectores residencial, comercial e industrial para el período abril del 2014 a marzo del 2015 y las series de tiempo multivariadas de índice de precios al consumidor en los ocho grupos, para el período de mayo 2014 a abril del 2015.

**Algunos trabajos futuros**:

El algoritmo AMUSE presenta el problema de elección del retraso "$k$" ya que una elección desafortunada puede evitar la estimación de las series latentes; por otro lado, el algoritmo SOBI plantea considerar más retrasos para evitar la limitación de AMUSE, sin embargo todavia queda el problema del número de retrasos a considerar.

En ICA no se puede determinar el orden de las series latentes, esto no es un problema al momento de trabajar con un número pequeño de series como se vio en este trabajo, pero cabe la posibilidad de que en un estudio específico se necesite saber el orden de las series latentes; a pesar de contar con algunos criterios de ordenamiento, no existe una teoría que sustente un ordenamiento general.

Con el proposito de realizar una comparación de las predicciones, se podrían usar diferentes metodos de predicción, dentro de la teoría de series de tiempo, estan los vectores autorregresivos de medias móviles (VARMA).

Bibliografía


[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *Signal Processing, IEEE Transactions on*, 45(2):434–444, 1997.

[2] G. E. Box, G. M. Jenkins, and G. C. Reinsel. *Time series analysis: forecasting and control*. John Wiley & Sons, 2013.

[3] P. J. Brockwell and R. A. Davis. *Introduction to time series and forecasting*, volume 1. Taylor & Francis, 2002.

[4] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

[5] A. Cichocki, S.-i. Amari, et al. *Adaptive blind signal and image processing*. John Wiley Chichester, 2002.

[6] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[7] J. de Planificación. Apéndice estadístico. *Informe Económico al Gobernador*, 1:86, 2013.

[8] A. García-Ferrer, E. González-Prieto, and D. Peña. Exploring ica for time series decomposition. *UC3M Working papers. Statistics and Econometrics 11-11*, 2011.

[9] M. P. González Casimiro. Análisis de series temporales: Modelos arima. *SARRIKO-ON; 04-09*, 2009.

[10] J. D. Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

[11] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[12] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

[13] D. Langlois, S. Chartier, and D. Gosselin. An introduction to independent component analysis: Infomax and fastica algorithms. *Tutorials in Quantitative Methods for Psychology*, 6(1):31–38, 2010.

[14] P. A. Morettin and C. Toloi. *Análise de séries temporais*. Blucher, 2006.

[15] G. R. Naik and D. K. Kumar. An overview of independent component analysis and its applications. *Informatica: An International Journal of Computing and Informatics*, 35(1):63–81, 2011.

[16] K. Nordhausen, J. Cardoso, J. Miettinen, H. Oja, E. Ollila, and S. Taskinen. Jade: Jade and other bss methods as well as some bss performance criteria. *R package version*, 1, 2012.

[17] D. Peña. *Análisis de datos multivariantes*, volume 24. McGraw-Hill Madrid, 2002.

[18] T. D. Popescu. Time series forecasting using independent component analysis. *Proceedings of World Academy of Science: Engineering & Technology*, 49, 2009.

[19] S. Romero Lafuente. *Reducción de artefactos en señales electroencefalográficas mediante nuevas técnicas de filtrado automático basadas en separación ciega de fuentes.* PhD thesis, Universitat Politècnica de Catalunya, 2010.

[20] J. V. Stone. Independent component analysis: an introduction. *Trends in cognitive sciences*, 6(2):59–64, 2002.

[21] J. V. Stone. *Independent component analysis*. Wiley Online Library, 2004.

[22] A. C. Tang, J.-Y. Liu, and M. T. Sutherland. Recovery of correlated neuronal sources from eeg: the good and bad ways of using sobi. *Neuroimage*, 28(2):507–519, 2005.

[23] L. Tong, V. Soon, Y. Huang, and R. Liu. Amuse: a new blind identification algorithm. In *Circuits and Systems, 1990., IEEE International Symposium on*, pages 1784–1787. IEEE, 1990.

[24] R. Yau. Macroeconomic forecasting with independent component analysis. In *Econometric Society 2004 Far Eastern Meetings*, volume 741. Econometric Society, 2004.

# APÉNDICES

# Apéndice A
# GRÁFICAS Y RESULTADOS ADICIONALES DE SARIMA-SOBI Y SARIMA-AMUSE

## A.1. Serie de Tiempo Consumo de Energía Eléctrica

### A.1.1. SARIMA-SOBI

Luego estimar las series latentes mediante SOBI, se realiza el ajuste SARIMA. Debido a que las series latentes presentan el componente estacional de período 12, se puede identificar que $s = 12$, es necesario realizar una diferencia estacional $D = 1$, por otro lado para remover la tendencia realizamos una diferencia regular $d = 1$, luego de extraer estas componentes, procedemos con la selección autorregresiva y media movil, regular y estacional.

Cuadro A–1: Identificación de los Modelos Candidatos SARIMA de las Series Latentes de Consumo de Energía Eléctrica

| Componentes | SARIMA$(p,d,q)(P,D,Q)_s$ | $\hat{\sigma}^2$ | log-likelihood | AIC |
|---|---|---|---|---|
| $S_1$ | SARIMA$(6,1,0)(0,1,1)_{12}$ | 0.137 | -76.48 | 168.96 |
| | SARIMA$(5,1,1)(0,1,1)_{12}$ | 0.1391 | -78.53 | 173.06 |
| | SARIMA$(6,1,1)(0,1,1)_{12}$ | 0.1332 | -74.29 | 166.57 |
| $S_2$ | SARIMA$(0,1,2)(0,1,1)_{12}$ | 0.2256 | -117.82 | 243.64 |
| | SARIMA$(1,1,1)(0,1,1)_{12}$ | 0.2263 | -118.59 | 245.19 |
| | SARIMA$(2,1,1)(0,1,1)_{12}$ | 0.2254 | -117.88 | 245.77 |
| $S_3$ | SARIMA$(6,1,0)(0,1,1)_{12}$ | 0.4265 | -173.98 | 363.96 |
| | SARIMA$(5,1,1)(0,1,1)_{12}$ | 0.4304 | -173.36 | 362.11 |
| | SARIMA$(5,1,0)(0,1,1)_{12}$ | 0.4234 | -174.12 | 363.23 |

El Cuadro A–1, muestra la fase de identificación de posibles modelos para las series latentes, se muestran por cada serie tres posibles modelos SARIMA identificables bajo los criterios de $\hat{\sigma}^2$, *log-likelihood* y AIC. Luego de la fase de indentificación de

modelos SARIMA, estimamos los parámetros del modelos seleccionados, mediante el método de máxima verosimilitud condicional, a continuación se muestran los estimados y sus correspondientes errores estándar de estimación.

1. $S_1$: Modelo SARIMA$(6,1,1)(0,1,1)_{12}$

```
Coefficients:
          ar1      ar2      ar3      ar4      ar5      ar6      ma1     sma1
      -0.6595  -0.5932  -0.4588  -0.3389  -0.3381  -0.2789  -0.4158  -0.7821
s.e.   0.1582   0.1644   0.1617   0.1378   0.1039   0.0854   0.1561   0.0724
```

Luego el modelo estadístico es de la forma:

$$(1+0.66L+0.59L^2+0.46L^3+0.34L^4+0.34L^5+0.28L^6)\Delta\Delta_{12}X_t = (1+0.41L)(1+0.78L^{12})\varepsilon_t \quad \text{(A.1)}$$

2. $S_2$: Modelo SARIMA$(0,1,2)(0,1,1)_{12}$

```
Coefficients:
          ma1     ma2     sma1
      -0.8864  0.2706  -0.8242
s.e.   0.0751  0.0871   0.0780
```

Luego el modelo estadístico es de la forma:

$$\Delta\Delta_{12}X_t = (1+0.8864L-0.2706L^2)(1+0.8242L^{12})\varepsilon_t \quad \text{(A.2)}$$

3. $S_3$: Modelo SARIMA$(5,1,1)(0,1,1)_{12}$

```
Coefficients:
          ar1      ar2      ar3      ar4      ar5      ma1     sma1
      -0.6030  -0.4610  -0.3152  -0.2608  -0.3141  -0.4220  -0.8864
s.e.   0.2131   0.2144   0.1855   0.1324   0.0917   0.2254   0.0883
```

Luego el modelo estadístico es de la forma:

$$(1+0.60L+0.46L^2+0.31L^3+0.26L^4+0.31L^5)\Delta\Delta_{12}X_t = (1+0.42L)(1+0.89L^{12})\varepsilon_t \quad \text{(A.3)}$$

Luego de estimar parametros, procedemos al diagnóstico de los modelos, mediante el análisis de los residuales usando los errores estandarizados, la función de autocorrelación estimada y los P-valores de la prueba de *Ljung-Box*.

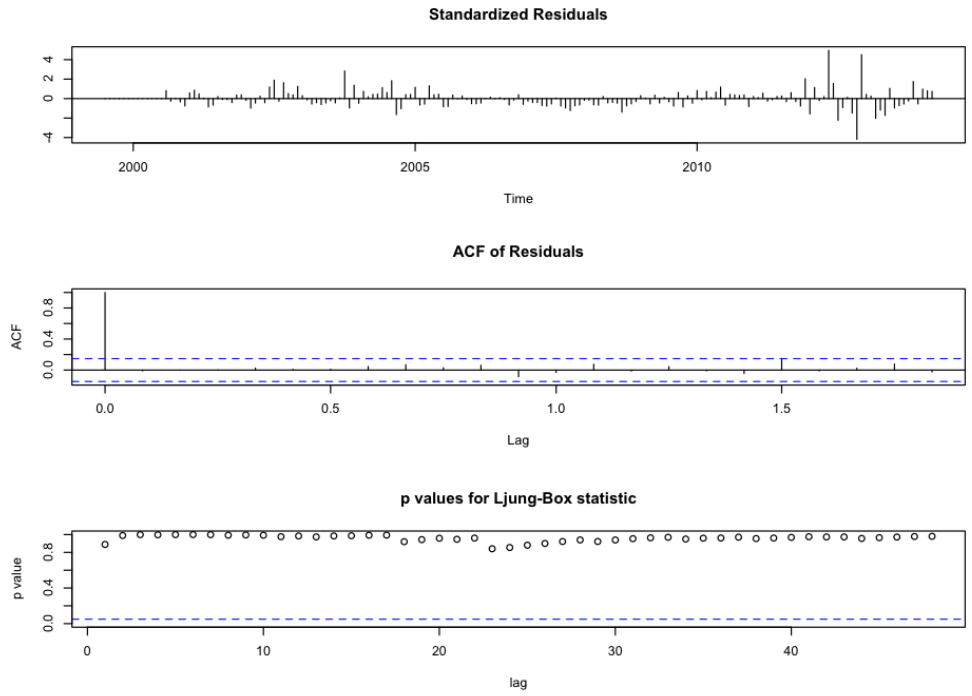


Figura A–1: Diagnóstico para la Serie Latente de Consumo de Energía Eléctrica $S_1$ modelo SARIMA$(6,1,1)(0,1,1)_{12}$. Residuales estandarizados, ACF de los residuales y p-valores para la prueba de *Ljung-Box*.

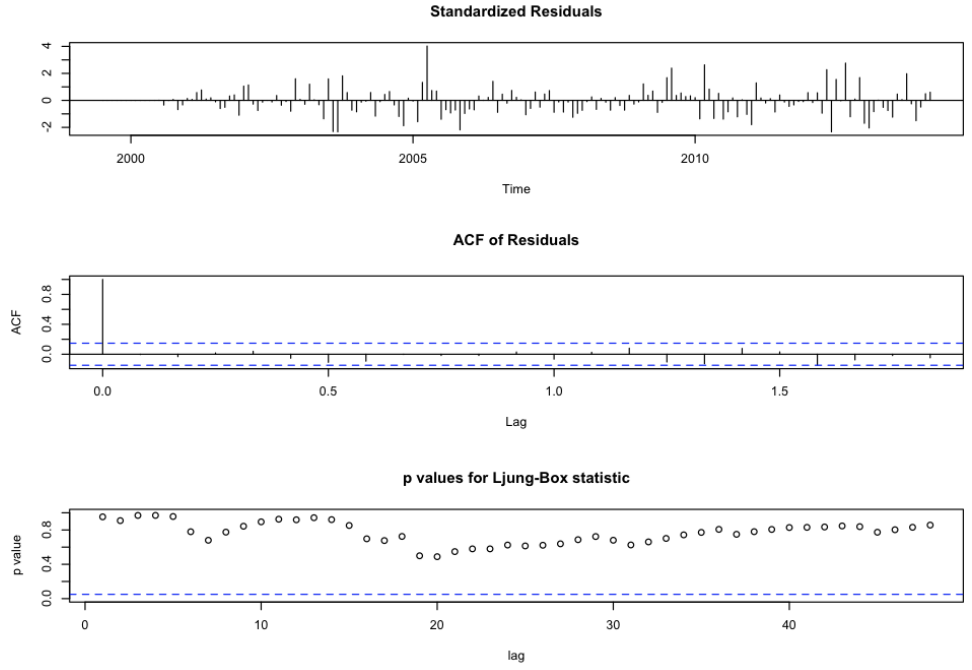


Figura A–2: Diagnóstico para la Serie Latente de Consumo de Energía Eléctrica $S_2$ modelo SARIMA$(0,1,2)(0,1,1)_{12}$. Residuales estandarizados, ACF de los residuales y p-valores para la prueba de *Ljung-Box*.

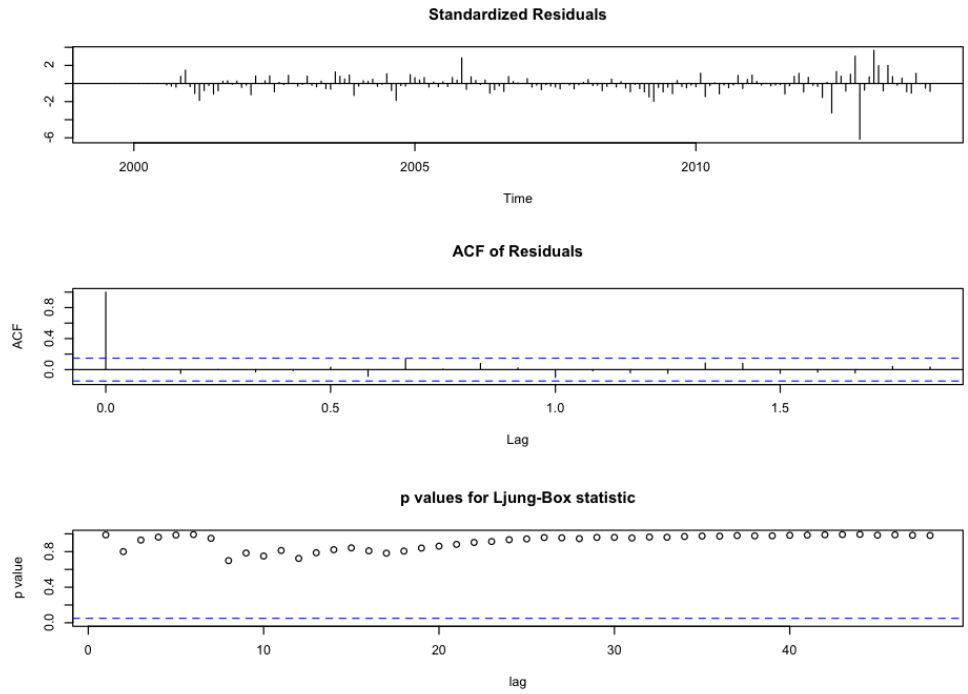


Figura A–3: Diagnóstico para la Serie Latente de Consumo de Energía Eléctrica $S_3$ modelo SARIMA$(5,1,1)(0,1,1)_{12}$. Residuales estandarizados, ACF de los residuales y p-valores para la prueba de *Ljung-Box*.

Las Figuras A–1, A–2 y A–3, muestran los p-valores para la prueba estadística de *Ljung-Box*, de los errores de cada modelo planteado de las series latentes. Se concluye que los errores de los modelos seleccionados SARIMA, se comportan como un ruido blanco, con una confianza del 95 %.

### A.1.2. SARIMA-AMUSE

Estimación de las series latentes se muestra en la Figura A–4 y matriz de separación $W$, con el retraso $k = 1$ se muestra en A.4.

$$W = \begin{pmatrix} 0.03687636 & -0.2807450 & 0.9567660 \\ 0.81635125 & 0.3481179 & -0.1816169 \\ -1.15890260 & 1.1830128 & 0.6604486 \end{pmatrix} \tag{A.4}$$

La Figura A–5, muestra las funciones de autocorrelación y correlaciones cruzadas de las series latentes usando el AMUSE con $k = 1$.

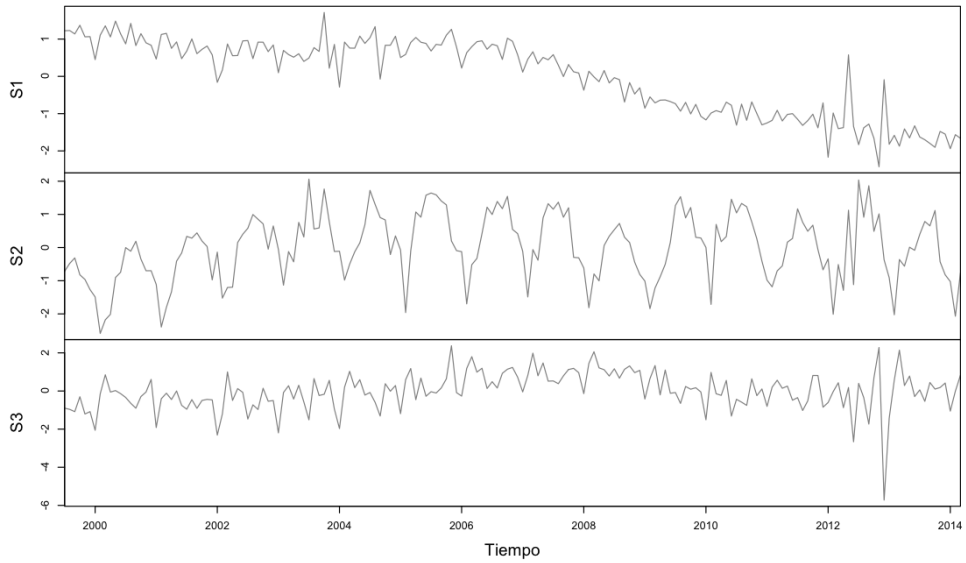


Figura A–4: Series Latentes de Consumo de Energía Eléctrica, estimadas mediante AMUSE con $k = 1$

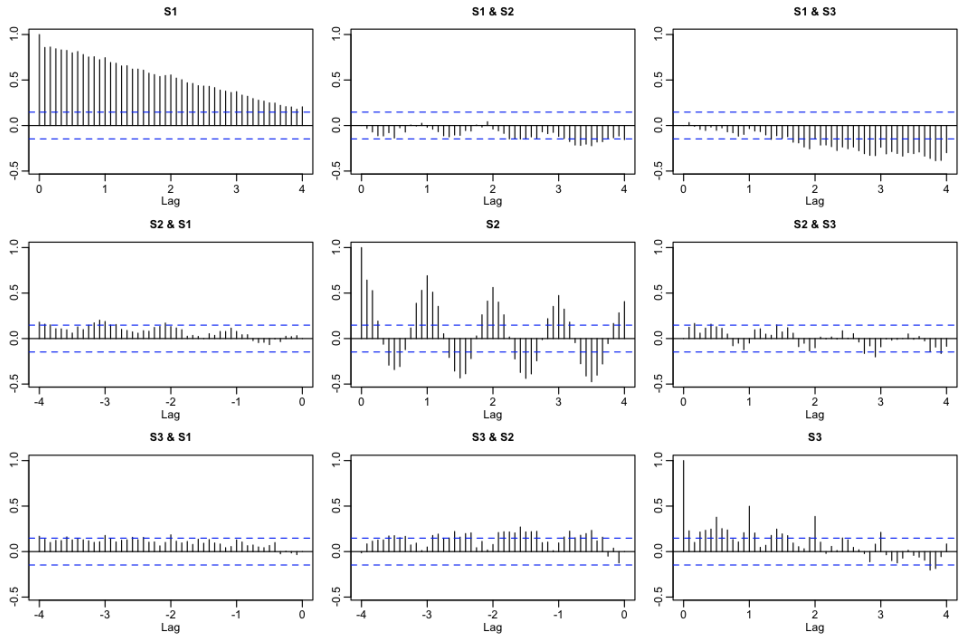


Figura A–5: Autocorrelaciones y Correlaciones Cruzadas de las Series Latentes de Consumo de Energía Eléctrica,Estimadas mediante AMUSE con $k = 1$

El ajuste de los modelos SARIMA para cada una de las series latentes, bajo AMUSE, es consideran similares, es decir que la identificación de parámetros son los mismos, el cambio se realiza en la estimación, claramente el diagnóstico cumple las suposiciones del modelo SARIMA planteado. La Figura A–6, muestra las predicciones para las series latentes con AMUSE $k = 1$.

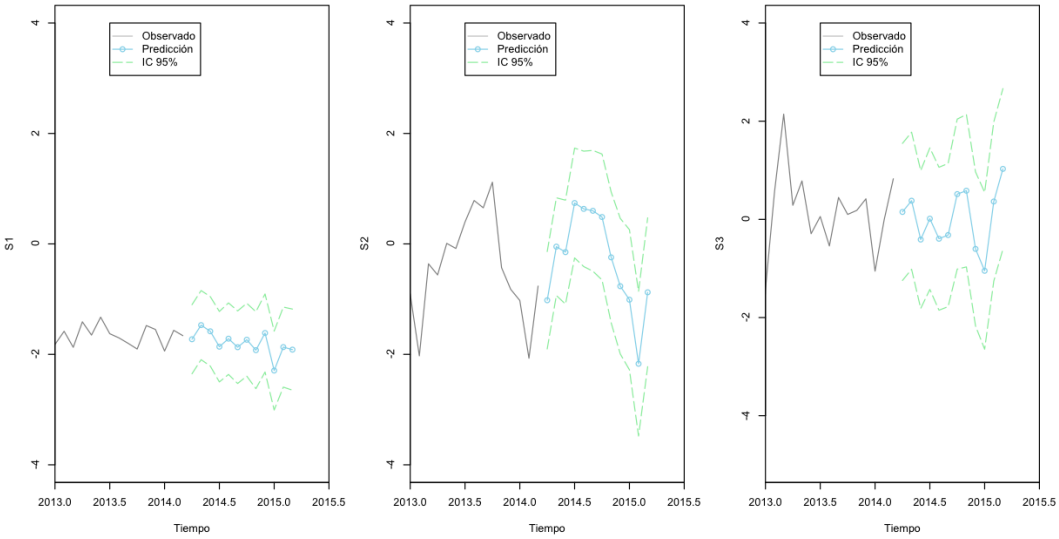


Figura A–6: Predicciones e Intervalos de Predicción del 95 % de Confianza de las Series Latentes de Consumo de Energía Eléctrica, para el período abril del 2014 a marzo del 2015 en Puerto Rico, con AMUSE $k = 1$

Las Figuras A–7, A–8 y A–9, muestran las predicciones para el período de abril del 2014 a marzo del 2015 de las series de tiempo de consumo de energía eléctrica para los sectores residencial, comercial e industrial respectivamente, con sus correspondientes intervalos de confianza del 95 %, los cuales presentan buena cobertura de las predicciones bajo el modelo SARIMA-AMUSE con $k = 1$.

El cuadro A–2, muestra los valores numéricos de las predicciones realizadas por le modelo SARIMA-AMUSE $k = 1$. de Consumo de Energía Eléctrica, para el período abril del 2014 a marzo del 2015, en Puerto Rico

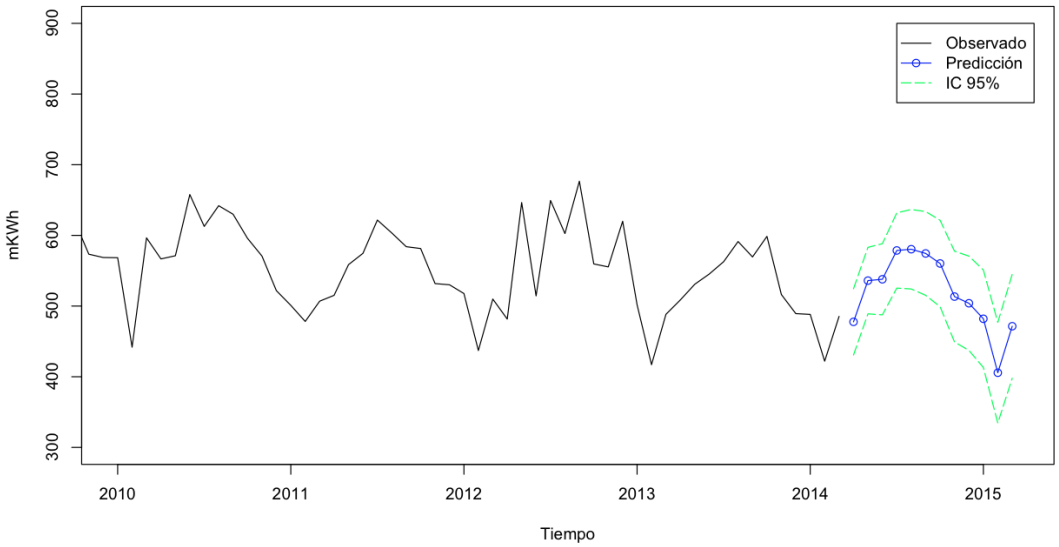


Figura A–7: Predicciones e Intervalos de Predicción del 95 % de Confianza de la Serie de Tiempo Consumo de Energía Eléctrica en el sector Residencial, para el período abril del 2014 a marzo del 2015 en Puerto Rico, con AMUSE $k = 1$

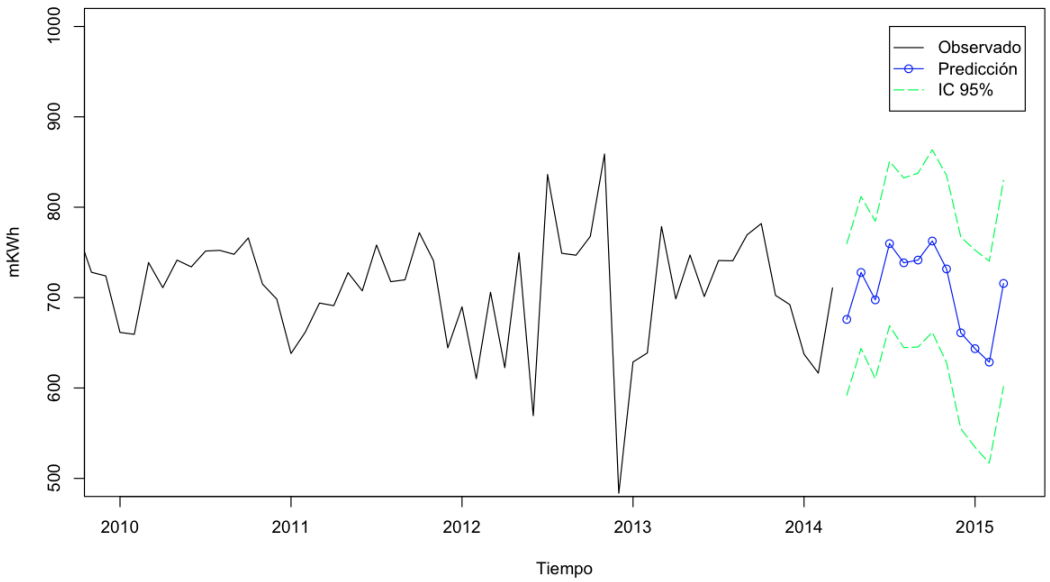


Figura A–8: Predicciones e Intervalos de Predicción del 95 % de Confianza de la Serie de Tiempo Consumo de Energía Eléctrica en el sector Comercial, para el período abril del 2014 a marzo del 2015 en Puerto Rico, con AMUSE $k = 1$

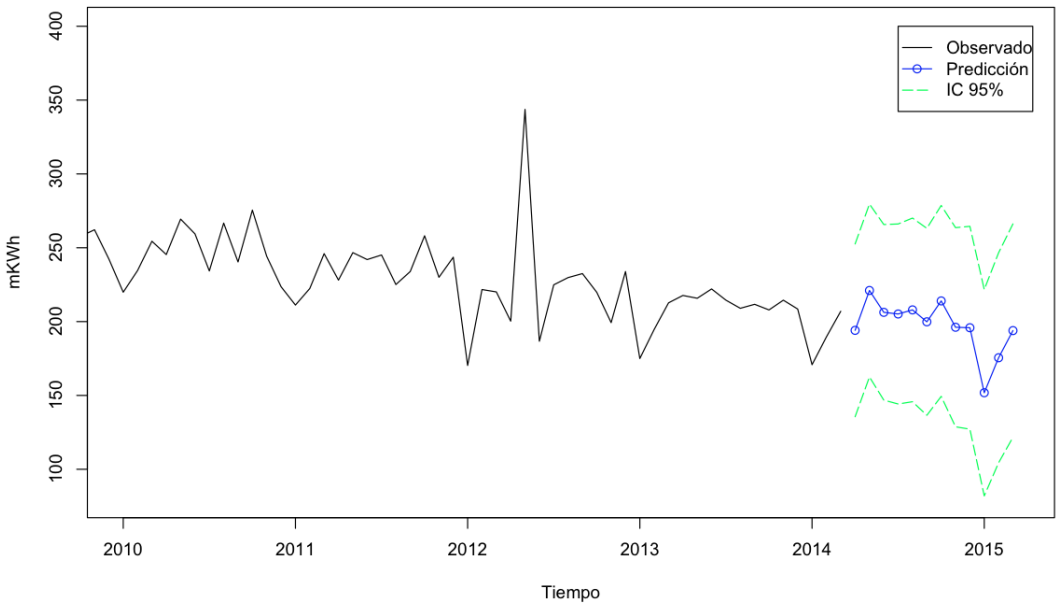


Figura A–9: Predicciones e Intervalos de Predicción del 95 % de Confianza de la Serie de Tiempo Consumo de Energía Eléctrica en el sector Industrial, para el período abril del 2014 a marzo del 2015 en Puerto Rico, con AMUSE $k = 1$

Cuadro A–2: Predicciones de Consumo de Energía Eléctrica, para el período abril del 2014 a marzo del 2015, en Puerto Rico, mediante el Modelo SARIMA-AMUSE $k = 1$

| período | Residencial | Comercial | Industrial |
|---|---|---|---|
| Apr 2014 | 477.66 | 675.89 | 194.07 |
| May 2014 | 536.02 | 727.83 | 221.07 |
| Jun 2014 | 537.90 | 697.51 | 206.27 |
| Jul 2014 | 578.66 | 759.75 | 205.15 |
| Aug 2014 | 580.42 | 738.51 | 207.92 |
| Sep 2014 | 574.45 | 741.55 | 199.85 |
| Oct 2014 | 560.25 | 762.62 | 213.99 |
| Nov 2014 | 513.40 | 731.73 | 196.17 |
| Dec 2014 | 503.94 | 661.06 | 195.84 |
| Jan 2015 | 482.04 | 643.50 | 151.82 |
| Feb 2015 | 405.67 | 628.63 | 175.52 |
| Mar 2015 | 471.47 | 715.68 | 193.91 |

## A.2. Serie de Tiempo de Índice de Precios al Consumidor

### A.2.1. SARIMA-SOBI

Las Figuras A–10 y A–11, presenta las funciones de autocorrelación y correlación cruzada de las series del IPC en los 8 grupos, estas muestran una clara estructura de dependencia temporal entre las series y dentro de cada serie, por lo que se hace necesario considerar el SOBI para separar las fuentes latentes del problema. Las Figuras A–12 y A–13, presenta las funciones de autocorrelación y correlación cruzada estimadas de las series latentes del IPC en los 8 grupos.

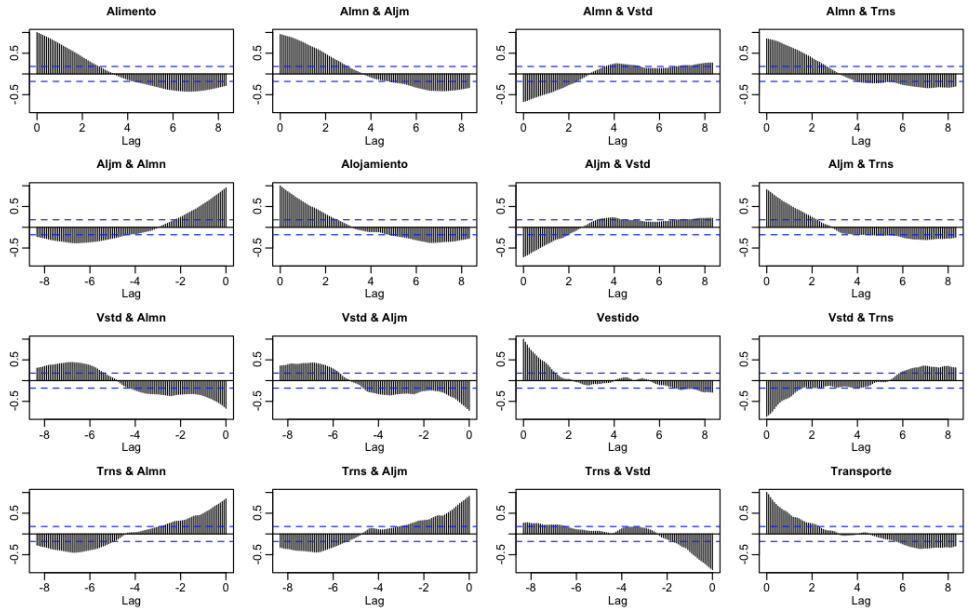


Figura A–10: Autocorrelaciones y Correlaciones Cruzadas de las Cuatro Primeras Series de Tiempo de Índice de Precios al Consumidor

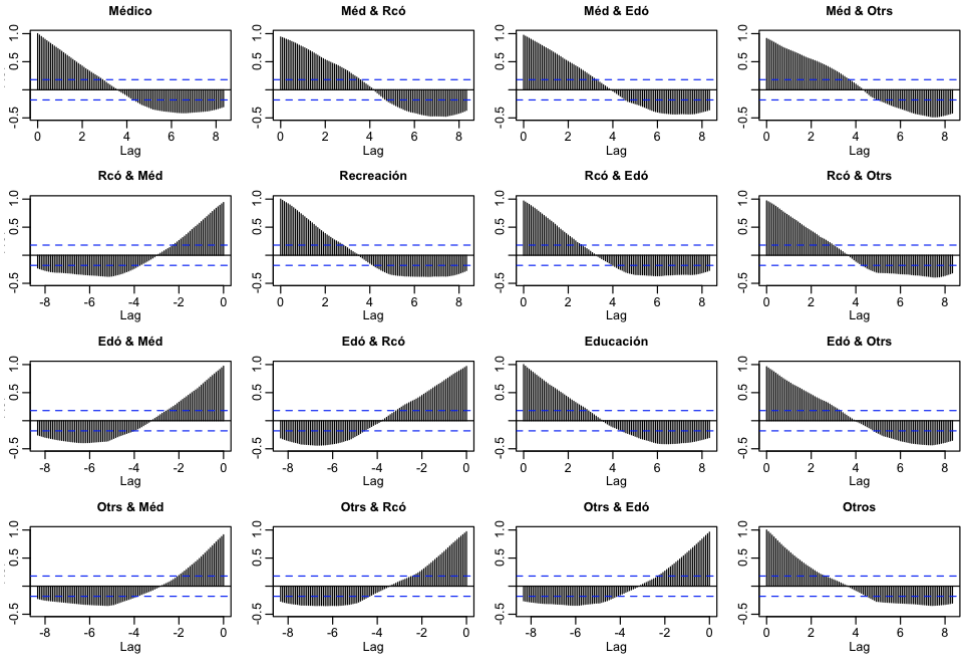


Figura A–11: Autocorrelaciones y Correlaciones Cruzadas de las Cuatro Últimas Series de Tiempo de Índice de Precios al Consumidor

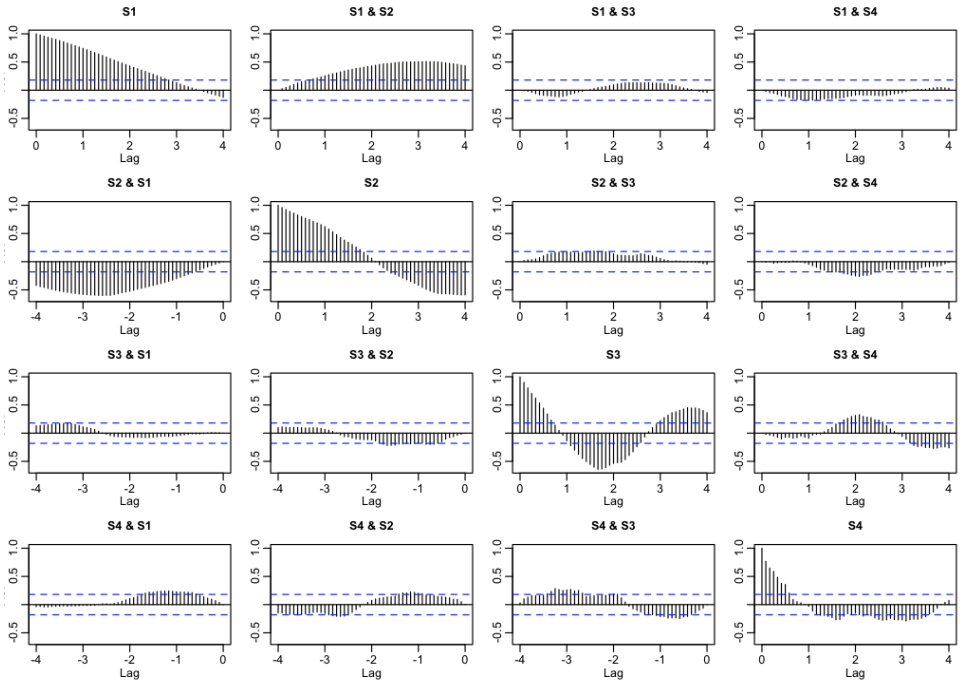


Figura A–12: Autocorrelaciones y Correlaciones Cruzadas de las Cuatro Primeras Series Latentes de Índice de Precios al Consumidor, Estimadas mediante SOBI $k = 100$

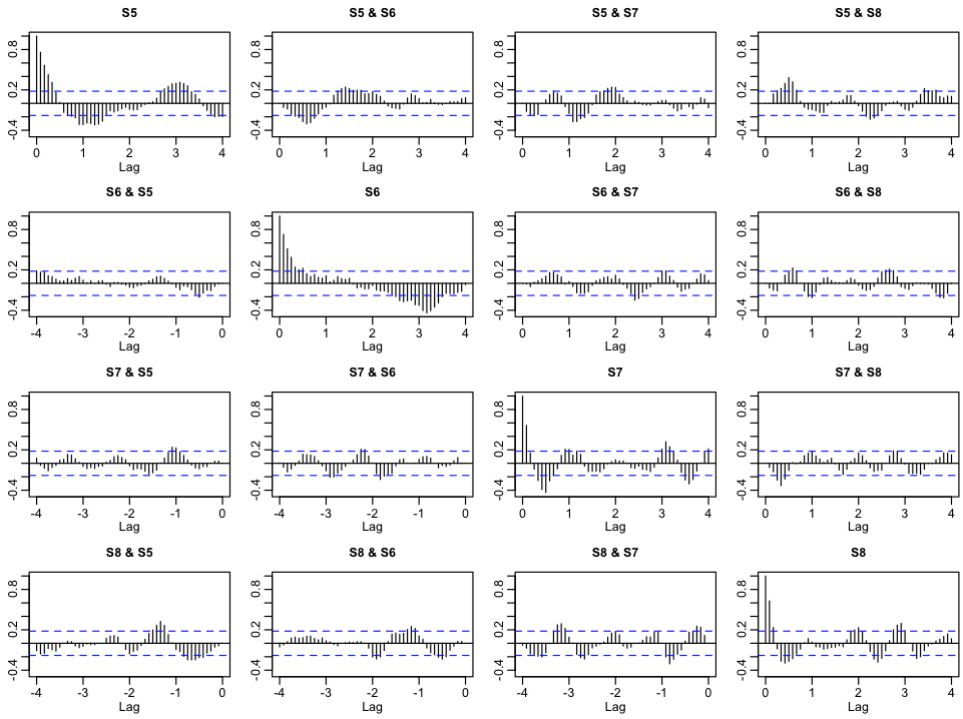


Figura A–13: Autocorrelaciones y Correlaciones Cruzadas de las Cuatro Últimas Series Latentes de Índices de Precios al Consumidor, Estimadas mediante SOBI $k = 100$

La estimación y diagnóstico de los modelos SARIMA, identificados en el Cuadro 4–5, se muestran a continuación:

1. $S_1$: SARIMA$(5,1,2)(0,0,0)_{12}$

```
Coefficients:
         ar1     ar2      ar3     ar4     ar5      ma1      ma2
      0.2026  0.5140  -0.1022  0.0580  0.3006  -0.2582  -0.6281
s.e.  0.1883  0.1805   0.1010  0.0903  0.0893   0.1838   0.1778
```

Luego el modelo estadístico es de la forma:

$$(1-0.2L-0.51L^2+0.1L^3-0.06L^4-0.3L^5)\Delta X_t = (1+0.26L+0.63L^2)\varepsilon_t \quad \text{(A.5)}$$

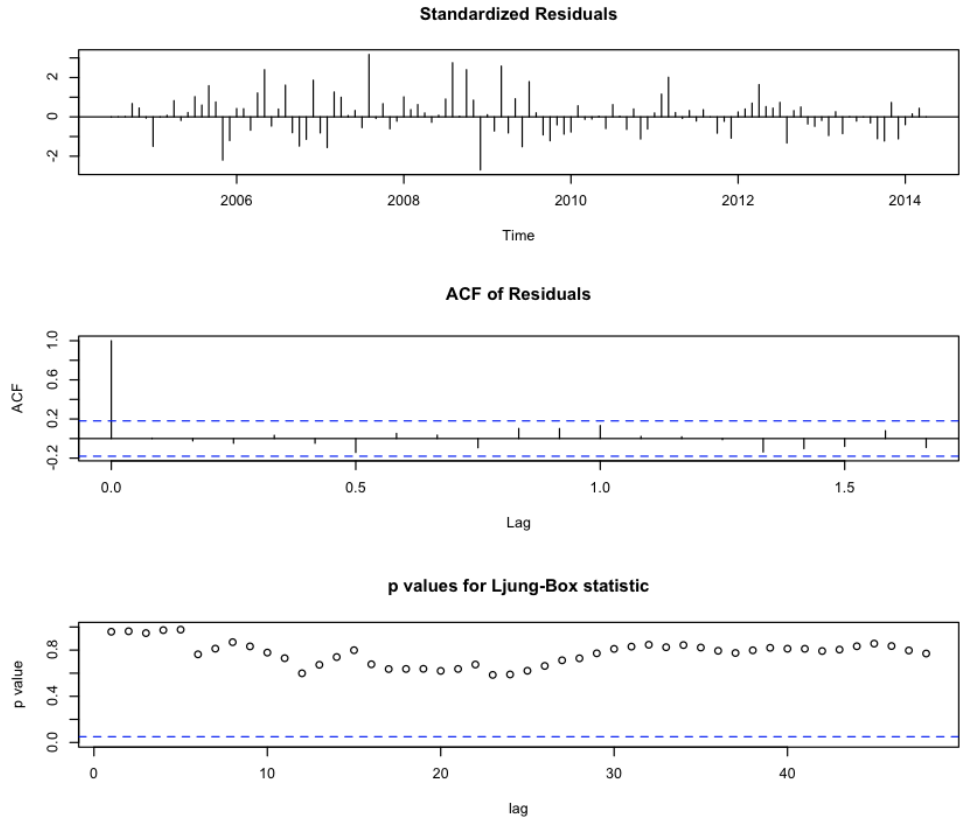


Figura A–14: Diagnóstico para la Series Latente $S_1$ de Índices de Precios al Consumidor, modelo SARIMA$(5,1,2)(0,0,0)_{12}$. Residuales estandarizados, ACF de los residuales y p-valores para la prueba de *Ljung-Box*.

2. $S_2$: SARIMA$(1,0,0)(1,0,0)_{12}$

```
Coefficients:
          ar1     sar1  intercept
       0.9676  0.2198    -0.4727
s.e.   0.0210  0.0895     0.6685
```

Luego el modelo estadístico es de la forma:

$$(1 - 0.9676L)(1 - 0.2198L^{12})\Delta X_t = -0.4727 + \varepsilon_t \tag{A.6}$$

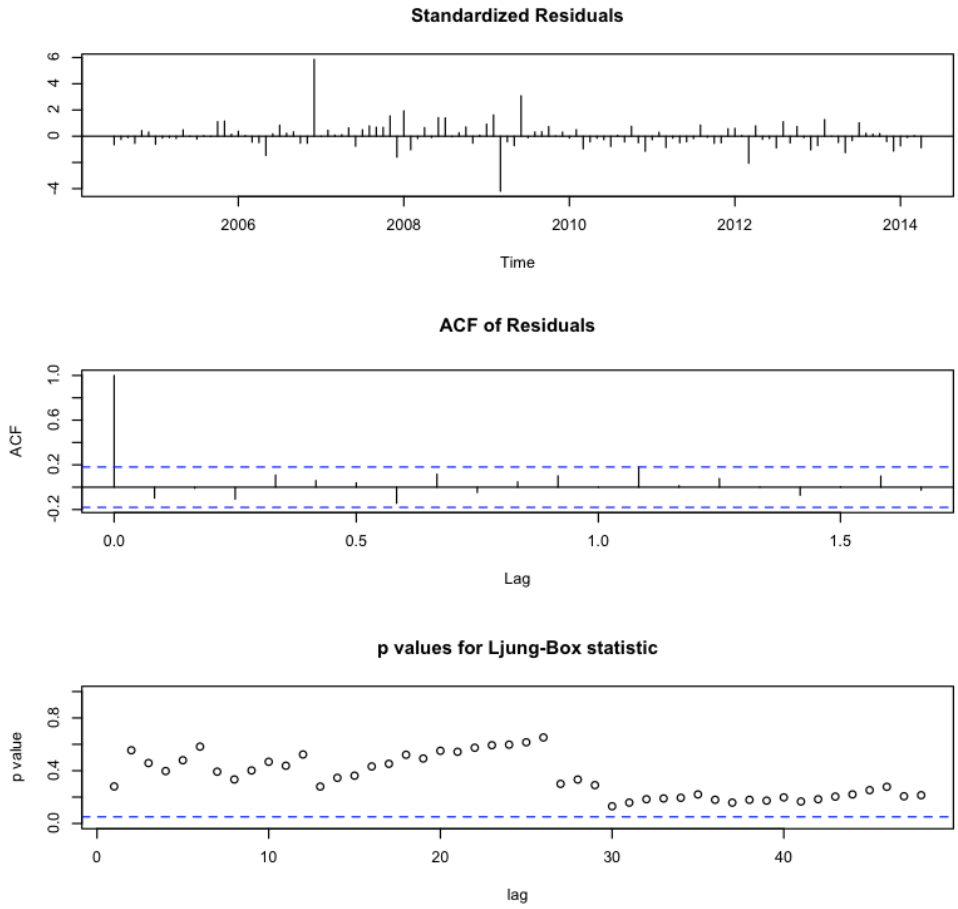


Figura A–15: Diagnóstico para la Series Latente $S_2$ de Índices de Precios al Consumidor, modelo SARIMA$(1,0,0)(1,0,0)_{12}$. Residuales estandarizados, ACF de los residuales y p-valores para la prueba de *Ljung-Box*.

3. $S_3$: SARIMA$(1,0,0)(0,0,0)_{12}$

```
Coefficients:
         ar1  intercept
      0.9057     0.0489
s.e.  0.0374     0.3795
```

Luego el modelo estadístico es de la forma:

$$(1 - 0.9057L)X_t = 0.0489 + \varepsilon_t \tag{A.7}$$

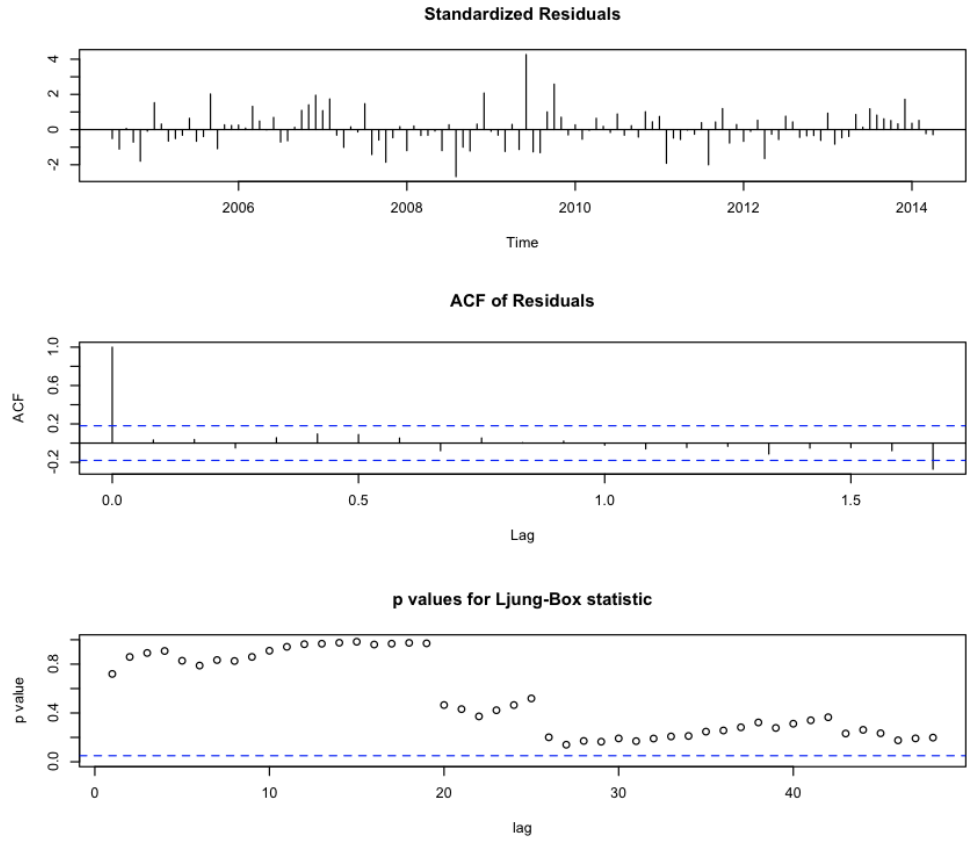


Figura A–16: Diagnóstico para la Series Latente $S_3$ de Índices de Precios al Consumidor, modelo SARIMA$(1,0,0)(0,0,0)_{12}$. Residuales estandarizados, ACF de los residuales y p-valores para la prueba de *Ljung-Box*.

4. $S_4$: SARIMA$(4,1,2)(0,0,0)_{12}$

```
Coefficients:
         ar1      ar2      ar3      ar4      ma1     ma2
      0.2985  -0.8884  -0.1997  -0.2615  -0.6437  0.9508
s.e.  0.1101   0.0963   0.0943   0.0932   0.0806  0.0566
```

Luego el modelo estadístico es de la forma:

$$(1 - 0.2985L + 0.8884L^2 + 0.1997L^3 + 0.2615L^4)\Delta X_t = (1 + 0.6437L - 0.9508L^2)\varepsilon_t \qquad \text{(A.8)}$$

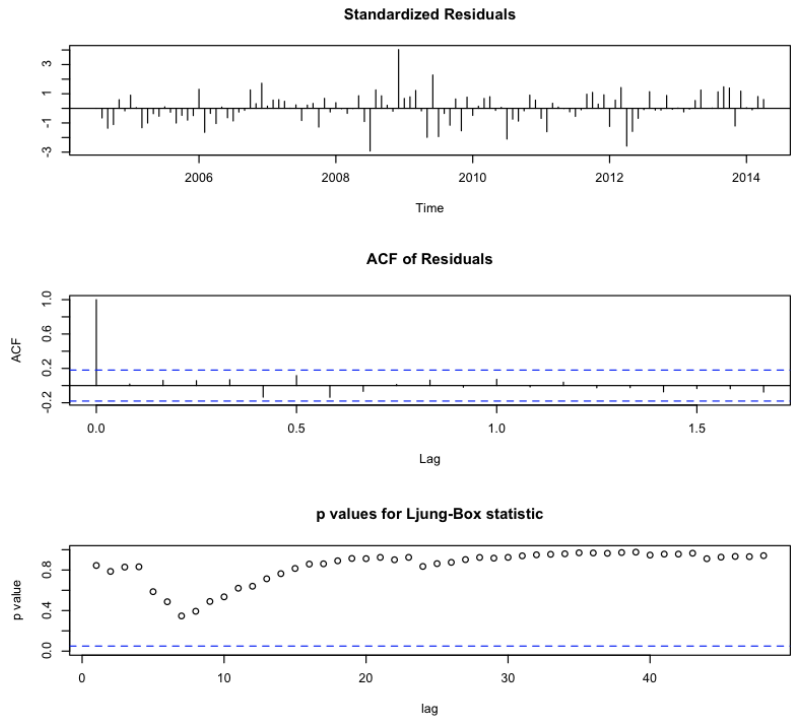


Figura A–17: Diagnóstico para la Series Latente $S_4$ de Índices de Precios al Consumidor, modelo SARIMA$(4,1,2)(0,0,0)_{12}$. Residuales estandarizados, ACF de los residuales y p-valores para la prueba de *Ljung-Box*.

5. $S_5$: SARIMA$(1,0,0)(0,0,0)_{12}$

```
Coefficients:
         ar1  intercept
      0.7778     0.0106
s.e.  0.0584     0.2547
```

Luego el modelo estadístico es de la forma:

$$(1 - 0.7778L)X_t = 0.0106 + \varepsilon_t \tag{A.9}$$

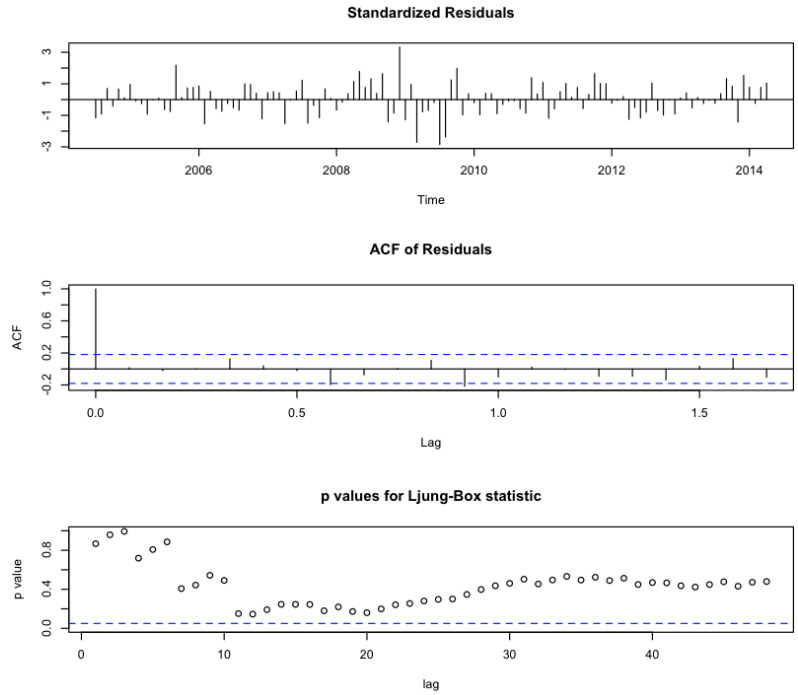


Figura A–18: Diagnóstico para la Series Latente $S_5$ de Índices de Precios al Consumidor, modelo SARIMA$(1,0,0)(0,0,0)_{12}$. Residuales estandarizados, ACF de los residuales y p-valores para la prueba de *Ljung-Box*.

6. $S_6$: SARIMA$(5,1,0)(0,0,0)_{12}$

```
Coefficients:
          ar1      ar2      ar3      ar4      ar5
      -0.1485  -0.2543  -0.0627  -0.2438  -0.1675
s.e.   0.0914   0.0894   0.0923   0.0894   0.0914
```

Luego el modelo estadístico es de la forma:

$$(1 + 0.1485L + 0.2543L^2 + 0.0627L^3 + 0.2438L^4 + 0.1675L^5)\Delta X_t = \varepsilon_t \quad \text{(A.10)}$$

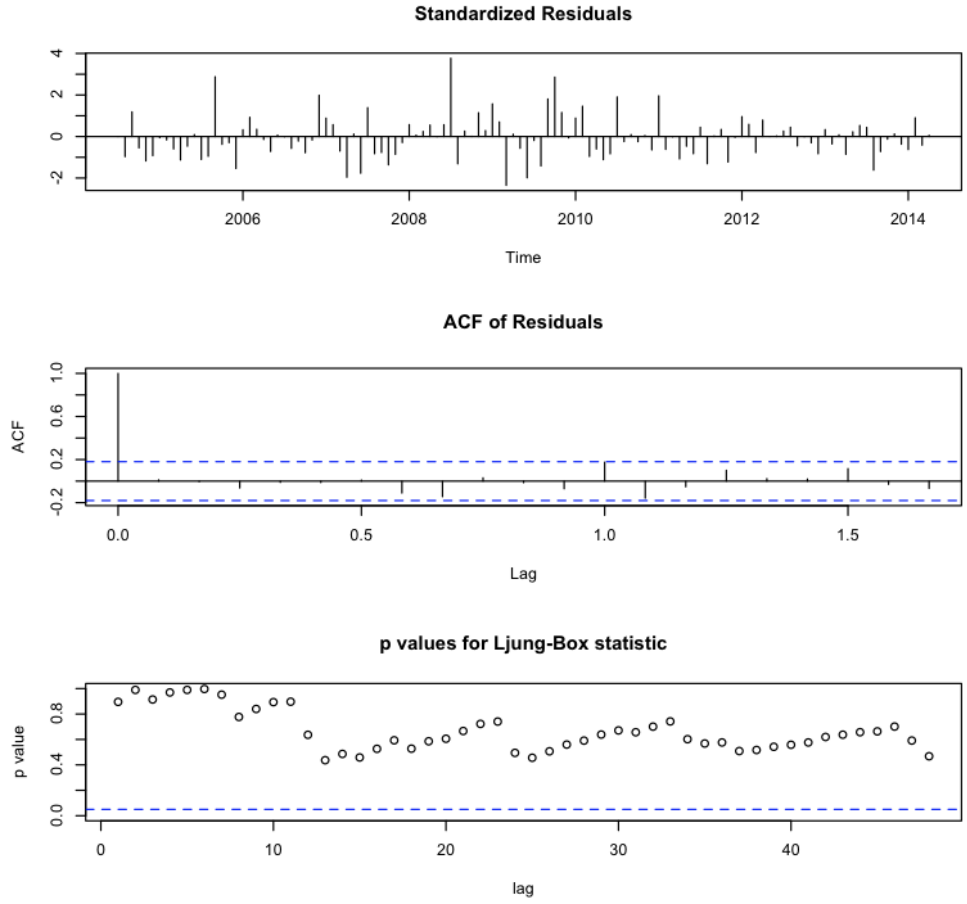


Figura A–19: Diagnóstico para la Series Latente $S_6$ de Índices de Precios al Consumidor, modelo SARIMA$(5,1,0)(0,0,0)_{12}$. Residuales estandarizados, ACF de los residuales y p-valores para la prueba de *Ljung-Box*.

7. $S_7$: SARIMA$(4,1,2)(0,0,0)_{12}$

```
Coefficients:
         ar1      ar2     ar3      ar4      ma1     ma2
      1.3727  -0.7010  0.1849  -0.1580  -1.8158  0.8158
s.e.  0.1361   0.1656  0.1551   0.1041   0.1198  0.1187
```

Luego el modelo estadístico es de la forma:

$$(1-1.37L+0.7L^2-0.1849L^3+0.158L^4)\Delta X_t = (1+1.8158L-0.8158L^2)\varepsilon_t \quad \text{(A.11)}$$

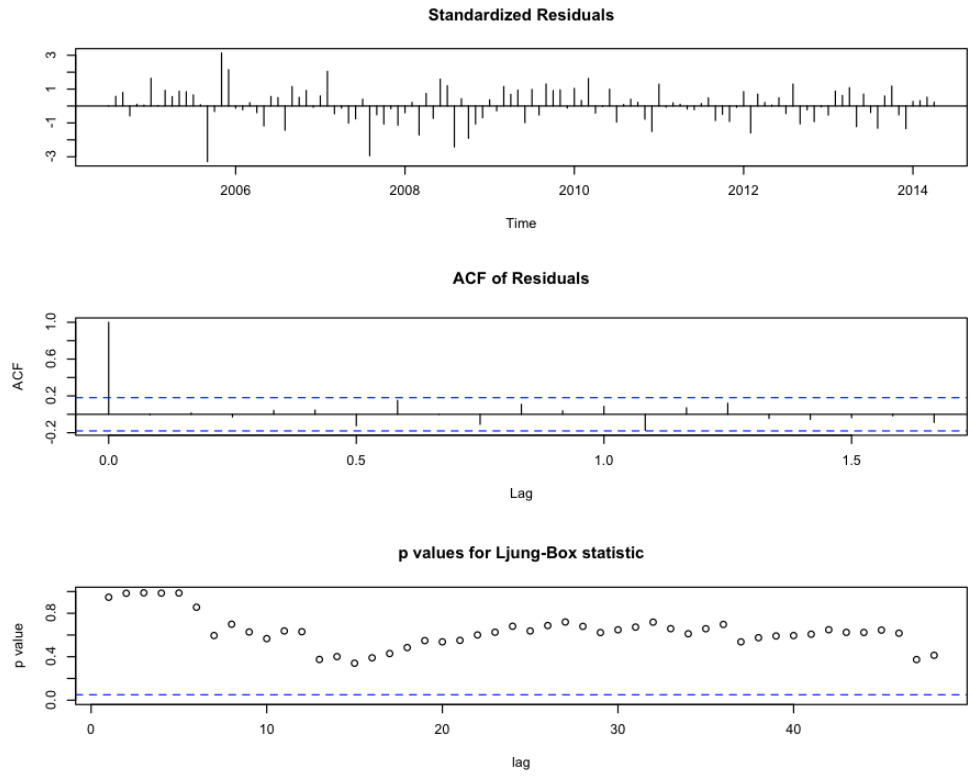


Figura A–20: Diagnóstico para la Series Latente $S_7$ de Índices de Precios al Consumidor, modelo SARIMA$(4,1,2)(0,0,0)_{12}$. Residuales estandarizados, ACF de los residuales y p-valores para la prueba de *Ljung-Box*.

8. $S_8$: SARIMA$(4,1,2)(0,0,0)_{12}$

```
Coefficients:
         ar1     ar2     ar3     ar4      ma1     ma2
      1.6242  -0.814  0.0170  0.0459  -1.9985  0.9996
s.e.  0.0947   0.176  0.1758  0.0933   0.0449  0.0449
```

Luego el modelo estadístico es de la forma:

$$(1-1.62L+0.814L^2-0.017L^3-0.0459L^4)\Delta X_t = (1+1.9985L-0.9996L^2)\varepsilon_t \quad \text{(A.12)}$$

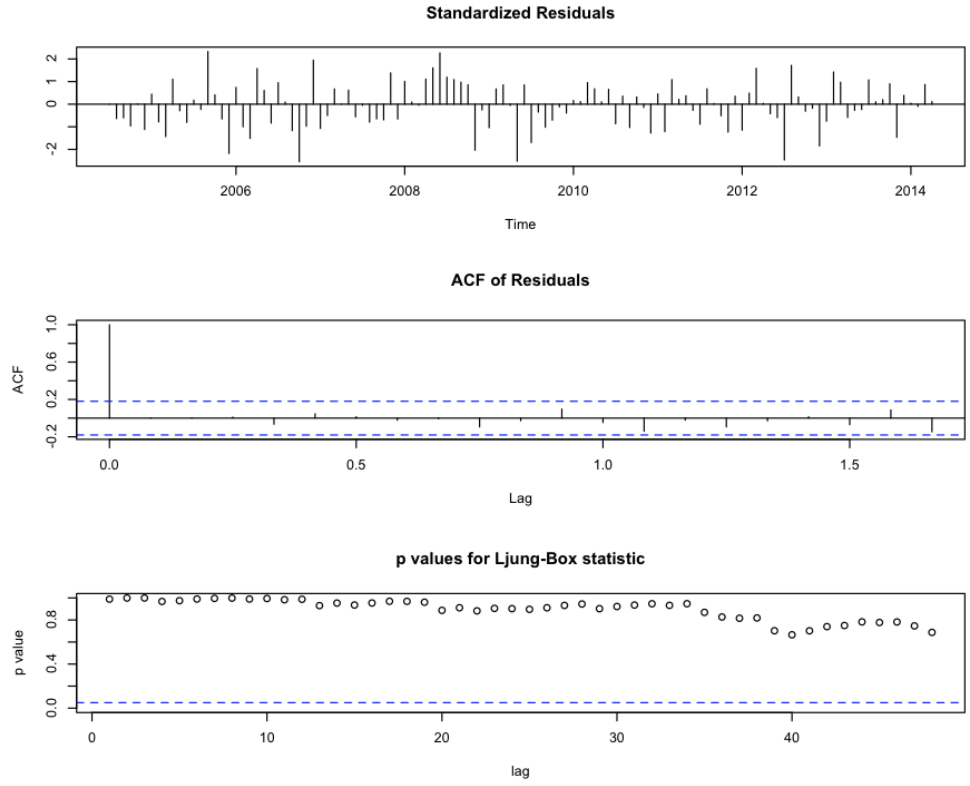


Figura A–21: Diagnóstico para laSeries Latente $S_8$ de Índices de Precios al Consumidor, modelo SARIMA$(4,1,2)(0,0,0)_{12}$. Residuales estandarizados, ACF de los residuales y p-valores para la prueba de *Ljung-Box*.

### A.2.2. SARIMA-AMUSE $k = 1$

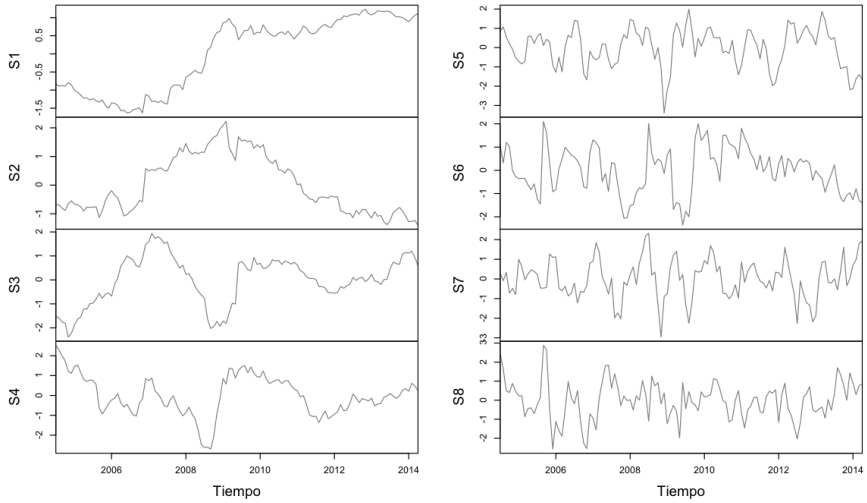


Figura A–22: Series Latentes de Índice de Precios al Consumidor, estimadas mediante AMUSE con $k = 1$

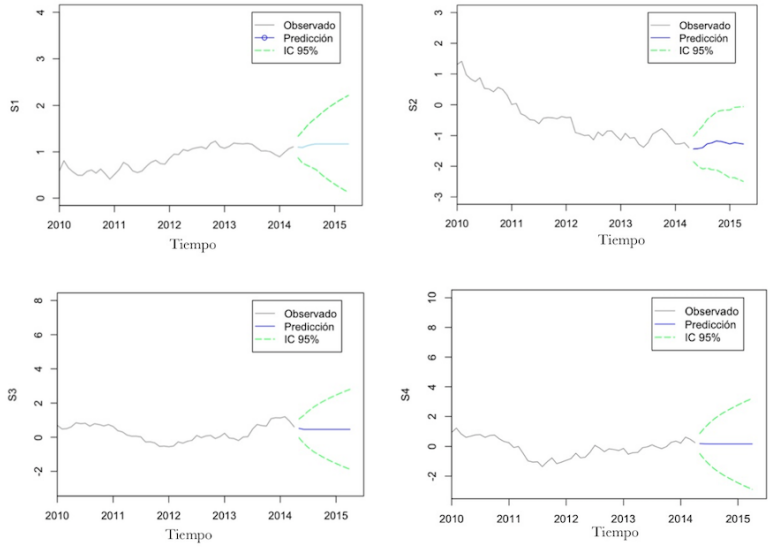


Figura A–23: Predicciones e Intervalos de Predicción del 95 % de Confianza de las cuatro primeras Series Latentes de Índice de Precios al Consumidor, para el período mayo del 2014 a abril del 2015, con AMUSE $k = 1$

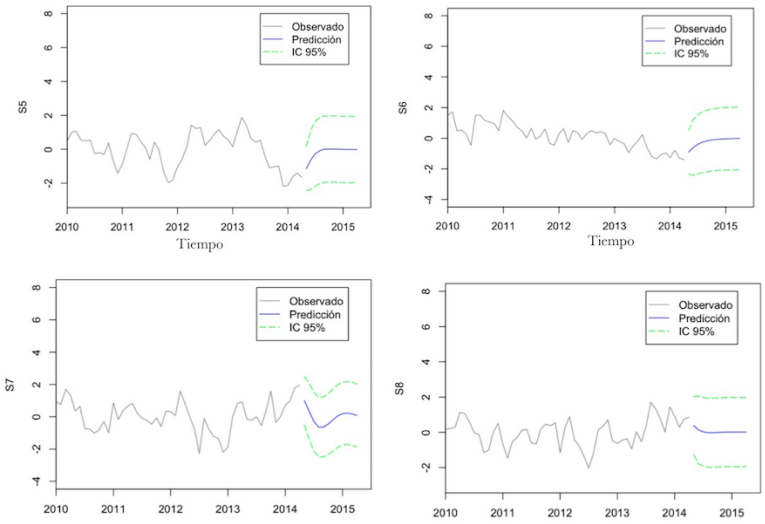


Figura A–24: Predicciones e Intervalos de Predicción del 95 % de Confianza de las cuatro últimas Series Latentes de Índice de Precios al Consumidor, para el período mayo del 2014 a abril del 2015, con AMUSE $k = 1$

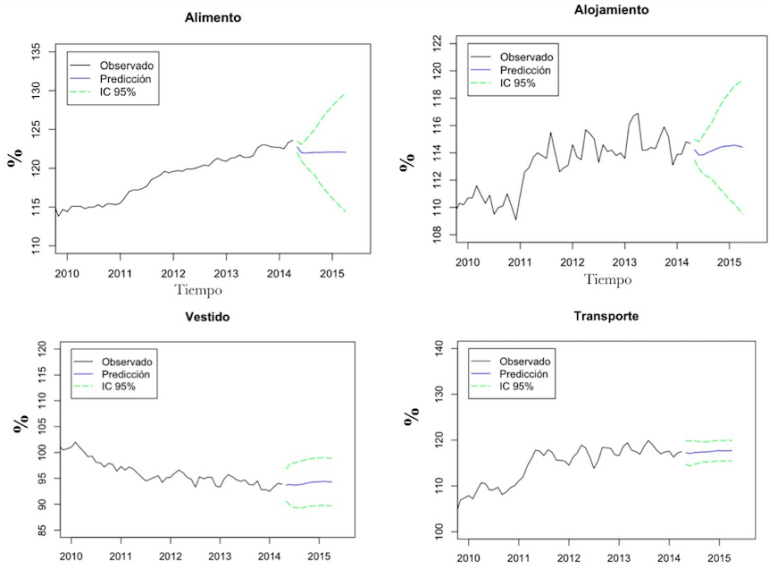


Figura A–25: Predicciones e Intervalos de Predicción del 95 % de Confianza de las cuatro primeras Series de Índice de Precios al Consumidor, para el período mayo del 2014 a abril del 2015, con AMUSE $k = 1$

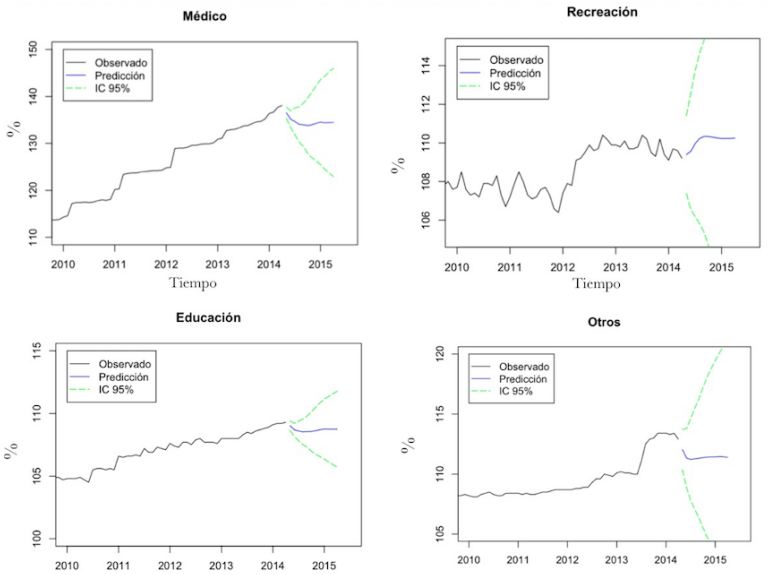


Figura A–26: Predicciones e Intervalos de Predicción del 95 % de Confianza de las cuatro últimas Series de Índice de Precios al Consumidor, para el período mayo del 2014 a abril del 2015, con AMUSE $k = 1$

# Apéndice B
# RUTINAS EN R PARA EL ANÁLISIS DE DATOS

A continuación se muestran las rutinas creadas en el lenguaje de programación R y el uso de las librerías *tseries* y *JADE*, que permiten realizar el ciclo iterátivo de *Box and Jenkins* y la estimación de las componentes independientes bajo los algoritmos SOBI y AMUSE, respectivamente. Consideramos únicamente la experimentación de la serie de tiempo multivariada consumo de energía eléctrica.

```
# Librerias requeridas
require(tseries); require(JADE)
# Descripcion de las series de tiempo
x<-read.table("ica.txt",header=T)
tsx<-ts(x,start=c(1999, 7), frequency=12)
plot.ts(tsx, main="Series de tiempo: Consumo de Energia Elelectrica",
xlab="Tiempo", xaxs = "i")
acf(tsx)
# Estimacion de los componentes independientes
res<-SOBI(tsx,k=100)
res<-AMUSE(tsx,k=3)  #candidatos k=3,k=9,k=20
S1=res$S[,1]; S2=res$S[,2]; S3=res$S[,3]
S<-cbind(S1,S2,S3)
plot(S, col="#808080",main="Series de Tiempo: Componentes Independientes",
xlab="Tiempo",xaxs = "i")
acf(S)
# Ciclo de Box and Jenkins (SARIMA) para las componentes
```

```
# Primera componente independiente
fS1=arima(S1, order = c(0,1,2), seasonal = list(order=c(0,1,1)))
tsdiag(fS1,gof.lag=48)
f1=predict(fS1, n.ahead = 12)
plot(S1, xlim=c(2010,2015.5),ylim=c(-4,4),col="#808080",
main="Prediccion de las Componentes Independientes",xaxs = "i",xlab="")
lines(f1$pred,type="o",col="skyblue")
U=f1$pred+1.96*f1$se
L=f1$pred-1.96*f1$se
lines(U,type="l",col="lightgreen", lty=5)
lines(L,type="l",col="lightgreen",lty=5)
# Segunda componente independiente
fS2=arima(S2, order = c(5,1,1), seasonal = list(order=c(0,1,1)))
tsdiag(fS2,gof.lag=48)
f2=predict(fS2, n.ahead = 12)
plot(S2,xlim=c(2010,2015.5),col="#808080",xaxs = "i",,xlab="")
lines(f2$pred,type="o",col="skyblue")
U=f2$pred+1.96*f2$se
L=f2$pred-1.96*f2$se
lines(U,type="l",col="lightgreen", lty=5)
lines(L,type="l",col="lightgreen",lty=5)
# Tercera componente independiente
fS3=arima(S3, order = c(6,1,1), seasonal = list(order=c(0,1,1)))
tsdiag(fS3,gof.lag=48)
f3=predict(fS3, n.ahead = 12)
plot(S3, xlim=c(2010,2015.5),ylim=c(-4,2),col="#808080",xaxs = "i",
 xlab="Tiempo")
lines(f3$pred,type="o",col="skyblue")
U=f3$pred+1.96*f3$se
L=f3$pred-1.96*f3$se
lines(U,type="l",col="lightgreen", lty=5)
lines(L,type="l",col="lightgreen",lty=5)
# Estimacion de las predicciones de las series originales
```

```
fS=cbind(f1$pred,f2$pred,f3$pred)
eS=cbind(f1$se,f2$se,f3$se)
f=fS%*%t(solve(res$W))
e=eS%*%t(solve(res$W))
f=ts(f,start=c(2014, 4), frequency=12)
e=ts(e,start=c(2014, 4), frequency=12)
x=read.table("ica.txt",header=T)
tsx=ts(x,start=c(1999, 7), frequency=12)
mean=apply(tsx,2,mean)
sd=apply(tsx,2,sd)
a=f[,1]*sd[1]+mean[1]
b=f[,2]*sd[2]+mean[2]
c=f[,3]*sd[3]+mean[3]
a1=(f[,1]+2*e[,1])*sd[1]+mean[1]
a2=(f[,1]-2*e[,1])*sd[1]+mean[1]
b1=(f[,2]+2*e[,2])*sd[2]+mean[2]
b2=(f[,2]-2*e[,2])*sd[2]+mean[2]
c1=(f[,3]+2*e[,3])*sd[3]+mean[3]
c2=(f[,3]-2*e[,3])*sd[3]+mean[3]
fx=cbind(a,b,c)
ux=cbind(a1,b1,c1)
lx=cbind(a2,b2,c2)
# Prediccion de la Serie de tiempo residencial
plot(tsx[,1],ylim=c(300,900), ylab="mKWh",xlab="Tiempo",xlim=c(2010,2015.2),
main="Prediccion del Consumo de Energia Electrica Residencial")
lines(fx[,1],col="blue",type="o")
lines(ux[,1],col="green",lty=5)
lines(lx[,1],col="green",lty=5)
leg.txt <- c("Observado", "Prediccion",  "IC 95%")
legend(2013.5,900,leg.txt,col=c("black","blue","green"),lty = c(1,1,5))
# Prediccion de la Serie de tiempo comercial
plot(tsx[,2], ylab="mKWh",ylim=c(500,1000),xlab="Tiempo",xlim=c(2010,2015.2),
main="Prediccion del Consumo de Energia Electrica Comercial")
```

```
lines(fx[,2],col="blue",type="o")
lines(ux[,2],col="green",lty=5)
lines(lx[,2],col="green",lty=5)
leg.txt <- c("Observado", "Prediccion",  "IC 95%")
legend(2013.5,1000,leg.txt,col=c("black","blue","green"),lty = c(1,1,5))
# Prediccion de la Serie de tiempo industrial
plot(tsx[,3],ylim=c(80,400), ylab="mKWh",xlab="Tiempo",xlim=c(2010,2015.2),
main="Prediccion del Consumo de Energia Electrica Industrial")
lines(fx[,3],col="blue",type="o")
lines(ux[,3],col="green",lty=5)
lines(lx[,3],col="green",lty=5)
leg.txt <- c("Observado", "Prediccion",  "IC 95%")
legend(2013.5,400,leg.txt,col=c("black","blue","green"),lty = c(1,1,5))
# validacion dejando 12 ultimas observaciones fuera.
x=read.table("ica1.txt",header=T); x=scale(x)
tsx=ts(x,start=c(1999, 7), frequency=12)
res1<-SOBI(tsx,k=100)
res<-AMUSE(tsx,k=3)   #candidatos k=3, k=9, k=20
S1=res$S[,3]; S2=res$S[,2]; S3=res$S[,1]
S=cbind(S1,S2,S3)
fS1=arima(S1, order = c(0,1,2), seasonal = list(order=c(0,1,1)))
tsdiag(fS1,gof.lag=48)
f1=predict(fS1, n.ahead = 12)
fS2=arima(S2, order = c(5,1,1), seasonal = list(order=c(0,1,1)))
tsdiag(fS2,gof.lag=48)
f2=predict(fS2, n.ahead = 12)
fS3=arima(S3, order = c(6,1,1), seasonal = list(order=c(0,1,1)))
tsdiag(fS3,gof.lag=48)
f3=predict(fS3, n.ahead = 12)
fS=cbind(f1$pred,f2$pred,f3$pred)
f=fS%*%t(solve(res$W))
x=read.table("ica1.txt",header=T)
tsx=ts(x,start=c(1999, 7), frequency=12)
```

```
mean=apply(tsx,2,mean)
sd=apply(tsx,2,sd)
a=f[,1]*sd[1]+mean[1]
b=f[,2]*sd[2]+mean[2]
c=f[,3]*sd[3]+mean[3]
fx=cbind(a,b,c)
#cargando las 12 dejadas fuera
xr=read.table("ica2.txt",header=T)
error=xr-fx
e2=error^2
abse=abs(error)
eab=abse/xr
rmse=sqrt(apply(e2,2,mean))
mape=apply(eab,2,mean)
```