PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Aplicación del modelo de espacio de estados con errores correlacionados a la tasa de desempleo en Perú

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN ESTADÍSTICA

Presentado por:

Rafael Visa Flores

Asesora: Anna Sikov

Miembros del jurado:
Dr. Luis Hilmar Valdivieso Serrano
Dra. Anna Sikov
Dra. Rocio Paola Maehara Aliaga

Lima, 2020

Resumen

En este trabajo se presenta los modelos de espacio de estados con errores correlacionados, propuesto por Pfeffermann y Tiller (2006), aplicado a datos reales de la tasa de desempleo para Lima Metropolitana, cuya información es recolectada mediante la Encuesta Permanente del Empleo - EPE por el Instituto Nacional de Estadística e Informática. Estos modelos permiten dar tratamiento a series de tiempo con errores de medición correlacionados, la estimación de los componentes del modelo se realiza mediante el algoritmo recursivo de Pfeffermann y Tiller, y cuando los errores son independientes se utiliza el algoritmo recursivo del filtro de Kalman.

Se realizó un estudio de simulación con series de tiempo con errores correlacionados con el objetivo de comparar las predicciones obtenidas con el algoritmo del filtro de Kalman y el algoritmo de Pfeffermann y Tiller, resultando este último con menores errores de predicción.

Con la finalidad de comparar la aplicación del modelo de espacio de estados con errores correlacionados con una metodología muy conocida como el desarrollado por Box and Jenkins, se ajustó los datos de la tasa de desempleo a un modelo ARIMA, se comparó las predicciones de ambos modelos con las verdaderas observaciones, donde los errores de las predicciones fueron similares, sin embargo, el menor error cuadrático medio se obtuvo con el modelo de espacio de estados con errores correlacionados.

Palabras-clave: Modelos de espacio de estados, series de tiempo con errores correlacionados, modelos ARIMA, filtro de Kalman, tasa de desempleo, Algoritmo de Pfeffermann y Tiller.

Abstract

This work presents the state space models with correlated errors, proposed by Pfeffermann and Tiller (2006), applied to real data on the unemployment rate for Metropolitan Lima, whose information is collected through the Permanent Employment Survey by the National Institute of Statistics and Informatics. These models allow to treat time series with correlated sampling errors, the estimation of the model components is performed using the Pfeffermann and Tiller recursive algorithm, and when the errors are independent, the Kalman filter recursive algorithm is used.

A simulation study with time series with correlated errors was carried out in order to compare the predictions obtained with the Kalman filter algorithm and the Pfeffermann and Tiller algorithm, the latter resulting in lower prediction errors.

In order to compare the application of the state space model with correlated errors with a well-known methodology such as that developed by Box and Jenkins, the unemployment rate data was adjusted to an ARIMA model, the predictions of both models were compared with the true observations, where the errors of the predictions were similar, however, the smallest mean squared error was obtained with the state space model with correlated errors.
Keywords: The state space models, time series with correlated sampling errors, ARIMA models, the Kalman filter, the unemployment rate, Pfeffermann and Tiller algorithm.

Índice general

Capítulo 1

Introducción

1.1. Consideraciones preliminares

Las oficinas gubernamentales de estadísticas recogen información longitudinal para medir el empleo y desempleo de forma continua, y publican los respectivos indicadores periódicamente, esta información sirve de insumo para que diferentes instituciones públicas y privadas puedan estudiar la oferta y demanda del mercado laboral, y tomar decisiones de política pública en base a estas evidencias.

En el Perú, las estadísticas del mercado laboral las produce el Instituto Nacional de Estadística e Informática - INEI, la información para elaborar estas estadísticas se recoge de forma trimestral mediante la Encuesta Permanente del Empleo - EPE, con una muestra representativa a nivel de Lima Metropolitana. Entre las estadísticas laborales se publica la tasa de desempleo, considerada como uno de los indicadores relevantes dentro del rubro debido a que permite medir la capacidad de absorción del empleo de la economía de un país. Precisamente por esta razón, la tasa de desempleo goza de mayor interés a nivel macroeconómico, conocer estas cifras de forma oportuna y en escenarios futuros, proporciona una ventaja sustancial a los tomadores de decisiones para prever medidas adecuadas. El Banco Central de Reserva del Perú y el INEI publican las cifras de la tasa de desempleo mensualmente, con un rezago de 1 a 2 meses y sin realizar un análisis sobre la dinámica de la serie, como la tendencia o los efectos de estacionalidad que puedan tener las observaciones, en este sentido, existe la necesidad de contar con un modelo de series de tiempo con capacidad de proveer buenas estimaciones y proyecciones que permita analizar la dinámica de la tasa de desempleo en el Perú.

Este trabajo intenta responder esta necesidad, sin embargo, existe una característica importante a considerar en el proceso de modelamiento, de la cual no existe antecedentes a nivel local, aquí se pone especial énfasis a ese detalle. Los datos longitudinales como los de la tasa de desempleo tienen la característica de presentar correlación entre los errores de muestreo, debido a que su diseño muestral se basa en un panel rotativo de unidades de muestreo, tal como es el caso del diseño de la EPE, cuyo panel de viviendas es rotado cada 2 años y cada vivienda participa en la muestra en 2 trimestres consecutivos por año, es justamente ahí donde se presenta la autocorrelación de los errores muestrales cuando se traslapa la información. En estos escenarios, por ejemplo, la Oficina de Estadísticas Laborales

de EE. UU. utiliza *modelos de espacio de estados* para estimar los componentes como la tendencia y estacionalidad de la serie, y producir predicciones mensuales de la tasa de empleo y desempleo, cuya información es recogida mediante la Encuesta de Fuerza Laboral (Labour Force Survey), el cual tiene en su diseño un esquema de muestreo de panel rotativo, esto es, los hogares están en la muestra por 4 meses sucesivos, luego son dejados fuera de la muestra por 8 meses y nuevamente son encuestados por 4 meses más, al tener este diseño hay meses que la muestra se traslapa entonces se presenta la autocorrelación de los errores muestrales.

Dentro del contexto descrito en los párrafos anteriores, en este trabajo se plantea aplicar a los datos reales de la tasa de desempleo producidos por la EPE, un *modelo de espacio de estados con errores correlacionados*. Para estimar los componentes del modelo utilizaremos el algoritmo del *filtro de Kalman* para el caso cuando el *modelo de espacio de estados* no considere los errores correlacionados y el algoritmo de *Pfeffermann y Tiller* cuando el modelo sí considere los errores correlacionados, luego se contrasta las estimaciones obtenidas entre ambos. Adicionalmente, se ajustará los datos a un modelo *ARIMA* con la finalidad de comparar los enfoques y las predicciones con los obtenidos mediante los *modelos de espacio de estados* descritos líneas arriba.

La metodología de espacio de estados sirve como un paraguas metafóricamente hablando, para representar una amplia gama de series de tiempo como univariantes, multivariantes sean estos estacionarios o no estacionarios. En esta metodología se asume que el desarrollo en el tiempo del sistema bajo estudio, es determinado por una serie de vectores $\alpha_1, \alpha_2, ..., \alpha_n$ no observados, las cuales son asociados a una serie de observaciones $y_1, y_2, ..., y_n$; esta relación entre los vectores $\alpha_t's$ y las observaciones $y_t's$ es especificado por el *modelo de espacio de estados*, cuyo principal propósito es inferir las propiedades relevantes de los $\alpha_t's$ a partir del conocimiento de las observaciones $y_1, y_2, ..., y_n$, con el propósito de extraer señales y estimar los componentes del modelo.

Sin embargo, previamente es necesario expresar la serie de observaciones $y_1, y_2, ..., y_n$ como un *modelo estructural básico*, el cual nos permitirá formular la serie directamente en términos de los componentes de interés, como la tendencia, la pendiente, los efectos estacionales y los términos irregulares (Harvey, 1989). Por ejemplo, la serie de observaciones de la tasa de desempleo bajo estudio, presenta tendencia que junto a la pendiente permitirán identificar los cambios que ha tenido la serie en el tiempo, también presenta picos elevados cada cierto periodo que parecen reflejar la existencia de patrones de estacionalidad. Así cuando una serie es formulado en componentes de un *modelo estructural básico*, es posible llevar a la forma de un *modelo de espacio de estados* y estimar cada componente, lo cual nos permitirá analizar la dinámica de los componentes estimados, que es esencial en el tratamiento de series de tiempo.

Complementando a lo indicado en el párrafo anterior, en el tratamiento de series de tiempo se presenta situaciones que exigen trabajar con datos longitudinales como los descritos en los párrafos iniciales, cuya característica a resaltar es que los datos presentan autocorrelación, para el tratamiento de este tipo de datos es necesario trabajar con *modelos de espacio de estados* que considere esta característica, el riesgo de no considerar la autocorrelación en el

tratamiento de la serie, es que puede llevar a conclusiones erróneas, específicamente se puede confundir las correlaciones con los efectos de estacionalidad, lo que llevaría tomar medidas inexactas y de ahí su importancia y motivación para ser el punto central de este trabajo.

En este sentido, una representación de un *modelo de espacio de estados* que considere los errores correlacionados (Pfeffermann and Tiller, 2006) es el siguiente, donde cabe mencionar que pueden haber muchas notaciones dentro del campo de estos modelos.

$$\begin{aligned} &y_t = Z_t\alpha_t + \varepsilon_t, \ \text{ es la ecuación de medición, donde} \\ &E(\varepsilon_t) = 0, \ \ E(\varepsilon_t\varepsilon_t') = \Sigma_{tt}, \ \ E(\varepsilon_\tau\varepsilon_t') = \Sigma_{\tau t}, \\ &\alpha_t = T_t\alpha_{t-1} + \eta_t, \ \text{ es la ecuación de transición, donde} \\ &E(\eta_t) = 0, \ \ E(\eta_t\eta_t') = Q, \ \ E(\eta_t\eta_{t-k}') = 0, \ k > 0. \end{aligned} \tag{1.1}$$

También se asume que $E(\eta_t\varepsilon_\tau') = 0$ para todo τ y t. En resumen se puede decir que, el problema de estimación en los *modelos de espacio de estados* se basa en que podemos observar la serie y_t, pero se desea conocer el estado α_t, el cual no observamos y debemos estimar con base a la información disponible hasta t.

Para la estimación de los $\alpha_t's$, como ya mencionamos, en primera instancia utilizaremos el *filtro de Kalman* definido como un procedimiento recursivo que permite calcular el estimador óptimo del vector de estados α_t en el tiempo t, basado en la información disponible hasta el tiempo t. Este es el escenario más frecuente donde se da tratamiento a las series de tiempo y existe una amplia literatura al respecto como, Harvey A. C. (1989), Durbin J. and Koopman S. J. (2012), Jacques J., F. Commandeur and S. J: Koopman (2007) por citar algunos de los mas relevantes.

Sin embargo, el algoritmo del *filtro de Kalman*, no supone la existencia de errores correlacionados, es decir, para cuándo $E(\varepsilon_\tau\varepsilon_t') = 0$ en (1.1). Considerando este detalle, *Pfeffermann y Tiller* (2006) desarrollaron un algoritmo de filtro recursivo (en adelante llamaremos como *algoritmo de Pfeffermann y Tiller*), que sí toma en cuenta los errores correlacionados, esto es, cuando $E(\varepsilon_\tau\varepsilon_t') = \Sigma_{\tau t}$ en el modelo (1.1). Para aplicar este algoritmo en la estimación de los componentes del modelo es necesario precisar el orden de las correlaciones, revisando los datos vemos que por ejemplo, en los datos de la EPE para el 2017 se identificó que, la muestra para el mes de Abril fue 1,468 viviendas de las cuales 715 viviendas ya habían participado en la muestra de Enero del mismo año, de forma similar la muestra de Mayo tenía 1,479 viviendas y de ellas 729 ya habían sido tomados en cuenta para la muestra de Febrero, todos dentro del mismo año. Estas traslapaciones parciales introducen correlaciones entre los estimadores de Enero y Abril, así como entre los estimadores de Febrero y Mayo respectivamente, esa misma lógica se presenta en los datos en general. Entonces la serie de datos de la tasa de desempleo pueden tener correlaciones de orden 3, 6, 9 y 12 conforme a los datos revisados y al diseño del panel de rotación de la muestra detallado en el tercer párrafo de este capítulo.

Los resultados empíricos muestran que las predicciones de las observaciones originales son más próximos a los datos reales de la tasa de desempleo, con las estimaciones realizadas con el

algoritmo de Pfeffermann y Tiller, dado que las observaciones de la tasa de desempleo tienen errores de medida correlacionados. Esto coincide con los hallazgos de Pfeffermann y Tiller (2006), donde se indica que para series de tiempo con errores de medida correlacionados, el algoritmo recursivo propuesto por ellos mismo ofrece mejores estimaciones. Complementando el análisis, se contrasta estos resultados empíricos con la metodología propuesta por Box and Jenkins (1970), ajustando los datos a un modelo ARIMA llegando a resultados muy próximos entre ambos.

La construcción de un *modelo de espacio de estados* es posible para los datos de la EPE, siempre y cuando, se conozca a profundidad su diseño muestral y los microdatos contengan la información que permita desagregar en áreas menores al área total representado por Lima Metropolitana, por ejemplo, al no contar con código de ubigeo en los microdatos no se pudo trabajar a nivel distrital, que era el escenario deseado, se limitó a aplicar el modelo por agrupación de distritos.

1.2. Objetivos

El objetivo general del estudio es ajustar el *modelo de espacio de estados con errores correlacionados* a los datos reales de la tasa de desempleo de Perú y contrastar los resultados con los obtenidos con los modelos *ARIMA*, previamente realizando un estudio de simulación en un escenario similar para comparar la precisión de las estimaciones aplicando el algoritmo del *filtro de Kalman* y el *algoritmo de Pfeffermann y Tiller*. De manera específica:

- Desarrollar el marco teórico adecuado para los modelos de espacio de estados y ARIMA.
- Estudiar el diseño muestral subyacente de la Encuesta Permanente del Empleo.
- Revisar, limpiar, procesar y validar los micro datos de la EPE para el modelado de la tasa de desempleo.
- Elaborar un estudio de simulación para comparar la eficiencia de los algoritmos.
- Aplicar el modelo estructural básico y ARIMA a la tasa de desempleo.
- Realizar las estimaciones y predicciones para comparar los modelos.
- Presentar los resultados y hallazgos encontrados.

1.3. Organización del Trabajo

En el Capítulo (2), presentamos los conceptos básicos sobre series de tiempo, estacionariedad y modelos lineales. Luego en el Capítulo (3) se presenta la teoría subyacente detrás de los modelos de espacio de estados, modelos estructurales de series de tiempo, *filtro de Kalman* y el *algoritmo de Pfeffermann y Tiller*. Seguidamente en el Capítulo (4) se presenta el diseño y características de la Encuesta Permanente del Empleo - EPE, así como la definición de la tasa de desempleo, mientras que el Capítulo (5) contiene un estudio de simulación, y finalmente en el Capítulo (6) se presenta la aplicación de los modelos para la tasa de desempleo de Perú.

Capítulo 2

Conceptos Básicos

2.1. Introducción

En este capítulo vamos a desarrollar los conceptos básicos que serán utilizados en el estudio, entre estos definimos las series de tiempo como una realización de un proceso estocástico y cuando estas se consideran estacionarios. En el tratamiento básico de series de tiempo es esencial que las series cumplan con la condición de estacionariedad dado que en función a esa característica los modelos son usados para realizar proyecciones, también definiremos algunos modelos lineales de series de tiempo como los modelos autoregresivos (AR), de medias móviles (MA) y la combinación de ambas (ARMA), modelos ARIMA y SARIMA cuando las series son no estacionarios, teniendo esto en mente será más entendible los conceptos más avanzados que desarrollaremos en el siguiente capítulo.

2.2. Procesos estocásticos

Un proceso estocástico es un sistema que evoluciona en el tiempo mientras que experimenta fluctuaciones o cambios al azar. Se puede describir dicho sistema definiendo una familia de variables aleatorias $\{X_t\}$, donde t es un punto en el espacio T llamado espacio de parámetros, y donde, para cada $t \in T$, X_t es un punto en el espacio S llamado espacio de estados (R. Coleman, 1974).

La familia $\{X_t\}$ puede ser considerado como el camino de una partícula que se mueve al (azar) en el espacio S, siendo su posición en el momento t, X_t un registro de estos caminos se denomina realización del proceso.

Un fenómeno estadístico que evoluciona en el tiempo de acuerdo con las leyes probabilísticas se denomina un proceso estocástico (Box and Jenkins, 1976). Una serie de tiempo entonces puede considerarse como una realización particular, producido por un mecanismo de probabilidad subyacente del sistema en estudio. Es decir, nos referimos a una serie de tiempo como una realización de un proceso estocástico.

2.3. Series de tiempo

Una serie de tiempo consiste en un conjunto de observaciones generados secuencialmente en el tiempo de una variable Z, si el conjunto cuenta con T observaciones se puede denotar

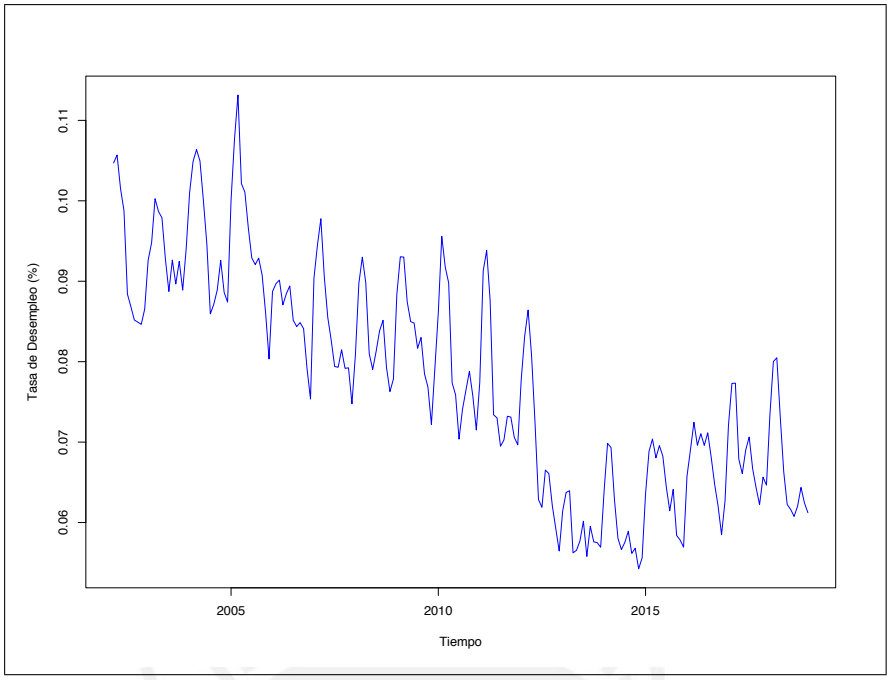


Figura 2.1: Tasa de desempleo de Lima Metropolitana según trimestres móviles 2002-2019

la serie de tiempo con Z_t donde $t = 1, 2, 3, ..., T$. Una serie de tiempo representado por Z_t es formulado en términos de los valores pasados de Z_t y/o errores, tomados sequencialmente en el tiempo, que pueden ser anual, mensual, trimestral, semanal o diario.

Las series de tiempo pueden considerarse continuos si t es continuo, y si t es discreto se dice que son series de tiempo discretos. En este trabajo solo consideraremos el caso discreto donde el intervalo de tiempo es fijo (t es mensual o trimestral), en adelante nos referiremos únicamente como series de tiempo al caso discreto, por ejemplo; la serie de tiempo que se muestra en la Figura (2.1) consta de observaciones trimestrales sobre la variación de la tasa de desempleo para Perú en el transcurso de los años de 2002 a 2019, la misma serie que será motivo de nuestro estudio.

Hay dos razones por las cuales se desea estudiar una serie de tiempo, descripción y modelamiento; el primero tiene por objetivo describir las principales características de la serie y el modelar una serie de tiempo tiene por objetivo permitir hacer proyecciones de futuros valores de la serie, las dos formas son complementarias, la misma información es procesada en diferentes formas lo que proporciona diferentes percepciones de la naturaleza de una serie de tiempo. Por ejemplo, se puede describir si la serie es o no estacionaria y si muestra patrones de estacionalidad, y con estas características modelar la serie buscando el modelo más optimo para realizar pronósticos.

2.4. Procesos estocásticos estacionarios

Un proceso estocástico se dice que es estacionario en sentido estricto, si sus propiedades no se ven afectadas por un cambio en el origen del tiempo, es decir, si la distribución de probabilidad conjunta asociada con m observaciones $Z_{t_1}, Z_{t_2}, ..., Z_{t_m}$, hecha en algún conjunto de tiem-

po $t_1, t_2, ..., t_m$, es la misma, como la asociada a las m observaciones $Z_{t_1+k}, Z_{t_2+k}, ..., Z_{t_m+k}$ hechas en los tiempos $t_1+k, t_2+k, ..., t_m+k$, para todo k entero, (Box, Jenkins, Reinsel and Ljung, 2015).

Para $m = 1$, la asunción de estacionariedad implica que la distribución de probabilidad $p(z_t)$ es el mismo para todos los tiempos t y puede ser escrito como $p(z)$, por lo que el proceso tiene una media constante que define el nivel de fluctuación[1](Box et al., 2015):

$$\mu = E[Z_t] = \int_{-\infty}^{\infty} zp(z)dz,$$

y una varianza constante, definido como:

$$\sigma_z^2 = E[(Z_t - \mu)^2] = \int_{-\infty}^{\infty} (z-\mu)^2 p(z)dz,$$

que mide la variación respecto del nivel de fluctuación. A partir de que la distribución de probabilidad $p(z)$ es el mismo para todo los tiempos t, su forma puede ser inferido formando el histograma con la observaciones de la serie de tiempo observada. La asunción de estacionariedad también implica que la distribución de probabilidad conjunta $p(z_{t_1}, z_{t_2})$ sea el mismo para todo los tiempos t_1 y t_2, que están separados por un intervalo constante. En particular, se deduce que la covarianza entre los valores Z_t y Z_{t+k} separados por k intervalos de tiempo o por k rezagos, debe ser el mismo para todo t bajo la asunción de estacionariedad. Esta covarianza se denomina como la *autocovarianza* con rezago k y se define como (Box et al., 2015):

$$\gamma_k = cov[Z_t, Z_{t+k}] = E[(Z_t - \mu)(Z_{t+k} - \mu)],$$

y de forma similar, la *autocorrelación* con rezago k es:

$$\begin{aligned} \rho_k &= \frac{E[(Z_t - \mu)(Z_{t+k} - \mu)]}{\sqrt{E[(Z_t - \mu)^2]E[(Z_{t+k} - \mu)^2]}} \\ &= \frac{E[(Z_t - \mu)(Z_{t+k} - \mu)]}{\sigma_z^2}. \end{aligned}$$

En un proceso estacionario, lo esencial es que sus propiedades no cambien en el tiempo, así la media μ permanece constante y es independiente del tiempo t, de igual forma para la varianza $\sigma_z^2 = \gamma_0$ es el mismo en el tiempo $t+k$ como en el tiempo t. Entonces la autocorrelación con rezago k, es la correlación entre Z_t y Z_{t+k}:

$$\rho_k = \frac{\gamma_k}{\gamma_0}$$

donde para $k = 0$, la autocorrelación sería $\rho_0 = 1$. ρ_k mide el grado de correlación entre dos observaciones que varia entre 1 y -1, y su utilidad es sumamente funcional porque a partir de los coeficientes de autocorrelación se define la función de autocorrelación, el cual se detalla más adelante.

[1]Fluctuación se refiere a la variación de un parámetro respecto al tiempo que se presenta de manera cambiante.

2.4.1. Series de tiempo estacionarias

Entonces para considerar cuando una serie de tiempo es estacionaria, se dice que una serie de tiempo Z_t es estacionaria en sentido amplio si cumple las siguientes condiciones:

a) $E[Z_t] = \mu$ es constante para todo t.

b) $E[(Z_t - \mu)^2] = \sigma^2$ es constante para todo t.

c) $cov[Z_t, Z_{t+k}] = \gamma_k$ depende únicamente de la separación o rezago k y no de t.

En una serie de tiempo estacionaria es más sencillo poder obtener predicciones, pues la media y varianza son constantes y se puede asumir que seguirán comportandose de la misma forma en el futuro para predecir una nueva observación. También se pueden obtener intervalos de confianza para las predicciones asumiendo que Z_t sigue una distribución conocida como la normal, son razones para buscar que una serie sea estacionaria.

Funciones de autocovarianza y autocorrelación

Cuando un proceso estocástico satisface las condiciones indicadas de estacionariedad, sus propiedades dinámicas pueden ser resumidas mediante el gráfico de γ_k contra k, dado que el coeficiente de autocovarianza mide la covarianza entre dos valores Z_t y Z_{t+k} separados por una distancia k. La secuencia γ_k se le conoce como la *función de autocovarianza* del proceso. Las autocovarianzas pueden ser estandarizadas al ser divididos por la varianzas γ_0 del proceso (Andrew C. Harvey, 1989), esto produce las *autocorrelaciones*. De forma similar, la secuencia de los coeficientes de autocorrelación ρ_k, es llamado como la *función de autocorrelación* del proceso.

La función de autocorrelación es adimensional, dado que es independiente de la escala de medición de las series de tiempo. A partir de $\gamma_k = \rho_k \sigma_z^2$, conocer la función de autocorrelación $\{\rho_k\}$ y la varianza σ_z^2 es equivalente a conocer la función de autocovarianza $\{\gamma_k\}$. La función de autocorrelación que se muestra en la Figura (2.2) como una gráfica de los diagonales de la matriz de autorrelación, revela como cambia la correlación entre dos valores de la serie a medida que cambia la separación.

Dado que $\rho_k = \rho_{-k}$, la función de autocorrelación es necesariamente simétrica respecto a cero. En la práctica sólo es necesario visualizar la parte positiva de la función, es decir la mitad, tal como se muestra en la Figura (2.3) que muestra la parte positiva del gráfico anterior. Cuando hablamos de la función de autocorrelación típicamente nos referimos a la parte positiva. De lo mostrado previamente en esta sección y en la sección anterior, un proceso normal estacionario Z_t es completamente caracterizado por su media μ y su función de autocovarianza $\{\gamma_k\}$, o equivalentemente por su media μ, varianza σ_z^2, y su función de autocorrelación $\{\rho_k\}$.

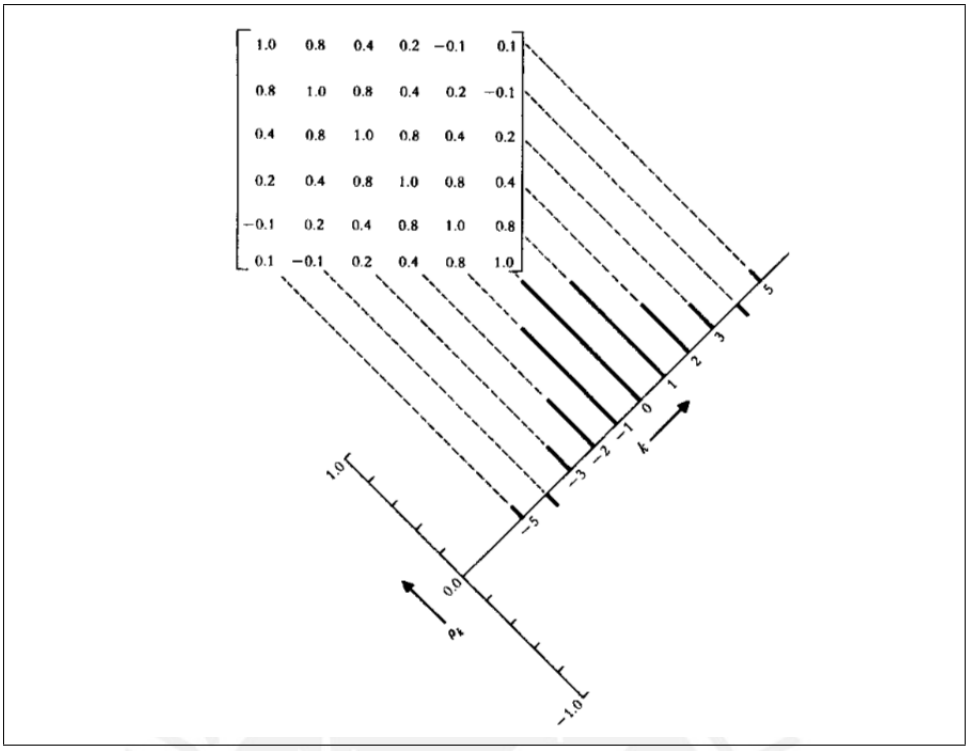


Figura 2.2: Matriz de autocorrelación y su correspondiente función de autocorrelación de un proceso estacionario (Box, Jenkins, Reinsel and Ljung, 2015)

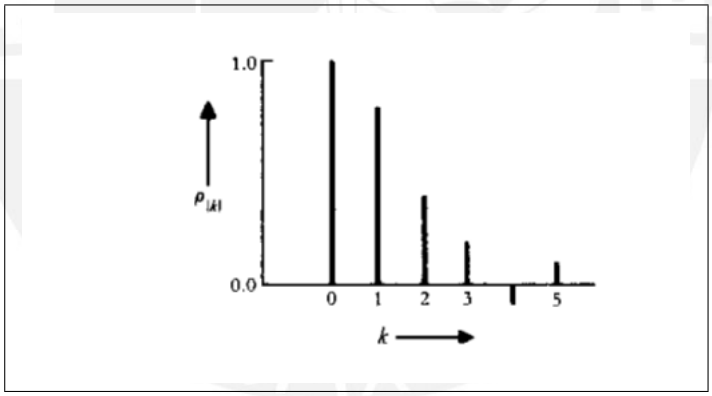


Figura 2.3: Parte positiva de la función de autocorrelación de la Figura 2.2

2.5. Procesos autorregresivos y de medias móviles

Los modelos estocásticos que utilizamos son basados originalmente de la idea de Yule (1927), que indica que, una serie de tiempo observable Z_t en la que los valores sucesivos son altamente dependientes, puede ser generada a partir de una serie independiente de "shocks" a_t. Estos shocks son generados aleatoriamente de una distribución fija, usualmente se asume que la distribución es normal con media cero y varianza σ_a^2, esta secuencia de variables aleatorias independientes $a_t, a_{t-1}, a_{t-2}, ...$ es llamado un proceso de perturbaciones o ruidos blancos.

El proceso de ruido blanco a_t es transformado a un proceso Z_t mediante un *filtro lineal*, como se puede ver en la Figura 2.4. La operación del filtro lineal consiste en tomar una suma

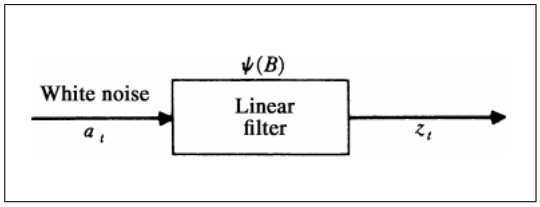


Figura 2.4: Representación de una serie de tiempo como resultado de un filtro lineal

ponderada de los shocks aleatorios anteriores de a_t, (Box et al., 2015) esto es:

$$\begin{aligned} Z_t &= \mu + a_t + \Psi_1 a_{t-1} + \Psi_2 a_{t-2} + ... \\ &= \mu + \Psi(B) a_t. \end{aligned} \tag{2.1}$$

En general, μ es un parámetro que determina el nivel del proceso y

$$\Psi(B) = 1 + \Psi_1 B + \Psi_2 B^2 + ...$$

es el operador lineal[2] que transforma a_t en Z_t y es denominado como la *función de transferencia* del filtro. La ecuación (2.1) también puede expresarse de la forma siguiente:

$$\begin{aligned} y_t &= a_t + \Psi_1 a_{t-1} + \Psi_2 a_{t-2} + ... \\ &= a_t + \sum_{j=1}^{\infty} \Psi_j a_{t-j}, \end{aligned} \tag{2.2}$$

donde $y_t = Z_t - \mu$ es la desviación del proceso desde algún origen o desde su media, si el proceso es estacionario. El *proceso lineal general* (2.2) permite representar y_t como la suma ponderada de los valores presentes y pasados de los disturbios o ruidos blancos a_t. Este proceso que consiste de una secuencia de variables aleatorias tiene media cero y varianza constante:

$$E[a_t] = 0 \quad y \quad var[a_t] = \sigma_a^2,$$

dado que las variables aleatorias a_t se asumen no correlacionadas, entonces su función de autocovarianza es:

$$\gamma_k = E[a_t a_{t+k}] = \begin{cases} \sigma_a^2 & k = 0, \\ 0 & k \neq 0, \end{cases}$$

así, la función de autocorrelación de los ruidos blancos tiene una forma particular simple:

$$\rho_k = \begin{cases} 1 & k = 0, \\ 0 & k \neq 0. \end{cases}$$

Un resultado fundamental en el desarrollo de procesos estacionarios es la de Wold (1938), quien estableció que cualquier proceso estacionario puramente no determinístico y_t posee una

[2]El operador de cambio hacia atrás B es definido por $Bz_t = z_{t-1}$; entonces en general $B^m z_t = z_{t-m}$

representación lineal como en la ecuación (2.2) con; $\sum_{j=0}^{\infty} \Psi_j^2 < \infty$, para asegurar que el proceso tenga varianza finita o constante. Los a_t con varianza común σ_a^2 son no correlacionados pero no necesitan ser independientes.

Para que y_t definido en la ecuación (2.2) represente un proceso estacionario válido, es necesario que los coeficientes Ψ_j sean *absolutamente sumables*, esto es; $\sum_{j=0}^{\infty} |\Psi_j| < \infty$, condición ligeramente más fuerte y necesario para algunos propósitos (ver Fuller, 1976). Bajo estas condiciones confortables y_t también puede ser expresado como la suma ponderada de los y_t's más un shock a_t (ver Koopmans, 1974), esto es:

$$\begin{aligned} y_t &= \pi_1 y_{t-1} + \pi_2 y_{t-2} + ... + a_t \\ &= \sum_{j=1}^{\infty} \pi_j y_{t-j} + a_t, \end{aligned} \tag{2.3}$$

tambiém puede ser escrito como

$$\begin{aligned} (1 - \sum_{j=1}^{\infty} \pi_j B^j) y_t &= a_t \\ \pi(B) y_t &= a_t, \end{aligned}$$

donde $\pi(B) = \Psi^{-1}(B)$. En la forma alternativa (2.3), la actual desviación y_t del nivel μ puede ser considerado como una regresión de sus desviaciones pasadas $y_{t-1}, y_{t-2}, ...$ del proceso.

2.5.1. Procesos Autorregresivos

La expresión de la ecuación (2.3) no es muy útil en la práctica si contiene un número infinito de parámetros π_j. Podemos describir una forma de introducir la parsimonia y llegar a un modelo que sea útil para las aplicaciones prácticas. Consideremos que en la ecuación (2.3) los primeros p parámetros son diferentes de cero (Box et al., 2015). Entonces el modelo puede ser expresado como:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + a_t,$$

donde ahora usamos los símbolos $\phi_1, \phi_2, ..., \phi_p$ para un conjunto finito de parámetros de ponderación. El proceso resultante es llamado como un *proceso autorregresivo* de orden p, denotado por $AR(p)$. Este modelo puede ser escrito de forma equivalente como:

$$\begin{aligned} (1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) y_t &= a_t \\ \phi(B) y_t &= a_t. \end{aligned}$$

Esto implica que

$$y_t = \phi^{-1}(B) a_t \equiv \Psi(B) a_t.$$

Por lo tanto, el proceso autorregresivo puede ser considerado como la salida y_t de un filtro lineal con función de transferencia $\phi^{-1}(B) = \Psi(B)$ cuando la entrada es un ruido blanco a_t.

El primer punto a establecer es, bajo qué condiciones un proceso AR es estacionario. Para ilustrar veamos como ejemplo, para el proceso $AR(1)$:

$$y_t = \phi_1 y_{t-1} + a_t, \quad t = 1, 2, ..., T.$$

A pesar que la serie es observada por primera vez cuando el tiempo es $t = 1$, se considera que el proceso empezó en algún punto en el pasado. Sustituyendo repetidamente los rezagos de y_t, se llega a obtener una expresión de la forma:

$$\begin{aligned} y_t &= \phi y_{t-1} + a_t \\ &= \phi(\phi y_{t-2} + a_{t-1}) + a_t \\ &= \phi^2 y_{t-2} + \phi a_{t-1} + a_t \\ &= \phi^2(\phi y_{t-3} + a_{t-2}) + \phi a_{t-1} + a_t \\ &= \phi^3 y_{t-3} + \phi^2 a_{t-2} + \phi a_{t-1} + a_t \\ &... \\ &= \phi^j y_{t-j} + \sum_{j=0}^{j-1} \phi^j a_{t-j}. \end{aligned} \tag{2.4}$$

El resultado consiste de dos partes, la primera parte depende de los valores de y_t en el tiempo $t-j$ y la segunda parte es la sumatoria de valores rezagados de los ruidos blancos cuya media es cero y varianza constante. Entonces tomando la esperanza en la expresión (2.4) y tratando como y_{t-j} como un número fijo, se tiene:

$$E(y_t) = E(\phi^j y_{t-j}) + E(\sum_{j=0}^{j-1} \phi^j a_{t-j}) = \phi^j y_{t-j}. \tag{2.5}$$

En la expresión (2.5) si $|\phi| \geq 1$, la esperanza depende de y_{t-j}. Por lo tanto, la expresión (2.4) contiene un componente determinista y el conocimiento de y_{t-j} permite hacer predicciones no triviales para valores futuros de la serie, sin importar cuán lejos esté, por tanto sería no estacionario. Por otro lado, si $|\phi| < 1$, el componente determinista es trivial a medida que j sea más grande. Sí $j \to \infty$ efectivamente este componente desaparece, por tanto si se considera que el proceso comenzó en algún momento en el pasado, la expresión (2.4) se puede escribir, como:

$$y_t = \sum_{j=0}^{\infty} \phi^j a_{t-j}, \quad t = 1, 2, ..., T.$$

Con está última expresión se puede ver qué un proceso $AR(1)$ con $|\phi| < 1$ es no determinístico. Por lo tanto, conforme a lo citado por Wold (1938) cualquier proceso puramente no determinista con $\sum_{j=0}^{\infty} \phi_j^2 < \infty$ es estacionario. El proceso $AR(1)$ es estacionario dado que la suma de los cuadrados de los coeficientes es una progresión geométrica, por tanto converge hacia un valor finito:

$$\sum_{j=0}^{\infty} \phi^{2j} = 1 + \phi^2 + \phi^4 + ... = \frac{1}{1 - \phi^2} < \infty.$$

2.5.2. Procesos de medias móviles

De manera similar que para los modelos AR, de la expresión definida en (2.2), en la práctica no es útil si contiene un número infinito de parámetros Ψ_j. Considerando los primeros q parámetros de Ψ diferentes de cero (Box et al., 2015). Entonces el proceso puede ser expresado de la siguiente forma:

$$y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q},$$

donde ahora usamos los símbolos $-\theta_1, -\theta_2, ..., -\theta_q$ para un conjunto finito de parámetros ponderadores. Este proceso es llamado como un *proceso de medias móviles* de orden q, el cual se expresa de forma abreviada como $MA(q)$ (por sus iniciales en ingles de Moving Average). Utilizando el operador de rezagos $Ba_t = a_{t-1}$, un proceso $MA(q)$ puede ser expresado de forma equivalente como:

$$\begin{aligned} y_t &= (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q) a_t \\ y_t &= \theta(B) a_t. \end{aligned}$$

Por lo tanto, un *proceso de medias móviles (MA)* puede ser considerado como una salida y_t de un filtro lineal con función de transferencia $\theta(B)$ cuando la entrada sea un ruido blanco a_t. Un proceso finito de medias móviles es siempre estacionario al satisfacer automáticamente las condiciones de estacionariedad (Harvey, 1993).

2.5.3. Procesos mixtos autorregresivos de medias móviles

En la construcción de modelos, puede ser necesario construir un modelo mixto para obtener una parametrización parsimoniosa, para esto es frecuente incluir ambos términos autorregresivo y medias móviles en el mismo modelo (Box et al., 2015). El modelo resultante sería:

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + a_t - \theta_1 a_{t-1} - ... - \theta_q a_{t-q}, \tag{2.6}$$

equivalentemente puede ser expresado como:

$$\begin{aligned} (1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) y_t &= (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q) a_t \\ \phi(B) y_t &= \theta(B) a_t \\ y_t &= \frac{\theta(B)}{\phi(B)} a_t, \end{aligned} \tag{2.7}$$

este proceso es llamado autorregresivo de medias móviles de orden (p,q), el cual en forma abreviada se denota por $ARMA(p, q)$. Vemos que un proceso mixto $ARMA$ puede considerarse como un resultado o salida y_t de un filtro lineal, cuya función de transferencia es la razón de dos operadores polinomiales $\theta(B)$ y $\phi(B)$, cuando la entrada o input es un ruido blanco a_t.

Ahora, que un proceso $ARMA$ sea estacionario o no, únicamente depende de la parte autorregresiva pues como se aprecia en la expresión (2.6), es claro que la parte de medias

móviles a la derecha no afecta los términos anteriores que establecen la condición de estacionario de un proceso autorregresivo. Entonces, para definir un proceso estacionario $ARMA$, de la expresión (2.7) la ecuación característica $\phi(B) = 0$ debe tener todas sus raíces fuera del círculo unitario[3].

2.6. Modelos ARIMA

En la práctica la mayoría de las series de tiempo son no estacionarios, para realizar el tratamiento de una serie no estacionaria es necesario eliminar la fuente de variación no estacionario. Si la serie observada es no estacionaria en la media, según lo definido en (2.4.1) entonces la serie presenta una tendencia no constante, es estos casos es posible tomar diferencias de las series hasta obtener una serie estacionaria. En este sentido, Box y Jenkins (1970) plantea los modelos ARIMA, esto se deriva de la ecuación (2.7) cuando la raíz de $\phi(B) = 0$ cae en el círculo unitario donde se pueden representar series de tiempo no estacionarios, para lo cual se considera el siguiente modelo:

$$\varphi(B)y_t = \theta(B)a_t,$$

donde $\varphi(B)$ es un operador autorregresivo no estacionario, tal que d de las raíces de $\varphi(B) = 0$ son unitarios y el resto permanecen fuera del círculo unitario, entonces el modelo puede ser escrito como:

$$\varphi(B)y_t = \phi(B)(1-B)^d y_t = \theta(B)a_t,$$

donde $\phi(B)$ es un operador autorregresivo estacionario, y $\nabla = 1 - B$ es el operador de diferenciación, entonces el modelo queda como:

$$\phi(B)\nabla^d y_t = \theta(B)a_t, \tag{2.8}$$

para $d \geq 1$ y diferenciando d veces puede obtenerse un proceso estacionario, aquí la expresión (2.8) se denomina modelo autorregresivo integrado y medias móviles *ARIMA(p,d,q)*, donde el operador autorregresivo $\phi(B)$ es de orden p, d es el número de diferencias tomadas y el operador de medias móviles $\theta(B)$ es de orden q.

2.7. Modelos SARIMA

Los modelos SARIMA captan el comportamiento puramente estacional de una serie, en la misma lógica de los modelos ARIMA se realiza las diferencias al componente no estacional de la serie. Entonces una serie influenciada por el componente puramente estacional puede ser descrito por un modelo SARIMA(P,D,Q), el cual se representa por la siguiente expresión:

$$\Phi_P(B^s)\nabla_s^D y_t = \Theta_Q(B^s)a_t, \tag{2.9}$$

donde $\Phi_P(B^s) = 1 - \Phi_s(B^s) - \Phi_{2s}(B^{2s}) - ... - \Phi_{Ps}(B^{Ps})$ es un polinomio autorregresivo estacional de orden P y $\Theta_Q(B^s) = 1 - \Theta_s(B^s) - \Theta_{2s}(B^{2s}) - ... - \Theta_{Qs}(B^{Qs})$ es un polinomio

[3]Es decir, todas las raíces de la ecuación indicada deben ser mayores a 1 en valor absoluto, $|B| > 1$

de medias móviles estacional de orden Q.

Es sabido que en la vida real lo que pasa con mucha frecuencia es que las series no siempre se presentan únicamente afectados por la tendencia o sólo por los efectos de estacionalidad, sino todo lo contrario, generalmente las series vienen afectadas por ambas, tendencia y estacionalidad. En este sentido, Box y Jenkins (1970) propone un modelo denominado multiplicativo, el cual tenga la capacidad de explicar el comportamiento de la serie afectada por la tendencia y estacionalidad, expresado como una combinación de los modelos (2.8) y (2.9):

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D y_t = \theta_q(B)\Theta_Q(B^s)a_t. \tag{2.10}$$

Entonces el modelo (2.10) se denomina modelo multiplicativo de orden $(p,d,q)\times(P,D,Q)$. Este modelo resulta ser muy útil en la práctica. Por ejemplo la serie de la tasa de desempleo para Lima Metropolitana presentado en el Figura (2.1) puede ser modelado por este modelo, debido a que en la gráfica se visualiza que la serie tiene tendencia pues presenta una tendencia negativa y al parecer tiene estacionalidad, con efectos estacionales de data mensual ($s = 12$). En el siguiente capítulo veremos los conceptos más avanzados para el tratamiento de los componentes mencionados de una serie de tiempo.

Capítulo 3

Marco Teórico

En este capítulo se desarrolla la teoría subyacente del tema de este estudio, donde desarrollaremos los conceptos esenciales referidos a; modelos de espacio de estados, modelos estructurales de series de tiempo y su descomposición en componentes, que es punto clave para la construcción de los modelos y que permite llevar a la forma de los modelos de espacio de estados. Luego corresponde presentar los métodos que serán utilizados para la estimación de los componentes del modelo, para el cual se recurrirá al uso del algoritmo del filtro de Kalman para el caso de series de tiempo con errores independientes y al algoritmo de Pfeffermann y Tiller para el caso de series de tiempo con errores no independientes. También se define la estimación por máxima verosimilitud utilizada para estimar los parámetros de los modelos, finalizamos con un ejemplo práctico utilizando los conceptos y métodos desarrollados en el capítulo.

3.1. Modelo de espacio de estados

La forma de los modelos de espacio de estados es una herramienta enormemente poderosa, el cual abre camino para manejar una amplia gama de modelos de series de tiempo (Andrew C. Harvey, 1989). Una vez que una serie de tiempo es puesto en la forma de espacio de estados, puede aplicarse el *filtro de Kalman* (el cual definiremos más adelante) para realizar las estimaciones.

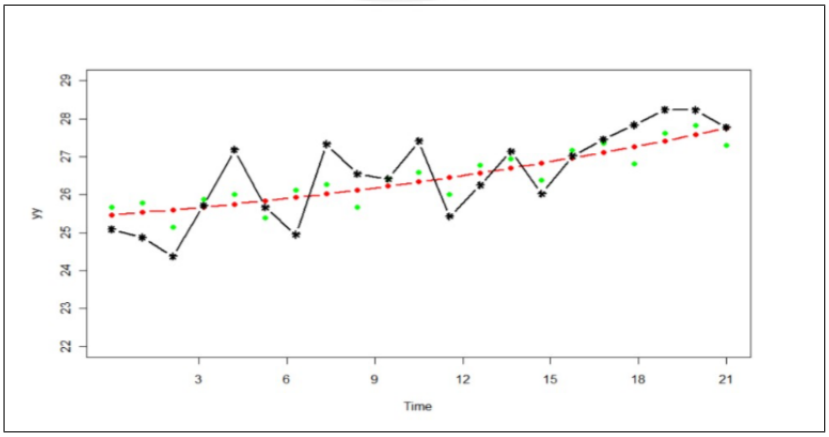


Figura 3.1: Serie de tiempo con componentes de tendencia, estacionalidad y error

Antes de presentar formalmente, partamos de la idea de una serie temporal como la que se muestra en la Figura (3.1). De este gráfico podemos indicar algunas características generales como la *tendencia*, el cual representa los movimientos de largo plazo de la serie, y patrones de *estacionalidad* que se repiten cada cierto periodo de tiempo. Estas son características que un modelo de serie temporal necesita capturar.

En una gráfica de series de tiempo como la que se muestra en (3.1), es usual ver como un todo la trayectoria de las observaciones de la serie (puntos negros) lo cual dificulta ver el comportamiento de cada componente de la serie, como la tendencia, estacionalidad y errores. Según la descomposición clásica de series de tiempo (Andrew C. Harvey, 1989), argumenta que con los modelos de espacio de estados es posible descomponer la serie y visualizar cada componente por separado como; la tendencia (línea punteada color rojo), pendiente obtenida de dos puntos continuos de la tendencia, los efectos de la estacionalidad que vendrían a ser la diferencia entre los puntos rojos y los puntos verdes, incluso es posible ver por separado los errores y ver si tienen un comportamiento aleatorio. De esa manera los modelos de espacio de estados facilitan la caracterización y análisis de las series de tiempo.

Partiendo de la descomposición clásica de una serie de tiempo definido en la sección (3.2) al que se denominará modelo estructural, es posible expresar la serie observada y_t en la forma de un modelo de espacio de estados, el cuál consiste de dos ecuaciones (Durbin and Koopman, 2012):

La *ecuación de observación* tiene la siguiente forma

$$y_t = Z_t\alpha_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H_t). \tag{3.1}$$

La *ecuación de estados* tiene la siguiente forma

$$\alpha_{t+1} = T_t\alpha_t + \eta_t, \quad \eta_t \sim N(0, Q_t), \quad t = 1, 2, 3, ..., n, \tag{3.2}$$

donde el vector de estados α_t puede particionarse como $\alpha_t = (\alpha'_{1t}, \alpha'_{2t}, ..., \alpha'_{mt})'$ siendo m el número de elementos del vector de estados y describiendo este vector la estructura característica de los modelos de espacio de estados. De otro lado, Z_t es un vector de orden $m \times 1$, T_t es la matriz de transición de orden $m \times m$, Q_t es la matriz defina como $E(\eta_t\eta'_t) = Q_t$ y $H_t = \sigma^2_\varepsilon$, donde se asume también que los errores son independientes entre sí, esto es; $E(\eta_t\varepsilon'_\tau) = 0$ para todo t y τ.

La idea subyacente del modelo es, que el desarrollo de un sistema a lo largo del tiempo está determinado por la evolución del vector de estados α_t de acuerdo a la ecuación (3.2), expresado como un proceso autorregresivo de primer orden $AR(1)$, pero como α_t no puede ser observado directamente, entonces debemos realizar el análisis en base a las observaciones y_t expresado en la ecuación (3.1) como una combinación lineal de los componentes no observados más el error de medición ε_t.

3.2. Modelo estructural de series de tiempo

Partiendo de la idea intuitiva expresada, de que una serie puede estar descompuesta en componentes de tendencia, estacionalidad y el término irregular, y representando la serie observada y_t como un conjunto de observaciones $y_1, y_2, ..., y_n$ ordenados en el tiempo, el modelo aditivo quedaría expresado de la siguiente forma (Durbin and Koopman, 2012):

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \quad t = 1, 2, 3, ..., n \tag{3.3}$$

donde, μ_t es un componente que varía lentamente llamado *tendencia*, γ_t es un componente de periodo fijo llamado *estacional* y ε_t es el componente irregular llamado *error* o *perturbación*. Entonces se denomina *modelo estructural de series de tiempo* al modelo expresado en (3.3) más otros componentes relevantes que son modelados explicitamente. A continuación desarrollamos cada uno de los componentes excepto el componente irregular debido a que se asume como una variable aleatoria con distribución normal.

3.2.1. Componente de tendencia

Consideremos una forma simple de expresar la ecuación (3.3) donde asumimos que ε_t tiene una varianza constante σ_ε^2, esto será:

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{3.4}$$

donde el componente de tendencia μ_t, es simplemente un nivel que fluctua hacia arriba y hacia abajo, conforme al paseo aleatorio no estacional y expresado como:

$$\mu_{t+1} = \mu_t + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2) \tag{3.5}$$

para $t = 1, 2, ..., n$ donde ε_t y η_t son mutuamente independientes. Sí a la tendencia expresada en (3.5) le agregamos el término de la pendiente ν_t, y que a su vez, es generada por paseo aleatorio, obtenemos el siguiente modelo:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, \qquad \xi_t \sim N(0, \sigma_\xi^2), \\ \nu_{t+1} &= \nu_t + \zeta_t, \qquad \zeta_t \sim N(0, \sigma_\zeta^2) \end{aligned} \tag{3.6}$$

Esta expresión (3.6) es llamado como *modelo de tendencia lineal local* (Durbin and Koopman, 2012) o simplemente como *modelo de nivel local* (Andrew C. Harvey, 1989). Un detalle a notar es, sí; $\xi_t = \zeta_t = 0$, entonces $\nu_{t+1} = \nu_t = \nu$ por lo que $\mu_{t+1} = \mu_t + \nu$ sería exactamente una tendencia lineal y la expresión (3.6) se reduce a una tendencia determinística lineal más el ruido. La expresión (3.6) con $\sigma_\xi^2 > 0$ y $\sigma_\zeta^2 > 0$ permite que el nivel de tendencia y la pendiente varíen con el tiempo.

3.2.2. Componente estacional

Para modelar el término estacional γ_t expresado en (3.3) supongamos que se tiene s meses al año. Así los efectos estacionales para un año pueden ser; $s = 12$ para observaciones mensuales, $s = 4$ para observaciones trimestrales, y $s = 7$ si las observaciones son de frecuencia diaria modelando semanalmente.

Si el patrón estacional es constante en el tiempo, las observaciones estacionales mensuales de 1 a s pueden ser modelados por las constantes $\gamma_1^*, ..., \gamma_s^*$ donde $\sum_{j=1}^{s} \gamma_j^* = 0$. Para el j-ésimo mes en el año i se tiene $\gamma_t = \gamma_s^*$ donde $t = s(i+1) + j$ para $i = 1, 2, ...$ y $j = 1, ..., s$. Resulta que $\sum_{j=0}^{s-1} \gamma_{t+1-j} = 0$ entonces $\gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j}$ con $t = s-1, s, ...$

En la práctica frecuentemente se desea que el patrón estacional cambie en el tiempo. Una manera simple de conseguir esto es agregando un término de error denotado por w_t a la relación y queda expresado como:

$$\gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j} + w_t, \qquad w_t \sim N(0, \sigma_w^2), \tag{3.7}$$

para $t = 1, ..., n$. Una alternativa sugerida por Harrison y Stevens (1976) es denotar el efecto estacional j en el tiempo t por γ_{jt} y este puede ser generado por un paseo cuasi-aleatorio:

$$\gamma_{j,t+1} = \gamma_{j,t} + w_{j,t}, \quad t = (i-1)s + j, \quad i = 1, 2, ..., \quad j = 1, ..., s,$$

con un ajuste para asegurar que cada conjunto sucesivo de s componentes estacionales sumen a cero (ver Harvey, 1989). Frecuentemente es preferible expresar los efectos estacionales en una forma trigonométrica, y una forma de expresar trigonométricamente la ecuación (3.7) es en la forma de un modelo de un paseo cuasi-aleatorio (Durbin and Koopman, 2012), de la siguiente forma:

$$\gamma_t = \sum_{j=1}^{[s/2]} \gamma_{j,t}, \tag{3.8}$$

donde $[s/2]$ es el entero más grande $\leq s/2$ y

$$\begin{aligned} \gamma_{j,t+1} &= \gamma_{j,t} cos\lambda_j + \gamma_{j,t}^* sen\lambda_j + w_{j,t}, \quad \lambda_j = \frac{2\pi j}{s}, \\ \gamma_{j,t+1}^* &= -\gamma_{j,t} sen\lambda_j + \gamma_{j,t}^* cos\lambda_j + w_{j,t}^*, \quad j = 1, ..., [s/2], \end{aligned} \tag{3.9}$$

donde los términos $w_{j,t}$ y $w_{j,t}^*$ son variables independientes distribuidos normalmente con media cero y varianza σ_w^2.

3.2.3. Modelo estructural básico

Cada uno de los modelos estacionales detallados en la sección anterior pueden combinarse con los modelos de tendencia mostrados en la sección (3.2.1) para dar un modelo estructural de series de tiempo, y estos a su vez, pueden ser expresados en la forma de modelos de espacio de estados como los que se muestran en las ecuaciones (3.1) y (3.2). Así, por ejemplo,

si combinamos el modelo de tendencia local (3.6) con el modelo estacional expresado en (3.8) y (3.9), se obtiene el modelo aditivo siguiente:

$$y_t = \mu_t + \gamma_t + \varepsilon_t; \quad t = 1, 2, 3, ..., n, \tag{3.10}$$

donde para representar en la forma del modelo de espacio de estados como en (3.1) y (3.2), podemos expresar el vector de estados como:

$$\alpha_t = (\mu_t, \nu_t, \gamma_t, \gamma_{t-1}, ..., \gamma_{t-s+2})',$$

y tomar el sistema de matrices como:

$$Z_t = \left(Z_{[\mu]}, Z_{[\gamma]}\right), \quad T_t = \left(T_{[\mu]}, T_{[\gamma]}\right), \quad Q_t = diag\left(Q_{[\mu]}, Q_{[\gamma]}\right),$$

donde

$$Z_{[\mu]} = (1, 0), \qquad Z_{[\gamma]} = (1, 0, ..., 0),$$

$$T_{[\mu]} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad T_{[\gamma]} = \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix},$$

$$Q_{[\mu]} = \begin{bmatrix} \sigma_\xi^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix}, \qquad Q_{[\gamma]} = \sigma_w^2 (1, 0, ..., 0)'.$$

Este modelo expresado en (3.10) desempeña un papel importante en el enfoque del análisis estructural de series temporales, al que Harvey (1989) denominó *Modelo Estructural Básico* que en forma abreviada se denota por BSM por sus iniciales en ingles (Basic Structural Model). La forma de espacio de estados de este modelo básico, por ejemplo para $s = 4$ será:

$$\alpha_t = (\mu_t, \nu_t, \gamma_t, \gamma_{t-1}, \gamma_{t-2})', \quad Z_t = (1, 0, 1, 0, 0),$$

$$Q_t = \begin{bmatrix} \sigma_\xi^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_\zeta^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_w^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad T_t = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Alternativamente el componente estacional en su forma trigonométrica expresado en (3.8) y (3.9) puede ser incorporado en el BSM con un vector de estados de dimensión $(s + 1) \times 1$ expresado como

$$\alpha_t = (\mu_t, \nu_t, \gamma_{1t}, \gamma_{1t}^*, \gamma_{2t}, \gamma_{2t}^*, ..., \gamma_{s/2,t})',$$

y las partes relevantes del sistema de matrices estará dado por

$$Z_{[\gamma]} = (1,0,1,0,1,...,1,0,1), \quad T_{[\gamma]} = diag\,(C_1,...,C_{s^*},-1), \quad Q_{[\gamma]} = \sigma_w^2 I_{s-1},$$

cuando asumimos que s es par, se tiene $s^* = s/2$ donde

$$C_j = \begin{bmatrix} cos\lambda_j & sen\lambda_j \\ -sen\lambda_j & cos\lambda_j \end{bmatrix}, \qquad \lambda_j = \frac{2\pi j}{s}, \qquad j = 1,...,s^*, \tag{3.11}$$

y cuando s es impar, se tiene $s^* = (s-1)/2$ con

$$Z_{[\gamma]} = (1,0,1,0,1,...,1,0), \quad T_{[\gamma]} = diag\,(C_1,...,C_{s^*}), \quad Q_{[\gamma]} = \sigma_w^2 I_{s-1}.$$

donde C_j es definido en (3.11) para $j = 1,...,s^*$, I es la matriz identidad.

Para el desarrollo del presente trabajo, haremos uso del modelo estructural básico presentado en esta sección para construir el modelo subyacente de la tasa de desempleo y presentar en la forma de un modelo de espacio de estados con los componentes del modelo estructural que han sido desarrollados en detalle, tanto para la tendencia (3.6) como para la estacionalidad. Sin embargo, se ha visto que para la estacionalidad hay varias formas de expresar, nosotros usaremos la forma trigonométrica (3.8) debido a su amplio uso como en (Pfeffermann y Tiller, 2006) y simplicidad de expresar en la forma matricial.

3.3. El filtro de Kalman

El filtro de Kalman es un conjunto de ecuaciones de recursión, que de forma iterativa determina las estimaciones con error cuadrático medio mínimo del vector de estados α_t como el expresado en (3.1) y (3.2), basandose en la información disponible hasta el tiempo t. Es decir, bajo condiciones de normalidad el filtro de Kalman proporciona la media y la matriz de covarianzas del vector de estados, condicional a la información disponible hasta el momento t. Para esto reformulamos convenientemente el modelo de espacio de estados de la siguiente manera (ver Durbin and Koopman, 2012):

$$\begin{aligned} y_t &= Z_t\alpha_t + \varepsilon_t, & \varepsilon_t &\sim N(0,H_t), \\ \alpha_{t+1} &= T_t\alpha_t + \eta_t, & \eta_t &\sim N(0,Q_t), \qquad t = 1,2,3,...,n \\ & & \alpha_1 &\sim N(a_1,P_1), \end{aligned} \tag{3.12}$$

tomando como punto de partida $t = 1$ en (3.12), para el caso donde el estado inicial α_1 es $N(a_1,P_1)$ donde a_1 y P_1 son conocidos, en la práctica a_1 suele ser vector de ceros y P_1 una matriz diagonal de varianzas, asumiendo inicialmente varianzas altas. Bajo la condiciones de normalidad el objetivo es obtener las distribuciones condicionales de α_t y α_{t+1} dado Y_t como se muestra en las ecuaciones abajo, donde Y_t estará definido por el vector $(y_1',...,y_t')'$, mientras Y_0 indica que no hay observación antes de $t = 1$, también Y_{t-1} denota al conjunto

de observaciones pasadas $y_1, ..., y_{t-1}$ para $t = 2, 3, ...,$

$$\begin{aligned} a_{t|t} &= E(\alpha_t|Y_t), \qquad P_{t|t} = Var(\alpha_t|Y_t), \\ a_{t+1} &= E(\alpha_{t+1}|Y_t), \qquad P_{t+1} = Var(\alpha_{t+1}|Y_t). \end{aligned}$$

Como todas las distribuciones son normales, entonces las distribuciones condicionales de sub conjuntos de variables dado otros sub conjuntos de variables también son normales, esto como consecuencia del siguiente Lema.

Lema 1

Supongamos que X y Y son vectores aleatorios que tienen distribución conjunta normal con (Durbin and Koopman, 2012):

$$E\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \qquad Var\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma'_{XY} & \Sigma_{YY} \end{bmatrix},$$

donde Σ_{YY} se asume que es una matriz no singular. Entonces la distribución condicional de X dado Y es normal con vector de medias y matriz de varianzas siguientes:

$$\begin{aligned} E(X|Y) &= \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y), \\ Var(X|Y) &= \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma'_{XY}. \end{aligned}$$

Por lo tanto, las distribuciones de α_t dado Y_t y α_{t+1} dado Y_t están dados por $N(a_{t|t}, P_{t|t})$ y $N(a_{t+1}, P_{t+1})$ respectivamente. Procediendo inductivamente con $N(a_t, P_t)$ como distribución de α_t dado Y_{t-1}, mostramos en detalle como calcular recursivamente los parámetros $a_{t|t}$, a_{t+1}, $P_{t|t}$ y P_{t+1} de las distribuciones dadas, iniciando desde a_t y P_t para $t = 1, ..., n$.

3.3.1. Ecuaciones de actualización

Sea la siguiente ecuación dada por:

$$\upsilon_t = y_t - E(y_t|Y_{t-1}) = y_t - E(Z_t\alpha_t + \varepsilon_t|Y_{t-1}) = y_t - Z_t a_t,$$

donde υ_t es el error de predicción de y_t dado Y_{t-1}, es decir es el error de ir un punto hacia adelante en el tiempo o simplemente innovación. Cuando Y_{t-1} y υ_t son fijos entonces Y_t es fijo y vice versa. Así $E(\alpha_t|Y_t) = E(\alpha_t|Y_{t-1}, \upsilon_t)$. Pero $E(\upsilon_t|Y_{t-1}) = E(y_t - Z_t a_t|Y_{t-1}) = E(\alpha_t + \varepsilon_t - Z_t a_t|Y_{t-1}) = 0$. Consecuentemente $E(\upsilon_t) = 0$ y $Cov(y_j, \upsilon_t) = E[y_j E(\upsilon_t|Y_{t-1})'] = 0$ para $j = 1, ..., t-1$. También

$$\begin{aligned} a_{t|t} = E(\alpha_t|Y_t) = E(\alpha_t|Y_{t-1}, \upsilon_t), \\ a_{t+1} = E(\alpha_{t+1}|Y_t) = E(\alpha_{t+1}|Y_{t-1}, \upsilon_t), \end{aligned}$$

ahora aplicando el Lema 1 para la distribución conjunta condicional de α_t dado Y_{t-1}, tomando X y Y del Lema 1 como α_t y υ_t, aquí. Esto da como resultado

$$a_{t|t} = E(\alpha_t|Y_{t-1}) + Cov(\alpha_t, \upsilon_t)[Var(\upsilon_t)]^{-1}\upsilon_t, \tag{3.13}$$

donde Cov y Var se refiere a la covarianza y varianza en la distribución conjunta condicional de α_t y υ_t dado Y_{t-1}. Aquí $E(\alpha_t|Y_{t-1}) = a_t$ por definción de a_t y

$$\begin{aligned} Cov(\alpha_t, \upsilon_t) &= E[\alpha_t(Z_t\alpha_t + \varepsilon_t - Z_t a_t)'|Y_{t-1}] \\ &= E[\alpha_t(\alpha_t - a_t)'Z_t'|Y_{t-1}] = P_t Z_t', \end{aligned}$$

por definición de P_t. Ahora sea

$$F_t = Var(\upsilon_t|Y_{t-1}) = Var(Z_t\alpha_t + \varepsilon_t - Z_t a_t|Y_{t-1}) = Z_t P_t Z_t' + H_t,$$

entonces reemplazando en (3.13) se tiene

$$a_{t|t} = a_t + P_t Z_t' F_t^{-1}\upsilon_t, \tag{3.14}$$

Por el Lema 1 para la varianza se tiene

$$\begin{aligned} P_{t|t} = Var(\alpha_t|Y_t) &= Var(\alpha_t|Y_{t-1}, \upsilon_t) \\ &= Var(\alpha_t|Y_{t-1}) - Cov(\alpha_t, \upsilon_t)[Var(\upsilon_t)]^{-1}Cov(\alpha_t, \upsilon_t)' \\ &= P_t - P_t Z_t' F_t^{-1} Z_t P_t. \end{aligned} \tag{3.15}$$

asumiendo que F_t es una matriz no singular. Las relaciones expresadas en (3.14) y (3.15) son denominadas *paso de actualización o ecuaciones de actualización del Filtro de Kalman* (Durbin and Koopman, 2012).

3.3.2. Ecuaciones de predicción

Ahora desarrollamos las recursiones para calcular a_{t+1} y P_{t+1}. Basándonos de la relación existente $\alpha_{t+1} = T_t\alpha_t + \eta_t$, se tiene

$$\begin{aligned} a_{t+1} &= E(T_t\alpha_t + \eta_t|Y_t) \\ &= T_t E(\alpha_t|Y_t), \end{aligned} \tag{3.16}$$

$$\begin{aligned} P_{t+1} &= Var(T_t\alpha_t + \eta_t|Y_t) \\ &= T_t Var(\alpha_t|Y_t)T_t' + Q_t, \end{aligned} \tag{3.17}$$

para $t = 1, ..., n$. Sustituyendo la expresión (3.14) en (3.16) se tiene

$$\begin{aligned} a_{t+1} &= T_t a_{t|t} \\ &= T_t a_t + K_t \upsilon_t, \qquad t = 1, ..., n, \end{aligned} \tag{3.18}$$

donde

$$K_t = T_t P_t Z_t' F_t^{-1}, \tag{3.19}$$

La matriz K_t es conocido como la *Ganancia de Kalman* (Durbin and Koopman, 2012). Podemos observar que a_{t+1} ha sido obtenido como una función lineal de los valores previos a_t y el error de predicción υ_t de y_t dado Y_{t-1}.

Ahora para hallar P_{t+1} sustituyendo la expresiones (3.15) y (3.19) en (3.17) se obtiene

$$P_{t+1} = T_t P_t (T_t - K_t Z_t)' + Q_t, \quad t = 1, ..., n. \tag{3.20}$$

Así la las expresiones (3.18) y (3.20) son conocidos como *pasos de predicción* o *ecuaciones de predicción* del filtro de Kalman.

3.3.3. Recursión del filtro de Kalman

Para obtener el proceso de recursión completo, juntando convenientemente las ecuaciones de filtrado (Durbin and Koopman, 2012), así se tiene que:

$$\begin{aligned} \upsilon_t &= y_t - Z_t a_t, & F_t &= Z_t P_t Z_t' + H_t, \\ a_{t|t} &= a_t + P_t Z_t' F_t^{-1} \upsilon_t, & P_{t|t} &= P_t - P_t Z_t' F_t^{-1} Z_t P_t, \\ a_{t+1} &= T_t a_t + K_t \upsilon_t, & P_{t+1} &= T_t P_t (T_t - K_t Z_t)' + Q_t, \end{aligned} \tag{3.21}$$

para $t = 1, ..., n$ donde $K_t = T_t P_t Z_t' F_t^{-1}$ con a_1 y P_1 como vector de medias y matriz de varianzas en el vector de estados inicial α_1. La recursión expresada en el sistema de ecuaciones (3.21) se denomina *Filtro de Kalman*. Mediante este procedimiento es posible actualizar nuestro conocimiento del sistema cada vez que una nueva observación ingresa en el sistema, es decir, el sistema nos permite calcular los valores para $t+1$ dada las observaciones hasta t. En otras palabras, el filtro de Kalman permite que la estimación del vector de estados se actualice continuamente a medida que hay nueva información disponible. El procedimiento recursivo sería como sigue, asumiendo que los parámetros del sistema (3.21) son conocidos:

- Dado la información disponible Y_{t-1} se predice el vector de estados $a_t = E(\alpha_t | Y_{t-1})$ cuya matriz de covarianza de error de predicción es $P_t = E[\alpha_t(\alpha_t - a_t)' | Y_{t-1}]$, entonces al predecir α_t con información hasta el tiempo $t-1$ se tiene la nueva observación y_t con error de predicción F_t.

- La nueva información y_t obtenida en el paso anterior se ingresa al sistema recursivo de donde se obtienen $a_{t|t}$ y la matriz de covarianza $P_{t|t}$ en las ecuaciones de actualización entonces se actualiza el vector de estados α_t.

- Una vez actualizado el sistema (3.21) se tiene $a_{t|t}$ y $P_{t|t}$, y reemplazando en las ecuaciones de predicción; $a_{t+1} = T_t a_{t|t}$ y $P_{t+1} = T_t P_{t|t} T_t' + Q_t$ se predice el próximo vector de estados α_{t+1}, haciendo esto de forma recursiva nuevamente se predice la nueva observación.

3.4. Estimación por máxima verosimilitud

En las secciones previas hemos desarrollado métodos para estimar los componentes que se encuentran en el vector de estados expresado en (3.1) y (3.2). En la práctica los modelos los dependen de parámetros adicionales, los cuales deben ser estimados de los datos. Por ejemplo, en las expresiones (3.4) y (3.5) los parámetros referidos a las varianzas σ_ε^2 y σ_η^2 son desconocidos y es necesario que sean estimados. Los estimadores de estos parámetros son hallados maximizando la función de verosimilitud $L(Y_n; \Psi)$ con respecto a Ψ que contiene los parámetros desconocidos.

Asumiendo que el estado inicial del vector de estados tiene distribución $N(a_1, P_1)$, donde a_1 y P_1 son conocidos, entonces la función de verosimilitud queda expresado como:

$$L(Y_n; \sigma_\varepsilon^2, \sigma_\xi^2, \sigma_\zeta^2, \sigma_w^2) = L(Y_n) = p(y_1, ..., y_n) = p(y_1) \prod_{t=2}^{n} p(y_t|Y_{t-1}),$$

donde $Y_t = (y_1', ..., y_t')'$, $p(y_1, ..., y_n)$ es la función de densidad de probabilidad conjunta y $p(y_t|Y_{t-1})$ denota la distribución condicional de y_t dada la información hasta el momento $t-1$. En la práctica generalmente se trabaja con el logaritmo de la función de verosimilitud el cual sería:

$$log[L(Y_n)] = \sum_{t=1}^{n} log[p(y_t|Y_{t-1})]. \tag{3.22}$$

Para el modelo estructural básico expresado en (3.10) y sus componentes detallados, por ejemplo, los parámetros a estimar mediante máxima verosimilitud serían (σ_ε^2, σ_ξ^2, σ_ζ^2 y σ_w^2).

Con la finalidad de hallar los parámetros indicados, haremos uso de la factorización de la distribución conjunta de las observaciones, aplicado en la función de densidad conjunta $f(y_t, ..., y_1)$:

$$\begin{aligned} p(y_t, ..., y_1) &= p(y_t|y_{t-1}, ..., y_1)p(y_{t-1}, ..., y_1) = p(y_t|y_{t-1}, ..., y_1)p(y_{t-1}|y_{t-2}, ..., y_1)p(y_{t-2}, ..., y_1) \\ &= p(y_t|y_{t-1}, ..., y_1)p(y_{t-1}|y_{t-2}, ..., y_1)p(y_{t-2}|y_{t-3}, ..., y_1)p(y_{t-3}, ..., y_1)...p(y_2|y_1)p(y_1) \\ &= \prod_{i=0}^{t-2} p(y_{t-i}|y_{t-i-1}). \end{aligned}$$

Entonces $y_t|y_{t-1}, y_{t-2}, ..., y_1 \sim N(E(y_t|Y_{t-1}), Var(y_t|Y_{t-1}))$, donde $p(y_1|Y_0) = p(y_1)$. De las ecuaciones de (3.1) y (3.2), se tiene $E(y_t|Y_{t-1}) = Z_t a_t$. Poniendo $\upsilon_t = y_t - Z_t a_t$, $F_t = Var(y_t|Y_{t-1})$ y sustituyendo $N(Z_t a_t, F_t)$ por $p(y_t|Y_{t-1})$ en la expresión (3.22), se obtiene

$$log[L(Y_n)] = -\frac{np}{2} log(2\pi) - \frac{1}{2} \sum_{t=1}^{n} (log|F_t| + \upsilon_t' F_t^{-1} \upsilon_t). \tag{3.23}$$

donde p viene a ser el número de series de tiempo, para el caso univariante $p = 1$. Las cantidades v_t y F_t son calculados rutinariamente por el filtro de Kalman expresado en (3.21), así $log[L(Y_n)]$ de (3.23) es calculado fácilmente a partir de la salida del filtro de Kalman. También se asume que F_t es no singular para $t = 1, ..., n$, si esta condición no se cumple inicialmente es usual redefinir el modelo para que se cumpla (Durbin and Koopman, 2012).

3.5. Modelo de espacio de estados con errores correlacionados

Hasta aquí hemos visto la construcción de modelos de espacio de estados basado en un modelo estructural básico de una serie de tiempo con errores independientes y como se estiman sus parámetros usando el filtro de Kalman, sin embargo, ¿Qué pasa cuando los errores de una serie de tiempo no son independientes?, es decir, cuando los errores están correlacionados, esto generalmente se presenta cuando la misma unidad de muestreo se usa repetitivamente para recoger información.

Por ejemplo, para las encuestas de empleo y desempleo con muestra tipo panel una unidad de vivienda de la muestra puede participar en la toma de información en 2 o más veces al año, entonces es aquí donde se presentan las series con errores correlacionados, dado que el error de muestreo se repite. Para estos casos en el 2006 Pfeffermann y Tiller plantearon ecuaciones como las expresadas en (3.1) y (3.2) pero considerando los errores correlacionados, al que se le denomina modelo de espacio de estados con errores correlacionados. Las ecuaciones planteadas son:

$$\begin{aligned} &y_t = Z_t\alpha_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \\ &E(\varepsilon_t) = 0, \quad E(\varepsilon_t\varepsilon_t') = \Sigma_{tt}, \quad E(\varepsilon_\tau\varepsilon_t') = \Sigma_{\tau t}, \\ &\alpha_t = T_t\alpha_{t-1} + \eta_t, \quad \eta_t \sim N(0, Q_t), \\ &E(\eta_t) = 0, \quad E(\eta_t\eta_t') = Q_t, \quad E(\eta_t\eta_{t-k}') = 0, \quad k > 0. \end{aligned} \tag{3.24}$$

Al indicar que $E(\varepsilon_\tau\varepsilon_t') = \Sigma_{\tau t}$ el modelo asume que los errores están correlacionados, en este escenario ya no es posible aplicar el filtro de Kalman para la estimación de los parámetros del vector de estados y es necesario buscar una alternativa que considere la correlación de los errores de medición. Pfeffermann y Tiller (2006) desarrollaron un algoritmo de filtro recursivo para modelos de espacio de estados con errores de medición autocorrelacionados, el algoritmo actualiza el predictor más reciente del vector de estados cada vez que ingresa al sistema una nueva información. El algoritmo puede ser aplicado en forma general a los modelos de espacio de estados que consideren tener errores correlacionados, y dado que el objetivo del estudio es desarrollar este tipo de modelos, presentaremos a continuación en detalle el algoritmo recursivo desarrollado por Pfeffermann y Tiller (2006).

3.6. Algoritmo de filtro recursivo de Pfeffermann y Tiller

Como ya hemos mencionado en la sección anterior para el tratamiento de modelos de espacio de estados con errores correlacionados, como los expresados en (3.24), Pfeffermann y Tiller desarrollaron este algoritmo en 2006, dado que el filtro de Kalman no era aplicable al

no considerar los errores correlacionados. A continuación presentamos en detalle el algoritmo:

Para tiempo $t = 1$

Sea $\hat{\alpha}_1 = (I - K_1 Z_1)T\hat{\alpha}_0 + K_1 y_1$ el estimador filtrado en el tiempo $t = 1$, donde $\hat{\alpha}_0$ es un estimador inicial con matriz de covarianzas $P_0 = E[(\hat{\alpha}_0 - \alpha_0)(\hat{\alpha}_0 - \alpha_0)']$ y con $K_1 = P_{1|0}Z_1'F_1^{-1}$ denominado *"ganancia de Kalman"*. Se asume por conveniencia que $\hat{\alpha}_0$ es independiente de las observaciones. La matriz $P_{1|0} = TP_0T' + Q$ es la matriz de covarianzas de los errores de predicción $\hat{\alpha}_{1|0} - \alpha_1 = T\hat{\alpha}_0 - \alpha_1$, y $F_1 = Z_1 P_{1|0} Z_1' + \Sigma_{11}$ es la matriz de covarianzas de las innovaciones (errores de predicción de un paso hacia adelante) $\nu_1 = y_1 - \hat{y}_{1|0} = y_1 - Z_1\hat{\alpha}_{1|0}$. Por que $y_1 = Z_1\alpha_1 + \varepsilon_1$, entonces reemplazando se tiene:

$$\hat{\alpha}_1 = (I - K_1 Z_1)T\hat{\alpha}_0 + K_1 Z_1 \alpha_1 + K_1\varepsilon_1. \tag{3.25}$$

Para tiempo $t = 2$

Sea $\hat{\alpha}_{2|1} = T\hat{\alpha}_1$ que define al predictor de α_2 en el tiempo $t = 1$ con matriz de covarianzas $P_{2|1} = E[(\hat{\alpha}_{2|1} - \alpha_2)(\hat{\alpha}_{2|1} - \alpha_2)']$. Un predictor insesgado $\hat{\alpha}_2$ de α_2 [es decir, $E(\hat{\alpha}_2 - \alpha_2) = 0$] basado en $\hat{\alpha}_{2|1}$ y la observación y_2 es el predictor de mínimos cuadrados generalizado (MCG) en el modelo de regresión:

$$\begin{pmatrix} T\hat{\alpha}_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} I \\ Z_2 \end{pmatrix}\alpha_2 + \begin{pmatrix} u_{2|1} \\ \varepsilon_2 \end{pmatrix}, \quad u_{2|1} = T\hat{\alpha}_1 - \alpha_2,$$

de donde

$$\hat{\alpha}_2 = \left[(I, Z_2')V_2^{-1}\begin{pmatrix} I \\ Z_2 \end{pmatrix}\right]^{-1}(I, Z_2')V_2^{-1}\begin{pmatrix} T\hat{\alpha}_1 \\ y_2 \end{pmatrix},$$

aquí

$$V_2 = var\begin{pmatrix} u_{2|1} \\ \varepsilon_2 \end{pmatrix} = \begin{pmatrix} P_{2|1} & C_2 \\ C_2' & \Sigma_{22} \end{pmatrix},$$

y $C_2 = cov[T\hat{\alpha}_1 - \alpha_2, \varepsilon_2] = TK_1\Sigma_{12}$ (sigue de (3.24) y (3.25)). Notar que V_2 es la matriz de covarianzas de los errores $u_{2|1}$ y ε_2, y no de los predictores $T\hat{\alpha}_1$ y y_2. El predictor MCG $\hat{\alpha}_2$ es el mejor predictor lineal insesgado (BLUP Best Linear Unbiased Predictor, por sus iniciales en Ingles cuya prueba se puede ver en el Apéndice A del artículo Pfeffermann y Tiller, 2006) de α_2 basado en $T\hat{\alpha}_1$ y y_2, con matriz de covarianza

$$E[(\hat{\alpha}_2 - \alpha_2)(\hat{\alpha}_2 - \alpha_2)'] = \left[(I, Z_2')V_2^{-1}\begin{pmatrix} I \\ Z_2 \end{pmatrix}\right]^{-1} = P_2.$$

Para el tiempo t

Sea $\hat{\alpha}_{t|t-1} = T\hat{\alpha}_{t-1}$ que define al predictor de α_t en tiempo $t-1$ con matriz de covarianzas $E[(\hat{\alpha}_{t|t-1} - \alpha_t)(\hat{\alpha}_{t|t-1} - \alpha_t)'] = TP_{t-1}T' + Q = P_{t|t-1}$, donde $P_{t-1} = E[(\hat{\alpha}_{t-1} - \alpha_{t-1})(\hat{\alpha}_{t-1} -$

$\alpha_{t-1})']$. Estableciendo el modelo de regresión de coeficientes aleatorios como:

$$\begin{pmatrix} T\hat{\alpha}_{t-1} \\ y_t \end{pmatrix} = \begin{pmatrix} I \\ Z_t \end{pmatrix}\alpha_t + \begin{pmatrix} u_{t|t-1} \\ \varepsilon_t \end{pmatrix}, \qquad u_{t|t-1} = T\hat{\alpha}_{t-1} - \alpha_t, \tag{3.26}$$

y definiendo la varianza como

$$V_t = var\begin{pmatrix} u_{t|t-1} \\ \varepsilon_t \end{pmatrix} = \begin{pmatrix} P_{t|t-1} & C_t \\ C_t' & \Sigma_{tt} \end{pmatrix}. \tag{3.27}$$

La matriz de covarianzas $C_t = cov[T\hat{\alpha}_{t-1} - \alpha_t, \varepsilon_t]$ es calculado como sigue. Sea $[I, Z_j']V_j^{-1} = [B_{j1}, B_{j2}]$ donde B_{j1} contiene las primeras q columnas de $[I, Z_j']V_j^{-1}$ y B_{j2} contiene el resto de las columnas, con $q = dim(\alpha_j)$. Definiendo $A_j = TP_jB_{j1}$, $\tilde{A}_j = TP_jB_{j2}$, $j = 2, ..., t-1$, $\tilde{A}_1 = TK_1$. Entonces

$$\begin{aligned} C_t &= cov[T\hat{\alpha}_{t-1} - \alpha_t, \varepsilon_t] \\ &= A_{t-1}A_{t-2}...A_2\tilde{A}_1\Sigma_{1t} + A_{t-1}A_{t-2}...A_3\tilde{A}_2\Sigma_{2t} \\ &\quad +... + A_{t-1}\tilde{A}_{t-2}\Sigma_{t-2,t} + \tilde{A}_{t-1}\Sigma_{t-1,t}. \end{aligned}$$

El predictor MCG (Mínimos Cuadrados Generalizados) de α_t basado en $T\hat{\alpha}_{t-1}$ y y_t, y la matriz de covarianzas de los errores de predicción son obtenidos de (3.26) y (3.27) como:

$$\begin{aligned} \hat{\alpha}_t &= \left[(I, Z_t')V_t^{-1}\begin{pmatrix} I \\ Z_t \end{pmatrix}\right]^{-1}(I, Z_t')V_t^{-1}\begin{pmatrix} T\hat{\alpha}_{t-1} \\ y_t \end{pmatrix}, \\ P_t &= E[(\hat{\alpha}_t - \alpha_t)(\hat{\alpha}_t - \alpha_t)'] = \left[(I, Z_t')V_t^{-1}\begin{pmatrix} I \\ Z_t \end{pmatrix}\right]^{-1}. \end{aligned} \tag{3.28}$$

3.7. Ejemplo práctico

En esta sección, con base a la teoría presentada en las secciones anteriores se desarrolla un ejemplo práctico con la finalidad de mostrar la construcción del modelo estructural básico, llevar a la forma de un modelo de espacio de estados con la finalidad de poder aplicar el algoritmo del filtro de Kalman y el algoritmo de Pfeffermann y Tiller para estimar los componentes del modelo estructural y explicar cuando se usa cada algoritmo.

Para esto, poniendo en el contexto práctico usaremos los datos la tasa de desempleo mensual de Lima Metropolitana. Sea y_t la serie de observaciones de la tasa de desempleo mensual con $n = 211$, primero veamos como expresar la serie como un modelo estructural básico, como el definido en (3.10), con sus componentes de la tendencia y los efectos de estacionalidad definidos en (3.6) y (3.8) respectivamente. Entonces tomando en cuenta cada

componente se tiene que:

$$
\begin{aligned}
&y_t = \mu_t + \gamma_t + \varepsilon_t; \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \\
&\mu_t = \mu_{t-1} + \nu_{t-1} + \xi_t; \quad \xi_t \sim N(0, \sigma_\xi^2), \\
&\nu_t = \nu_{t-1} + \zeta_t; \quad \zeta_t \sim N(0, \sigma_\zeta^2), \\
&\gamma_t = \sum_{j=1}^{6} \gamma_{j,t} \\
&\gamma_{j,t} = \gamma_{j,t-1} \cos \lambda_j + \gamma_{j,t-1}^* \operatorname{sen} \lambda_j + w_{j,t}; \quad w_{j,t} \sim N(0, \sigma_w^2), \\
&\gamma_{j,t}^* = -\gamma_{j,t-1} \operatorname{sen} \lambda_j + \gamma_{j,t-1}^* \cos \lambda_j + w_{j,t}^*; \quad w_{j,t}^* \sim N(0, \sigma_w^2), \\
&\lambda_j = \frac{2\pi j}{12}, \quad j = 1, ..., 6,
\end{aligned}
\tag{3.29}
$$

donde los términos ε_t, ξ_t, ζ_t, $w_{j,t}$, y $w_{j,t}^*$ son series de ruidos blancos independientes y los componentes μ_t, ν_t, γ_t, y ε_t definen el nivel de la tendencia en el tiempo, la pendiente de la tendencia, los efectos estacionales ($s = 12$ dada las observaciones mensuales), y el término irregular de la serie respectivamente, operando en el tiempo t. Como se puede ver la ecuación (3.29) está compuesto por las ecuaciones de cada componente, así la tendencia μ_t está en función de la misma tendencia μ_{t-1} y la pendiente ν_{t-1} de un periodo antes más un ruido blanco ξ_t, de forma similar, la pendiente ν_t depende de la misma pendiente un periodo antes ν_{t-1} más el ruido blanco ζ_t, mientras que para obtener los efectos de estacionalidad γ_t usamos la forma trigonométrica expresado en (3.8), considerando que los datos son mensuales $s = 12$.

Una vez que la serie y_t es formulada en los componentes de un modelo estructural básico, es posible llevarlo a la forma de un modelo de espacio de estados como fue definido en (3.1) y (3.2), donde $H_t = \sigma_\varepsilon^2$. Expresado así, ya es posible aplicar los algoritmos del filtro de Kalman y de Pfeffermann y Tiller para realizar las estimaciones, pero antes es necesario construir los vectores y matrices del modelo de espacio de estados.

Entonces el vector de estados α_t estará compuesto por elementos del modelo estructural básico como se ve a continuación:

$$
\alpha_t = (\mu_t, \nu_t, \gamma_{1t}, \gamma_{1t}^*, \gamma_{2t}, \gamma_{2t}^*, \gamma_{3t}, \gamma_{3t}^*, \gamma_{4t}, \gamma_{4t}^*, \gamma_{5t}, \gamma_{5t}^*, \gamma_{6t})'_{13\times 1},
$$

mientras Z_t será un vector fila de (1×13) constante y la matriz T_t, que también es constante, tendrá una dimensión de (13×13). Estos quedan definidos de la forma siguiente:

$$
Z_t = (1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1)_{1\times 13}\,,
$$

$$T_t = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & cos(\pi/6) & sen(\pi/6) & \cdots & 0 \\ 0 & 0 & -sen(\pi/6) & cos(\pi/6) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 \end{bmatrix}_{13\times 13},$$

en tanto que la matriz de varianzas y covarianzas Q_t es una matriz diagonal de (13×13) cuya diagonal principal está compuesto por los parámetros de la ecuación de la tendencia, de la pendiente y de la estacionalidad, y queda expresado de la forma:

$$Q_t = \begin{bmatrix} \sigma_\xi^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_\zeta^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_w^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_w^2 \end{bmatrix}_{13\times 13},$$

donde los parámetros σ_ξ^2, σ_ζ^2 y σ_w^2 en la práctica no son conocidos y se estiman previamente usando las observaciones de la serie. En este ejemplo sin embargo, supondremos que estos valores son conocidos y dados por $\sigma_\xi^2 = 0.0024$, $\sigma_\zeta^2 = 0.0004$ y $\sigma_w^2 = 0.0000001$.

Una vez construido los elementos del sistema de ecuaciones del modelo de espacio de estados, queda predispuesto para aplicar los algoritmos y estimar los componentes del vector de estados. Primero estimaremos el vector de estados α_t y su respectiva matriz de covarianzas P_t utilizando el filtro de Kalman expresado en el sistema de ecuaciones (3.21), para iniciar el cálculo se supone valores iniciales conocidos, entonces α_0 será un vector de ceros y P_t una matriz diagonal de (13×13) con varianzas igual a 10,000 (valores altos para ver como disminuyen).

Con las estimaciones del vector de estados obtenida con el filtro de Kalman realizamos el gráfico que se muestra en la Figura (3.2). En ella se puede visualizar las observaciones de la tasa de desempleo, también la tendencia, así como la suma de la tendencia más los efectos estacionales, de donde también se puede obtener el componente irregular de la serie, desagregación que ofrece muchas ventajas para el análisis de una serie de tiempo, como por ejemplo, focalizar el análisis sólo en la tendencia o en los otros componentes.

Para completar el ejemplo, explicamos como se realizan las estimaciones cuando el modelo de espacio de estados es para una serie de tiempo con errores correlacionados como los expresados en el modelo (3.24), donde $E(\varepsilon_\tau \varepsilon_t') = \Sigma_{\tau t}$ indica que los errores están correlacionados. En estos casos es preciso utilizar el algoritmo de Pfeffermann y Tiller dado que supone que existen correlaciones entre los errores de muestreo, conforme fue detallado en la sección (3.6). Para el ejemplo supondremos que existen ciertas correlaciones, a fin de mostrar la aplicación del algoritmo de Pfeffermann y Tiller. En la práctica las correlaciones se determinan en función a las observaciones y el diseño muestral utilizado para la recolección

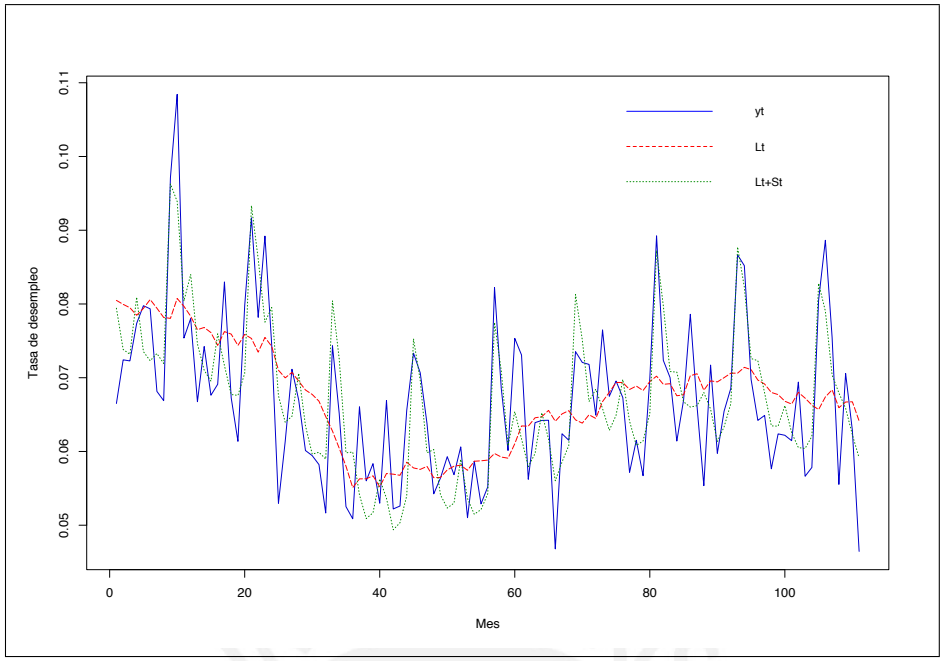


Figura 3.2: Tasa de desempleo mensual Lima Metropolitana (2002-2018) cuyos componentes fueron estimados mediante el *Filtro de Kalman*, donde L_t es el nivel de tendencia y S_t son los efectos estacionales.

de las observaciones.

Para aplicar el algoritmo de Pfeffermann y Tiller al modelo (3.24), las matrices y vectores del sistema se construyen tal cual fue construido para aplicar el filtro de Kalman, sin embargo, para estimar el vector de estados α_t y la matriz de covarianzas P_t definido en (3.28), es necesario construir la matriz V_t que considera los errores correlacionados, entonces veamos la construcción:

$$V_t = \begin{pmatrix} P_{t|t-1} & C_t \\ C_t' & \Sigma_{tt} \end{pmatrix}_{14\times 14},$$

donde los componentes $P_{t|t-1} = TP_{t-1}T' + Q_t$ y $\Sigma_{tt} = \sigma_\varepsilon^2$ son conocidos a partir de lo ya definido para el filtro de Kalman, pero C_t depende de las correlaciones de los errores que la serie y_t tenga y como dijimos eso se determina en función a las observaciones originales. Para el ejemplo supongamos que los primeros 3 errores están correlaciones, $corr(\varepsilon_t, \varepsilon_{t-1}) = 0.23$, $corr(\varepsilon_t, \varepsilon_{t-2}) = 0.17$ y $corr(\varepsilon_t, \varepsilon_{t-3}) = 0.11$, entonces C_t se plantea como una combinación lineal, de la forma siguiente:

$$C_t = A_{t-1}A_{t-2}\tilde{A}_{t-3}\Sigma_{t-3,t} + A_{t-1}\tilde{A}_{t-2}\Sigma_{t-2,t} + \tilde{A}_{t-1}\Sigma_{t-1,t},$$

donde A y $\tilde{A}$ se define en función de la matriz $[I, Z_j']V_j^{-1} = [B_{j1}, B_{j2}]$ con dimensión 13×14, B_{j1} serán las primeras 13 columnas y B_{j2} será la columna 14. Así, $A_j = TP_jB_{j1}$ y $\tilde{A}_j = TP_jB_{j2}$, donde $I_{13\times 13}$ es matriz identidad y $j = 2, ..., t-1$.

Con lo definido y aplicando el algoritmo de Pferffermann y Tiller se pueden estimar el vector de estados α_t definido en (3.28) y con los componentes estimados del vector de estados realizar el gráfico de la Figura (3.3). Este es muy similar al gráfico (3.2), debido a que las

correlaciones de los errores que hemos supuesto no corresponden a la serie de la tasa de desempleo, sin embargo, podemos indicar que las correlaciones se pueden llegar a confundir con los efectos de estacionalidad.

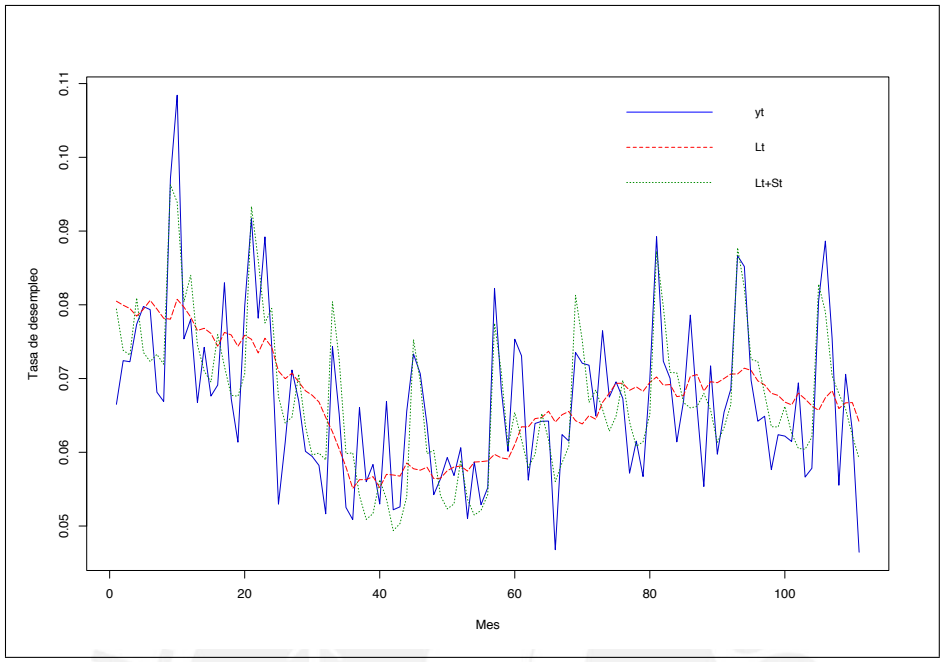


Figura 3.3: Tasa de desempleo mensual Lima Metropolitana (2002-2018) cuyos componentes fueron estimados mediante el *algoritmo de Pfeffermann y Tiller*, donde L_t es el nivel de tendencia y S_t son los efectos estacionales.

Para conocer el orden de los errores de muestreo correlacionados se necesita conocer a profundidad el diseño muestral utilizado para la recolección de los datos de la tasa de desempleo así como las características de los datos, con esa finalidad en el siguiente capítulo se verán en detalle estos temas.

Capítulo 4

Encuesta Permanente del Empleo - EPE

En el presente capítulo el objetivo es conocer en detalle el diseño longitudinal de la Encuesta Permanente del Empleo - EPE, en especial su diseño muestral, las variaciones y modificaciones que ha sufrido el diseño a lo largo del tiempo, así como otras características de la encuesta dado que nos servirá de insumo para estudiar las características y modelar la serie generada por las observaciones de la tasa de desempleo, el cual forma parte de la EPE y es uno de los indicadores más relevantes del mercado laboral.

La EPE es ejecutada por el Instituto Nacional de Estadística e Informática - INEI desde inicios del año 2001 ante la necesidad de brindar información estadística sobre las principales características del mercado laboral, dado que por razones presupuestales se dejo de ejecutar la encuesta nacional de hogares en forma trimestral. Ante esto, el INEI en coordinación con el Ministerio de Trabajo y Promoción del Empleo, y el Ministerio de Economía y Finanzas diseñaron la EPE con la finalidad de brindar información estadística mensual que sirva como insumo para el análisis del panorama laboral local.

Entre las principales indicadores del mercado laboral se encuentra la tasa de desempleo, dado que refleja la capacidad de la economía de un país para absorber el desempleo, este indicador se estima mensualmente con data acumulada de los tres últimos meses de la EPE válido sólo para Lima Metropolitana, esta información servirá de insumo para el modelado y el análisis de la serie formada desde sus inicios hasta la fecha.

4.1. Antecedentes

- Según los documentos revisados del INEI, la EPE se inicia en el mes de marzo del 2001 sobre la base de una muestra anual. Esta muestra estaba conformada por 150 conglomerados elegidos aleatoriamente del marco del Pre-censo 1999-2000. En cada conglomerado se seleccionaron 3 sub-muestras, cada una compuesta de 11 viviendas contiguas al que se denominará más adelante como grupo compacto, de estas aleatoriamente 1 fue entrevistada por mes hasta completar las 3 sub-muestras en el primer trimestre (marzo, abril y mayo). Las mismas 3 sub-muestras de viviendas, fueron nuevamente visitadas en los trimestres siguientes del año, es decir, de un trimestre a otro la superposición de la muestra de viviendas al 100 por ciento.

  Cada muestra mensual estaba conformado de 1,650 viviendas, completando las 4,950

viviendas en cada trimestre acumulado. En febrero del 2002, concluyó la primera ronda anual de la encuesta, en este periodo cada vivienda de la muestra fue visitada en 4 oportunidades.

- Para la segunda ronda de la encuesta, iniciada a partir de marzo del 2002 se diseño una muestra maestra, la cual además de asegurar la obtención de indicadores del empleo y desempleo, debía tener una vigencia no menor a 2 años. Esta muestra maestra estaba conformada por 600 conglomerados elegidos aleatoriamente, además, en cada conglomerado se seleccionaron 4 sub-muestras de 8 viviendas cada una.

  Cada muestra mensual estaba conformado por 1,600 viviendas que estaban en 200 conglomerados, completando las 4,800 viviendas por trimestre en los 600 conglomerados, tamaño suficiente de muestra según su diseño para la estimación de indicadores del mercado laboral mediante promedios móviles. También se tuvo en cuenta suavizar el empalme de las muestras (de un diseño a otro) y así evitar que las estimaciones de la encuesta de vieran afectadas por el cambio de muestra.

- Para el 2004, se diseño una muestra maestra de 1,200 conglomerados, de similares características pero independiente de las anteriores mencionadas. En cada conglomerado de la muestra se seleccionaron aleatoriamente 4 sub-muestras de 4 viviendas cada una. Cada muestra mensual estaba conformado por 1,600 viviendas (400 conglomerados) completando las 4,800 viviendas por trimestre en los 1,200 conglomerados. Con la finalidad de suavizar el empalme entre muestras, durante los meses de marzo, abril y mayo del 2004 se efectuó este proceso de tal manera que desde junio del mismo año la muestra se había renovado totalmente.

- Para el 2006, se realizó la Encuesta Continua (ENCO) en todo el territorio nacional, esta encuesta de propósitos múltiples, se basó en el marco muestral del censo nacional 2005 (X de población y V de vivienda). Dentro de la ENCO, el ámbito de Lima Metropolitana constituyó un dominio de estudio independiente al resto de los dominios considerados. La muestra para este dominio fue de 1,600 viviendas mensuales (200 conglomerados) completando las 4,800 viviendas por trimestre (600 conglomerados).

- Para el año 2007, se redujo el presupuesto de la ENCO Lima Metropolitana, que nuevamente se denominó EPE, lo que implicó una reducción de la muestra trimestral de 4,800 viviendas, como en anteriores ediciones, a 3,000 viviendas. La muestra para el año 2010 es la misma muestra de conglomerados del 2007.

- La muestra maestra de conglomerados y viviendas para el periodo 2011 - 2016 tuvo su termino en diciembre del 2016, mes donde la última sub-muestra de viviendas programada fue visitada.

- Para el periodo 2017 - 2020, se ha seleccionado una muestra maestra de conglomerados de similares características a diseños anteriores. el marco muestral se basa en información estadística y cartográfica del censo de población y vivienda del 2007 actualizada con el empadronamiento distrital de población y vivienda del 2013.

Mediante el plan de rotación y empalme, la muestra maestra vigente fue renovada durante el 2017, de tal manera que las estimaciones de la encuesta no perderán la comparabilidad temporal y confiabilidad estadística.

4.2. Finalidad

Suministrar información estadística mensual de seguimiento del mercado laboral con datos agregados del último trimestre, a través de indicadores de empleo, desempleo e ingresos y otros referentes a la disponibilidad y utilización de los recursos humanos en el Área Metropolitana de Lima y Callao (Instituto Nacional de Estadística e Informática [INEI], 2020).

4.3. Objetivos

Los objetivos planteados por el INEI (2020), son los siguientes:

- Generar indicadores de empleo, desempleo e ingresos en el Área Metropolitana de Lima y Callao.
- Desarrollar indicadores anticipatorios de la evolución del empleo, para fines prospectivos.
- Servir de fuente de información a instituciones públicas y privadas, así como a investigadores.
- Permitir la comparabilidad con investigaciones similares, en relación con las variables investigadas.

4.4. Población objetivo

"La población bajo estudio está constituido por el conjunto de viviendas particulares y sus ocupantes con residencia habitual ubicadas en el Área Metropolitana de Lima y Callao. Se excluye del estudio a los miembros de las fuerzas armadas que viven en cuarteles, campamentos, barcos, etc. Además, se excluye también a las viviendas colectivas (hoteles, hospitales, asilos, claustros religiosos, cárceles etc.)" (INEI, 2020).

4.5. Cobertura

"La encuesta se realiza en el área Metropolitana de Lima y Callao, constituida por 43 distritos de la Provincia de Lima y 6 distritos de la Provincia Constitucional del Callao" (INEI, 2020).

4.6. Características del diseño muestral

4.6.1. Marco muestral

En adelante nos referimos al diseño muestral la que está actualmente en vigencia, cuyo periodo es 2017 - 2022, este marco muestral tiene como fuente principal a la información estadística y cartográfica de los Censos Nacionales del 2007: XI de Población y VI de Vivienda,

el cual fue actualizado con el empadronamiento distrital de Población y Vivienda del año 2013 del Sistema de Focalización de Hogares (SISFOH)[1].

4.6.2. Estratificación implícita

La estratificación implícita en un procedimiento sencillo (a diferencia de la estratificación clásica, elimina la necesidad de establecer estratos explícitos, suprimiendo así, la necesidad de asignar muestra a dichos estratos), que implica organizar de forma adecuada las Unidades Primarias de Muestreo - UPM's entre sub-grupos importantes de la población, como son; el orden geográfico, socioeconómico, etc. para en seguida seleccionar la muestra de forma sistémica con igual probabilidad o con probabilidad proporcional al tamaño (Naciones Unidas [NU], 2008).

Para la EPE previa a la selección de la muestra, se ordenaron los conglomerados por estrato socioeconómico y en serpentín distrital de norte a sur, obteniendo así una estratificación implícita del marco de donde la muestra fue seleccionada sistemáticamente con probabilidad proporcional al tamaño.

4.6.3. Unidades de muestreo

Las unidades de muestreo definidos para la EPE por el INEI (2020) son:

- **Unidad Primaria de Muestreo (UPM)**.- La unidad primaria de muestreo es el conglomerado, que se define como el área geográfica conformada por una o más manzanas contiguas, en promedio cada conglomerado tiene 140 viviendas particulares.
- **Unidad Secundaria de Muestreo (USM)**.- La unidad secundaría es la vivienda particular que existe dentro de una UPM. En las viviendas que finalmente resulten seleccionadas se procede a investigar a todas las personas que tienen su residencia habitual en ella.

4.6.4. Tipo de muestreo

La estrategia de muestreo utilizado por la EPE es de tipo probabilística, estratificada y bietápica.

"La muestra es probabilística, porque las unidades son seleccionadas aleatoriamente, lo cual permite efectuar inferencias a la población objetivo en base a la teoría de probabilidades" (INEI, 2020).

La muestra es estratificada, porque previamente a la selección, los conglomerados son ordenados por estrato socioeconómico y los distritos de norte a sur, con la finalidad de mejorar la representatividad y la eficiencia del diseño. La estratificación implícita exige el uso de la selección sistémica en la primera etapa de muestreo (NU, 2008).

[1]SISFOH es un sistema intersectorial e intergubernamental que provee información socioeconómica a las Intervenciones Públicas Focalizadas, actualmente se encuentra bajo la dirección del Ministerio de Desarrollo e Inclusión Social - MIDIS (http://www.sisfoh.gob.pe/)

Bietápica, en la primera etapa del muestreo se seleccionan los conglomerados utilizando el método sistémico con probabilidad proporcional al tamaño (PPT) de viviendas en los conglomerados. En la segunda etapa, se seleccionaron las viviendas, utilizando el método sistémico simple con arranque aleatorio (INEI, 2020). En cada conglomerado de la muestra se seleccionaron 5 sub-muestras (grupo compacto o contigua de viviendas), cada una compuesta por 4 viviendas contiguas geográficamente.

4.6.5. Tamaño de la muestra

De forma similar que en los diseños anteriores, se ha seleccionado una muestra maestra de 2,400 conglomerados, esta muestra se espera que tenga un periodo de vida de al menos 4 años. Los 2,400 conglomerados son distribuidos de manera aleatoria en 24 grupos de igual tamaño (100 conglomerados), y cada conglomerado está conformado por 5 sub-muestras (m1, m2, m3, m4 y m5) seleccionadas aleatoriamente y a su vez, cada sub-muestra está conformada por 4 viviendas contiguas, tal como puede ver en el cuadro de distribución (4.1).

Grupos de conglomerados	Nro de conglomerados por grupo	Sub muestras de viviendas por conglomerados					Total de viv. por sub muestras	Total de viviendas
	N	m1	m2	m3	m4	m5		
G1	100	4	4	4	4	4	20	2000
G2	100	4	4	4	4	4	20	2000
G3	100	4	4	4	4	4	20	2000
G4	100	4	4	4	4	4	20	2000
G5	100	4	4	4	4	4	20	2000
G6	100	4	4	4	4	4	20	2000
G7	100	4	4	4	4	4	20	2000
G8	100	4	4	4	4	4	20	2000
G9	100	4	4	4	4	4	20	2000
G10	100	4	4	4	4	4	20	2000
G11	100	4	4	4	4	4	20	2000
G12	100	4	4	4	4	4	20	2000
G13	100	4	4	4	4	4	20	2000
G14	100	4	4	4	4	4	20	2000
G15	100	4	4	4	4	4	20	2000
G16	100	4	4	4	4	4	20	2000
G17	100	4	4	4	4	4	20	2000
G18	100	4	4	4	4	4	20	2000
G19	100	4	4	4	4	4	20	2000
G20	100	4	4	4	4	4	20	2000
G21	100	4	4	4	4	4	20	2000
G22	100	4	4	4	4	4	20	2000
G23	100	4	4	4	4	4	20	2000
G24	100	4	4	4	4	4	20	2000
Totales	2400	96	96	96	96	96	480	48000

Cuadro 4.1: Cuadro de distribución de sub muestras, según grupo de conglomerados

Entonces, de esta distribución se tiene que, el tamaño muestral mensual es de 1,600 viviendas obtenidas de 400 conglomerados, por ende, el acumulado trimestral sería de 4,800 viviendas obtenidas de 1,200 conglomerados. Así mismo, el tamaño muestral anual quedaría conformado por 19,200 viviendas obtenidas de un total de 4,800 conglomerados.

4.7. Probabilidad de selección de la muestra

Dado que el diseño muestral es de dos etapas, entonces la probabilidad de selección de este diseño puede expresarse según el muestreo con probabilidad proporcional al tamaño (NU, 2008), mediante la siguiente ecuación:

$$P(\alpha\beta) = \underbrace{P(\alpha)}_{prob\ fase\ 1} \overbrace{P(\beta/\alpha)}^{prob\ fase\ 2}$$

donde:

$P(\alpha\beta)$: es la probabilidad del hogar β de ser seleccionado en el conglomerado α.

$P(\alpha)$: es la probabilidad de un conglomerado α de ser seleccionado.

$P(\beta/\alpha)$: es la probabilidad condicional de seleccionar el hogar β en la segunda etapa teniendo en cuenta, que el conglomerado α fue seleccionado en la primera etapa (p. 57).

Con la finalidad de resolver la ecuación dada, se fija el tamaño total de la muestra en número de viviendas, para esto es necesario una muestra de n viviendas con igual probabilidad, de una población con N viviendas. Así, la tasa de muestreo total será $\frac{n}{N}$ que será igual a $P(\alpha\beta)$ tal como se define en la ecuación siguiente, pero antes se define los siguientes términos (NU, 2008):

a : es el número de conglomerados que se quiere incluir en la muestra.

b : es el número de viviendas que se quiere seleccionar en cada conglomerado, independiente del tamaño de los conglomerados seleccionados.

m_i : es el tamaño del α-ésimo conglomerado.

entonces $P(\beta/\alpha)$ es igual a b/m_i y por tanto:

$$P(\alpha\beta) = [P(\alpha)] \left[\frac{b}{m_i}\right], \tag{4.1}$$

dado que $n = ab$, reemplazando se tiene

$$\frac{ab}{N} = [P(\alpha)] \left[\frac{b}{m_i}\right],$$

resolviendo esta última expresión para $P(\alpha)$, se obtiene (p. 58)

$$P(\alpha) = \frac{am_i}{N}. \tag{4.2}$$

Teniendo en cuenta que $N = \sum m_i$ de forma que la probabilidad de seleccionar un conglomerado es proporcional a su tamaño, entonces reemplazando la ecuación (4.1) en (4.2), se obtiene la probabilidad final de seleccionar una vivienda en la muestra, como se muestra en la siguiente expresión:

$$P(\alpha\beta) = \left[\frac{am_i}{\sum m_i}\right]\left[\frac{b}{m_i}\right] = \left[\frac{ab}{\sum m_i}\right]. \tag{4.3}$$

4.8. Factor de expansión

Para que las estimaciones de la EPE sean representativas de la población, es necesario multiplicar los datos de cada vivienda de la muestra contenidos en la base de datos por el peso o factor de expansión calculada según el diseño muestral (INEI, 2020). El factor de expansión final para cada registro tiene 2 componentes; el factor básico de muestreo y los factores de ajuste por la no respuesta.

El factor básico de muestreo para cada vivienda muestral es determinado por el diseño muestral, que equivale, al inverso del producto de las probabilidades de la primera y segunda etapa, de la ecuación (4.3) se tiene:

$$P = \underbrace{\left[\frac{am_i}{\sum m_i}\right]}_{prob\ etapa\ 1} \overbrace{\left[\frac{b}{m_i}\right]}^{prob\ etapa\ 2},$$

donde P es la probabilidad final de selección de las viviendas.

Entonces el factor de expansión básico de puede expresar, como

$$W_i = \frac{1}{P}.$$

A esta última ecuación es importante ajustar por la magnitud de la no respuesta. Dado que los factores de expansión son calculados a nivel de cada UPM seleccionada, entonces el factor de expansión final para la i-ésima UPM seleccionada se puede expresar como

$$W_i' = (W_i)\left(\frac{m_i}{m_i'}\right),$$

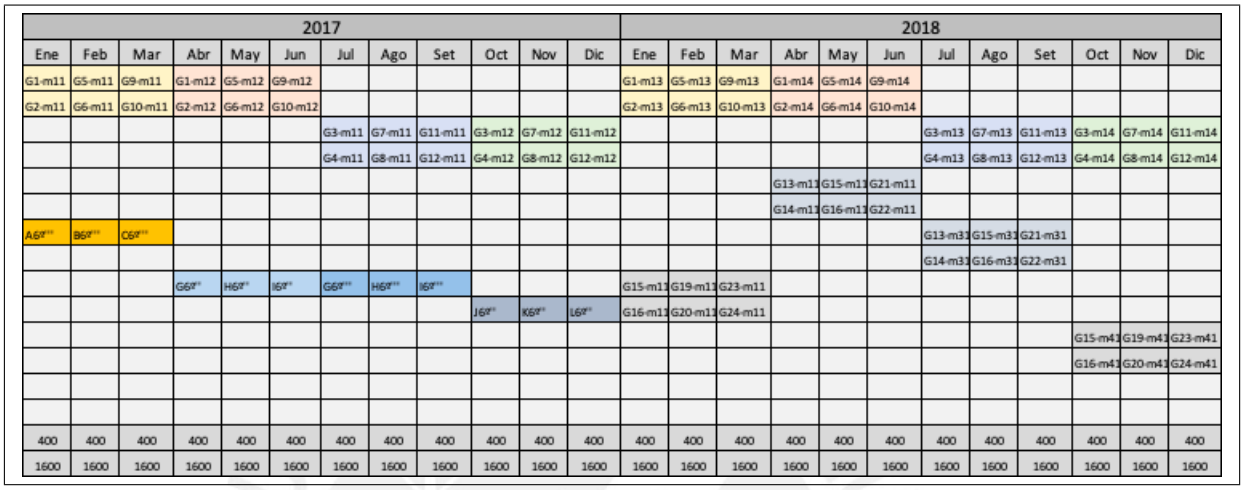
donde

m_i : viviendas seleccionadas en la i-ésima UPM

m_i' : viviendas entrevistadas en la i-ésima UPM

4.9. Rotación de la muestra

El esquema de rotación se plantea de la siguiente forma, dado que mensualmente se entrevista a 1,600 viviendas particulares, de las cuales, 800 son visitadas por primera vez y 800 son visitadas por segunda vez (viviendas panel), el esquema de rotación para los años 2017 y 2018 se puede visualizar Figura (4.1).



2017												2018											
Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov	Dic	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov	Dic
G1-m11	G5-m11	G9-m11	G1-m12	G5-m12	G9-m12							G1-m13	G5-m13	G9-m13	G1-m14	G5-m14	G9-m14						
G2-m11	G6-m11	G10-m11	G2-m12	G6-m12	G10-m12							G2-m13	G6-m13	G10-m13	G2-m14	G6-m14	G10-m14						
						G3-m11	G7-m11	G11-m11	G3-m12	G7-m12	G11-m12							G3-m13	G7-m13	G11-m13	G3-m14	G7-m14	G11-m14
						G4-m11	G8-m11	G12-m11	G4-m12	G8-m12	G12-m12							G4-m13	G8-m13	G12-m13	G4-m14	G8-m14	G12-m14
															G13-m11	G15-m11	G21-m11						
															G14-m11	G16-m11	G22-m11						
A6$^{a'''}$	B6$^{a'''}$	C6$^{a'''}$																G13-m31	G15-m31	G21-m31			
																		G14-m31	G16-m31	G22-m31			
			G6$^{a''}$	H6$^{a''}$	I6$^{a''}$	G6$^{a'''}$	H6$^{a'''}$	I6$^{a'''}$				G15-m11	G19-m11	G23-m11									
									J6$^{a''}$	K6$^{a''}$	L6$^{a''}$	G16-m11	G20-m11	G24-m11									
																					G15-m41	G19-m41	G23-m41
																					G16-m41	G20-m41	G24-m41
400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400	400
1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600

Figura 4.1: Esquema de rotación de la muestra para los años 2017 y 2018

Por ejemplo, observando el esquema de rotación; en el primer trimestre del año 2018, se realizó una tercera visita a las viviendas visitadas por primera vez, en el primer trimestre móvil de enero a marzo en 2017, en los grupos: G1-G2, G5-G6 y G9-G10 de la sub-muestra 01, lo cual permitirá resultados trimestrales. Así mismo, se realizó la primera visita de la muestra nueva en los grupos: G15-G16, G19-G20 y G23-G24 también de la sub-muestra 01.

En el primer trimestre del año 2017, se visitó por primera vez a las viviendas de los grupos; G1-G2, G5-G6 y G9-G10 de la sub-muestra 01, la cuarta y última visita a las viviendas de los grupos: $A6^{a'''}$, $B6^{a'''}$, $C6^{a'''}$ estas sub-muestras fueron visitadas por primera vez en el último trimestre del año 2015.

Otro ejemplo, en el tercer trimestre del año 2018, se realizó la cuarta visita a las viviendas visitadas por primera vez en el trimestre julio a setiembre de 2017, en los grupos: G3-G4, G7-G8 y G11-G12 de la sub-muestra 01, lo cual permite comparar los resultados trimestrales anuales. Así mismo, se realizó la primera visita de la muestra nueva en los grupos: G13-G14, G17-G18 y G21-G22 de la sub-muestra 03.

Bajo este esquema de rotación, las estimaciones trimestrales se basan en una muestra trimestral móvil de 4,800 viviendas particulares, de las cuales, la mitad (2,400 viviendas) son panel y los otros (2,400 viviendas) se renueva cada trimestre, entonces bajo este esquema es posible realizar comparaciones anuales y trimestrales.

4.10. Unidad de investigación

La EPE tiene como unidad de investigación la vivienda definido por INEI (2020) que está constituida por:

- Los integrantes del hogar familiar.
- Los trabajadores del hogar con cama adentro que reciban o no pago por sus servicios.
- Los integrantes de una pensión familiar que tienen como máximo 9 pensionista, y
- Las personas que no son miembros del hogar familiar, pero que estuvieron presentes en el hogar los últimos 30 días.

4.11. Tasa de desempleo

La tasa de desempleo es probablemente el indicador más conocido del mercado laboral y uno de los más citados por los medios en muchos países. La tasa de desempleo es una medida útil de la sub-utilización de la oferta de trabajo. Refleja la incapacidad de una economía para generar empleo para aquellas personas que desean trabajar pero no lo hacen, aunque estén disponibles para el empleo y busquen trabajo activamente. Por lo tanto, es visto como un indicador de la eficiencia y efectividad de una economía para absorber su fuerza laboral y del desempeño del mercado laboral (International Labour Organization [ILO], 2019).

Dada su utilidad para transmitir información valiosa sobre la situación del mercado laboral de un país y del hecho de que es ampliamente reconocido como un indicador principal del mercado laboral, se incluyó como uno de los indicadores propuestos para medir el progreso hacia el logro de los Objetivos de Desarrollo Sostenible - ODS, propuestos por las Naciones Unidas bajo el Objetivo 8 (Promover el crecimiento económico sostenido, inclusivo y sostenido, el empleo pleno y productivo, y el trabajo decente para todos)[2].

4.11.1. Definición

La tasa de desempleo es calculado expresando el número de personas desempleadas como un porcentaje del número total de personas en la fuerza laboral. La fuerza laboral (formalmente conocido como la Población Económicamente Activa - PEA) es la suma del número de personas que se encuentran activamente empleadas más el número de personas desempleadas. Así, para la medición de la tasa de desempleo se requiere conocer la cantidad de empleados y desempleados (ILO, 2019).

Los desempleados son todas aquellas personas que se encuentran en edad de trabajar y estaban (ILO 2019):

[2]El indicador 8.5.2 se refiere a la tasa de desempleo desglosado por sexo, edad y personas con discapacidad. Para ver la lista oficial de los indicadores de los ODS propuestos, ver: http://unstats.un.org/sdgs/indicators/indicators-list/ y para Perú lo mide el INEI, ver: http://ods.inei.gob.pe/ods/objetivos-de-desarrollo-sostenible

- Sin trabajar durante el periodo de referencia, es decir, no estaban en un empleo remunerado o por cuenta propia (auto-empleo).
- Disponibles para trabajar, es decir, estaban disponibles para un empleo remunerado o por cuenta propia en el periodo de referencia, y
- Buscando trabajo, también se incluyen, personas quienes no estaban buscando trabajo pero que en un futuro cercano (no más de tres meses) tendrán una participación en el mercado laboral (haciendo arreglos para empezar un trabajo) son también considerados como desempleados, así como los participantes en programas de capacitación en programas de promoción del empleo, quienes sobre esta base, no estaban (con empleo), no estaban (disponibles actualmente) y no (buscaron empleo) porque tenían una oferta de trabajo para comenzar en un futuro cercano (no mayor a tres meses) y personas (sin empleo) que realizaron actividades para migrar al extranjero con el fin de trabajar por un sueldo o ganancia, pero que todavía estaban esperando la oportunidad para irse.

Los empleados son todas aquellas personas que se encuentran en edad de trabajar y quienes durante un periodo específico de tiempo, tal como, un día o una semana, estaban en la siguientes categorías (ILO, 2019):

- Con empleo remunerado (ya sea, se encuentre trabajando o de permiso temporal pero con empleo) ó
- Como trabajador por cuenta propia (ya sea, se encuentre trabajando o que tiene una empresa aunque no esté trabajando)

4.11.2. Cálculo de la tasa de desempleo

Conforme a las normas internacionales establecidas por la OIT y que el INEI los toma como referencia, para obtener la tasa de desempleo se usa la siguiente expresión (INEI, 2020), gráficamente se puede observar el procedimiento seguido para obtener los componentes para el cálculo de la tasa de desempleo (ver Figura 4.2):

$$\text{Tasa de Desempleo } (\%) = \frac{\text{Personas Desocupadas}}{\text{Población Economicamente Activa}} * 100$$

4.11.3. Población en Edad de Trabajar (PET)

Es aquella población definida por las norma internacionales (OIT), como apta en cuanto a edad para ejercer funciones productivas (de 14 años y más de edad en Perú). Esta se subdivide en población económicamente activa (PEA) y población económicamente inactiva (NO PEA) (INEI, 2020).

$$\text{PET} = \text{PEA} + \text{NO PEA (fuera de la fuerza de trabajo)}$$

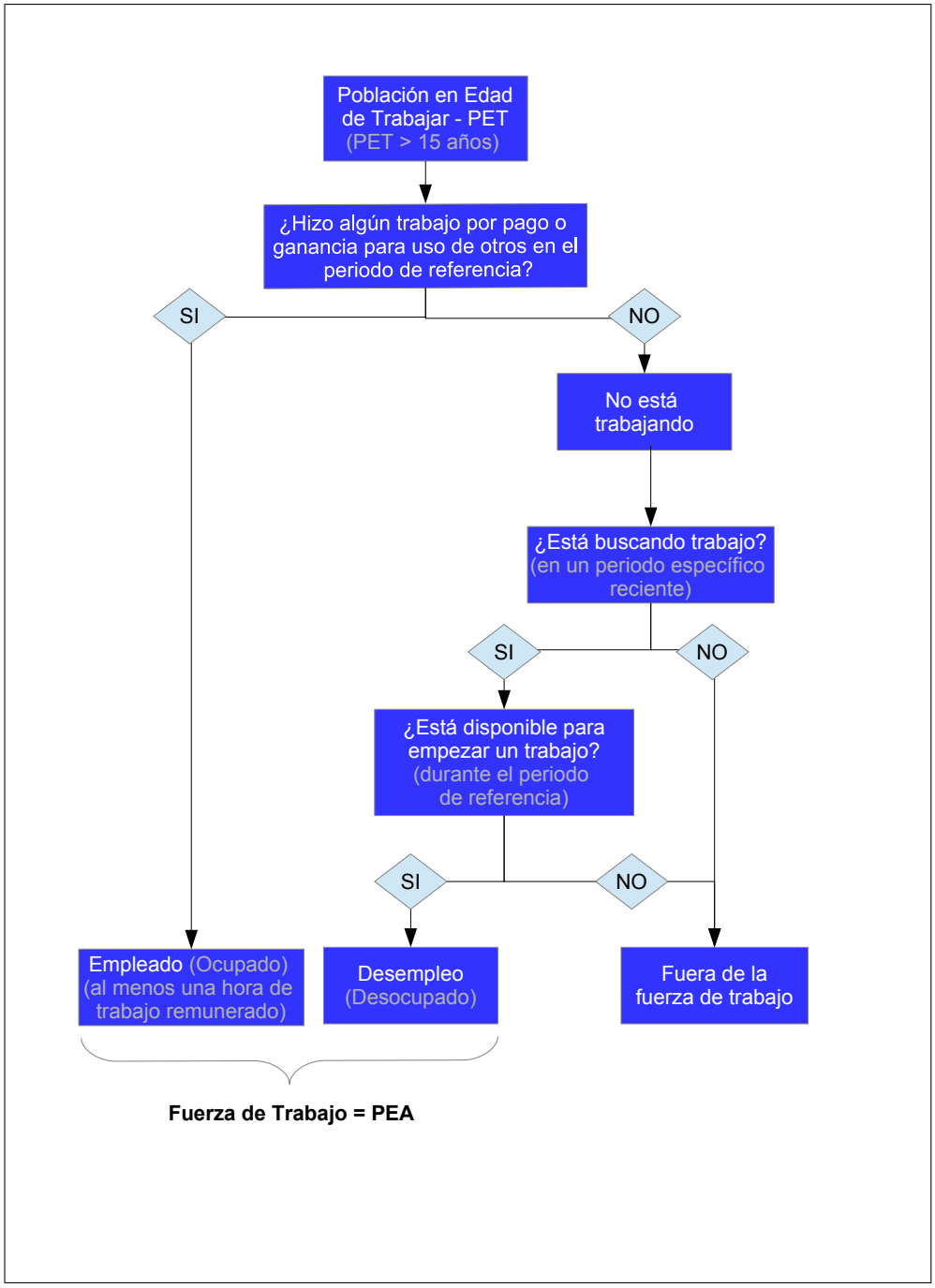


Figura 4.2: Gráfica del procedimiento estructural para obtener los componentes de la tasa de desempleo

4.11.4. Población Económicamente Activa (PEA)

También conocido como fuerza de trabajo, la PEA comprende a todas las personas de catorce (14) años y más de edad que en la semana de referencia se encontraban (INEI, 2020):

- Trabajando.
- No trabajando pero tenían trabajo.
- Se encontraban buscando activamente un trabajo.

$$\text{PEA} = \text{Ocupados} + \text{Desocupados}$$

4.11.5. Ocupados

Los ocupados se determinan según los cuatro criterios definidos por el INEI (2020):

- Ocupados son las personas de 14 años y más de edad que estuvieron participando en alguna actividad económica, en el periodo de referencia.
- Los trabajadores dependientes, que teniendo empleo fijo, no trabajaron en el periodo de referencia por hallarse de vacaciones, huelga, licencia por enfermedad, licencia pre y post natal, etc. todas ellas pagadas.
- Los trabajadores independientes, que estuvieron temporalmente ausentes del trabajo durante el periodo de referencia, pero la empresa o negocio siguió funcionando.
- A las personas que no estuvieron en ninguna de las condiciones anteriores, se les indaga si realizaron alguna actividad económica en el periodo de referencia, al menos una hora, por lo cual recibirá pago en dinero y/o especie (el objetivo es recuperar las actividades realizadas, pero que no son consideradas como trabajo por las personas).

También se consideran a las personas que trabajaron 15 horas o más como trabajador familiar no remunerado, a los practicantes con o sin remuneración y a los oficiales y suboficiales de las Fuerzas Armadas y las Fuerzas Policiales.

4.11.6. Desocupados

Se definen como desocupados (según OIT - 2013) a todas aquellas personas, de uno u otro sexo, que durante el periodo de referencia cumplen en forma simultanea con los 3 requisitos siguientes (INEI, 2020):

- Sin empleo, es decir; que no tienen ningún empleo como asalariado o independiente.
- Corrientemente disponible para trabajar, es decir; con disponibilidad para trabajar en un empleo asalariado o independiente, durante el periodo de referencia.
- En busca de empleo, es decir; que habían tomado acciones concretas para buscar un empleo asalariado o independiente, en un periodo de tiempo especificado.

En la definición se consideran tanto a personas que buscaron trabajo pero que trabajaron antes (cesantes), como a los que buscaron trabajo por primera vez (aspirantes). Para mayor detalle de las definiciones (ver ficha técnica de la EPE).

Capítulo 5

Estudio de Simulación

En este capítulo nuestro objetivo es poner en relevancia el tratamiento de series de tiempo con errores de medición correlacionados, estos errores se presentan, por ejemplo, en los datos longitudinales donde la información se recoge de las mismas unidades de muestreo en diferentes puntos del tiempo, donde el error de medición de la unidad de muestreo se repite cada vez que es incluido en la muestra. Entonces se introduce la correlación entre estos errores al considerar en la muestra la misma unidad en repetidas ocasiones.

Para el objetivo planteado, se ha diseñado un estudio de simulación con 5, 000 series de tiempo con errores correlacionados de 200 observaciones por serie, las series fueron generadas utilizando el modelo estructural básico representado en la forma de un modelo de espacio de estados con errores correlacionados, donde los errores fueron generados mediante una distribución condicional normal. En tanto que, para la estimación de los componentes de cada serie simulada se han utilizado, el algoritmo del Filtro de Kalman y el algoritmo de Pfeffermann y Tiller propuesto en 2006.

Se utilizan dos algoritmos diferentes para la estimación de los componentes, con la idea de ver si hay diferencias en las estimaciones obtenidas, entre el algoritmo del filtro de Kalman que supone errores no correlacionados y el algoritmo Pfeffermann y Tiller que supone errores correlacionados. Para concluir si existen diferencias se calculan la magnitud de los errores de predicción, en un primer momento se calculan estos errores comparando los componentes del vector de estados originales de las series generadas en la simulación con los componentes estimados con los algoritmos en mención, luego como segundo acto, se calculan los mismos errores comparando las observaciones originales de las series simuladas con las observaciones proyectadas con cada algoritmo. Las proyecciones con el algoritmo de Pfeffermann y Tiller incluyen además de las observaciones futuras, las proyecciones de los errores correlacionados, dado que suponemos conocer la distribución que los genera y el orden de las correlaciones.

Consideraremos los siguientes errores de predicción para la comparación; el error medio, el error absoluto medio y el error cuadrático medio. Comparando estos errores podremos afirmar o no, si existen las diferencias en la precisión y con cuál de los algoritmos se obtienen menores errores. Precisar que para el objetivo del estudio nosotros suponemos que los parámetros del modelo estructural básico son conocidos, pero que en la práctica es necesario calcular previamente mediante la estimación por Máxima Verosimilitud el cual no detallaremos por

no ser objeto del presente estudio.

5.1. Series de tiempo generadas con errores correlacionados

Con la finalidad de poder generar 5,000 series de tiempo y_t con errores e_t correlacionados, vamos a suponer que conocemos las correlaciones y la varianza de los errores, adicionalmente también daremos por conocido los parámetros del modelo estructural básico (para mayor detalle del modelo ver (3.29)) que usamos para generar las series. Teniendo estas consideraciones en cuenta planteamos el siguiente modelo de espacio de estados para la generación de las series:

$$\begin{aligned} y_t &= Z_t\alpha_t + e_t, \qquad e_t|e_{t-1}, e_{t-2}, e_{t-3} \sim N(\tilde{\mu}, \tilde{\Sigma}), \\ \alpha_t &= T\alpha_{t-1} + \eta_t, \quad \eta_t \sim N(0, Q). \end{aligned} \tag{5.1}$$

donde $t = 1, 2, 3, ..., 200$; η_t y e_t son independientes, suponiendo que conocemos las siguientes correlaciones $corr(e_t, e_{t-1}) = 0.58$, $corr(e_t, e_{t-2}) = 0.31$, $corr(e_t, e_{t-3}) = 0.05$ y la varianza de los errores $\sigma_e^2 = 2.8$. Además Z_t, T y Q son matrices fijos del sistema de ecuaciones (5.1), α_t es el vector de estados que contiene los componentes del modelo básico estructural utilizado, tal como fue detallado en la sección (3.2.3), y se formula de la siguiente manera:

$$\alpha_t = (\mu_t, \nu_t, \gamma_{1t}, \gamma_{1t}^*, \gamma_{2t}, \gamma_{2t}^*, \gamma_{3t}, \gamma_{3t}^*, \gamma_{4t}, \gamma_{4t}^*, \gamma_{5t}, \gamma_{5t}^*, \gamma_{6t})'_{13\times 1}$$

El vector de estados tiene una dimensión (13×1) dado que las series son mensuales, en base a este vector se definen las matrices Z_t, T y Q de la misma forma que fue definido en el ejemplo práctico (3.7), donde los valores de los parámetros de la matriz Q serán; $\sigma_\xi^2 = 0.0024$, $\sigma_\zeta^2 = 0.0004$ y $\sigma_w^2 = 0.0000001$.

Para completar el modelo de simulación (5.1), es importante detallar como se generan los errores e_t correlacionados, dado que estos tienen una distribución normal multivariada, de la forma:

$$\begin{bmatrix} e_t \\ e_{t-1} \\ e_{t-2} \\ e_{t-3} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \tag{5.2}$$

para poder calcular la media y su varianza aplicando el Lema 1, es necesario llevar a la forma de una normal condicionada como a continuación se detalla;

$$\begin{aligned} &X = e_t, \quad \mu_X = 0, \\ &Y = \begin{pmatrix} e_{t-1} \\ e_{t-2} \\ e_{t-3} \end{pmatrix}, \quad \mu_Y = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \text{entonces} \quad X|Y \sim N(\tilde{\mu}, \tilde{\Sigma}), \end{aligned}$$

donde la condicional de X dado Y tiene una distribución normal con vector de medias $\tilde{\mu}$ y matriz de varianzas $\tilde{\Sigma}$, entonces para determinar el vector de medias y la matriz de covarianzas de la distribución normal, utilizaremos la matriz Σ expresada en (5.2) y para calcular sus covarianzas vamos utilizar las correlacionados y la varianza de los errores considerados en el

planteamiento y expresar de la siguiente forma:

$$\Sigma = \begin{bmatrix} \sigma_e^2 & \rho_1\sigma_e^2 & \rho_2\sigma_e^2 & \rho_3\sigma_e^2 \\ \rho_1\sigma_e^2 & \sigma_e^2 & \rho_1\sigma_e^2 & \rho_2\sigma_e^2 \\ \rho_2\sigma_e^2 & \rho_1\sigma_e^2 & \sigma_e^2 & \rho_1\sigma_e^2 \\ \rho_3\sigma_e 2 & \rho_2\sigma_e^2 & \rho_1\sigma_e^2 & \sigma_e^2 \end{bmatrix} = \begin{bmatrix} 2.8 & 1.62 & 0.87 & 0.14 \\ 1.62 & 2.8 & 1.62 & 0.87 \\ 0.87 & 1.62 & 2.8 & 1.62 \\ 0.14 & 0.87 & 1.62 & 2.8 \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

donde particionando la matriz Σ se tiene

$$\Sigma_{11} = 2.8; \quad \Sigma_{12} = (1.62; 0.87; 0.14),$$
$$\Sigma_{21} = \begin{pmatrix} 1.62 \\ 0.87 \\ 0.14 \end{pmatrix}, \quad \Sigma_{22} = \begin{pmatrix} 2.8 & 1.62 & 0.87 \\ 1.62 & 2.8 & 1.62 \\ 0.87 & 1.62 & 2.8 \end{pmatrix},$$

por lo tanto, para obtener los parámetros de la distribución normal $\tilde{\mu}$ y $\tilde{\Sigma}$ se tendrá

$$\tilde{\mu} = \mu_X + \Sigma_{12}\Sigma_{22}^{-1}(Y - \mu_Y),$$
$$\tilde{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

reemplazando los valores se tiene que:

$$\tilde{\mu} = 0 + (1.62; 0.87; 0.14) \begin{pmatrix} 2.8 & 1.62 & 0.87 \\ 1.62 & 2.8 & 1.62 \\ 0.87 & 1.62 & 2.8 \end{pmatrix}^{-1} \left(\begin{pmatrix} e_{t-1} \\ e_{t-2} \\ e_{t-3} \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right),$$
$$\tilde{\mu} = 0.59e_{t-1} + 0.07e_{t-2} - 0.17e_{t-3},$$
$$\tilde{\Sigma} = 2.8 - (1.62; 0.87; 0.14) \begin{pmatrix} 2.8 & 1.62 & 0.87 \\ 1.62 & 2.8 & 1.62 \\ 0.87 & 1.62 & 2.8 \end{pmatrix}^{-1} \begin{pmatrix} 1.62 \\ 0.87 \\ 0.14 \end{pmatrix},$$
$$\tilde{\Sigma} = 1.8$$

reemplazando estos valores en (5.1) se realizan las simulaciones de las series, donde los errores están correlacionados y tienen una distribución N($\tilde{\mu}$,$\tilde{\Sigma}$). A continuación explicamos el detalle del proceso de generación de las series:

a) Como punto de partida suponemos que los valores iniciales para $t = 1, 2, 3$ de las tendencias y los efectos de estacionalidad son ceros, es decir; $\alpha_1 = \alpha_2 = \alpha_3 = (0, ..., 0)'_{13\times 1}$, de forma similar para los errores $e_1 = e_2 = e_3 = 0$.

b) Generar el vector $\eta_4 \sim N(0, Q)$, considerando en Q los parámetros planteados.

c) Generar el vector de estados $\alpha_4 = T\alpha_3 + \eta_4$.

d) Generar el error e_4 con la distribución normal, $e_4|e_1, e_2, e_3 \sim N(\tilde{\mu}, \tilde{\Sigma})$.

e) Generar la observación $y_4 = Z_t\alpha_4 + e_4$, dado que las primeras observaciones $y_1 = y_2 = y_3 = 0$.

f) Así sucesivamente hasta generar las 300 observaciones por cada serie, sin embargo es necesario descartar las primeras 100 observaciones para que las series no estén influidos por los valores iniciales, quedándonos solo con las 200 restantes.

g) Todo el proceso se repite hasta generar las 5,000 series.

Como resultado de este proceso de simulación, en la Figura (5.1) podemos observar las gráficas de algunas series generadas mediante el modelo de espacio de estados (5.1) cuyos errores no son independientes y tienen distribución condicional normal.

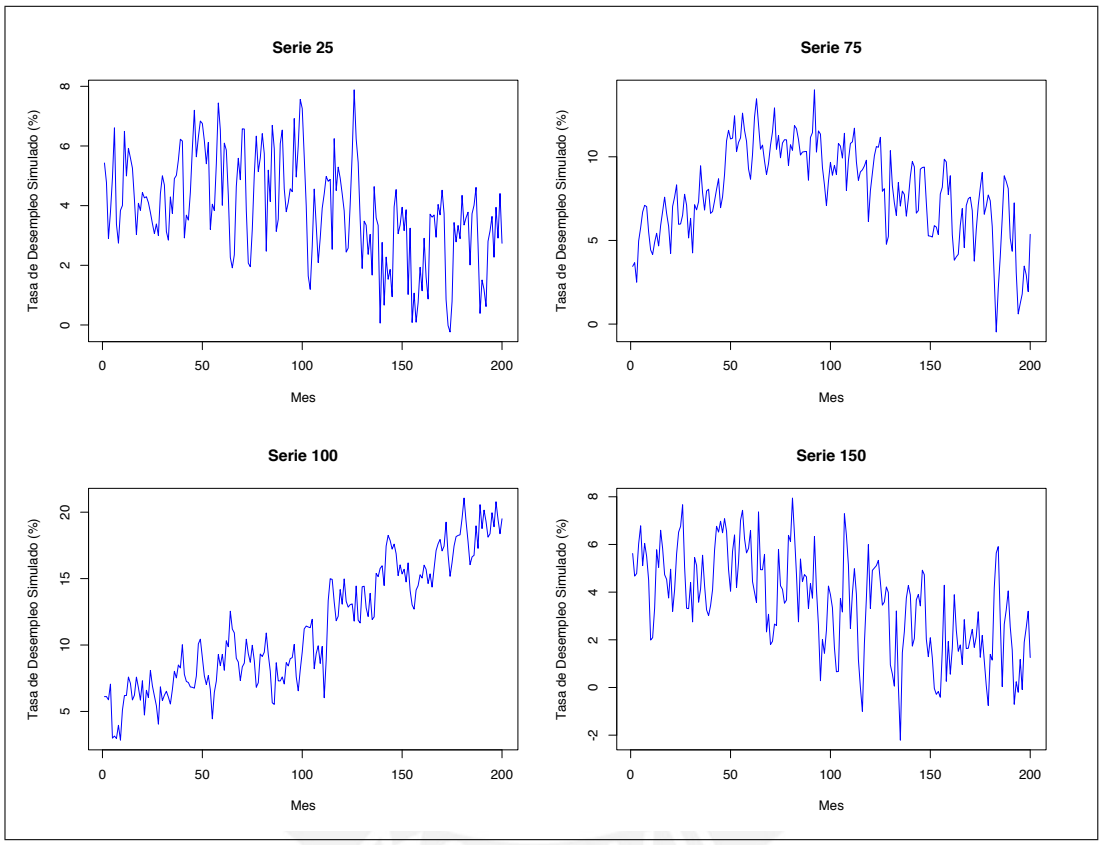


Figura 5.1: Series de tiempo con errores correlacionados generados mediante el modelo estructural básico expresado en la forma de modelo de espacio de estados.

5.2. Resultados

A continuación se realiza un análisis comparativo del desempeño obtenido en las estimaciones de los $\alpha_t's$ entre el algoritmo del filtro de Kalman $\alpha_{t(FK)}$ y el algoritmo de Pfeffermann y Tiller $\alpha_{t(PT)}$.

Previamente vamos explicar el proceso de estimación de los vectores de estados aplicando los algoritmos indicados. Primero, hemos estimado los vectores $\alpha_{t(FK)}$ y sus respectivas covarianzas $P_{t(FK)}$ para cada tiempo t con las series generadas con (5.1), para ello hemos aplicado el sistema recursivo del filtro de Kalman expresado en (3.21) que supone que las series no tienen errores correlacionados, luego de forma similar, se ha estimado los vectores $\alpha_{t(PT)}$ y sus correspondientes covarianzas $P_{t(PT)}$ para cada tiempo t, aplicando el algoritmo

recursivo de Pfeffermann y Tiller expresado en (3.28) el cual sí supone la existencia de errores correlacionados en las observaciones de las series.

Seguidamente, con los vectores de estados estimados con cada método se obtienen las proyecciones respectivas, esto es; $\hat{Y}_{ti(FK)} = Z_t\hat{\alpha}_{ti(FK)}$ aplicando el filtro de Kalman y $\hat{Y}_{ti(PT)} = Z_t\hat{\alpha}_{ti(PT)}$ aplicando algoritmo de Pfeffermann y Tiller, donde $t = 1, 2, 3, ..., 200$ e $i = 1, 2, 3, ..., 5000$. Adicionalmente necesitamos calcular la predicción de los errores $E(e_{ti}|e_{t-1,i}, e_{t-2,i}, e_{t-3,i})$, con la finalidad de poder agregar estas proyecciones a las proyecciones obtenidas con el algoritmo de Pfeffermann y Tiller $\hat{Y}_{ti(PT)} = Z_t\hat{\alpha}_{ti(PT)} + E(e_{ti}|e_{t-1,i}, e_{t-2,i}, e_{t-3,i})$ debido a que como ya hemos mencionado, este algoritmo supone errores correlacionados. Así podremos comparar con las proyecciones obtenidas con el algoritmo de filtro de Kalman que no consideran estos errores, por ello a continuación se muestra como se obtienen estas proyecciones de los errores:

- Primero determinamos errores de predicción para $t = 1, 2, 3$, con las observaciones generadas y_{ti} y los componentes del vector de estados α_{ti} se cumple que: $e_{1i} = y_{1i} - \mu_{1i} - \gamma_{1i}$, $e_{2i} = y_{2i} - \mu_{2i} - \gamma_{2i}$ y $e_{3i} = y_{3i} - \mu_{3i} - \gamma_{3i}$.
- Para $t = 4$, la predicción de los errores será de la forma $E(e_{4i}|e_{3i}, e_{2i}, e_{1i}) = 0.59e_{1i} + 0.07e_{2i} - 0.17e_{3i}$.
- Para $t = 5$, serán $E(e_{5i}|e_{4i}, e_{3i}, e_{2i}) = 0.59e_{2i} + 0.07e_{3i} - 0.17e_{4i}$ y así sucesivamente hasta completar la predicción de errores para todas las series i y todos los tiempos t.

Para medir la precisión de las proyecciones vamos a calcular; el Error Medio (EM), el Error Absoluto Medio (EAM) y el Error Cuadrático Medio (ECM) cuyas expresiones matemáticas están detalladas en (5.3). Para el calculo de estos errores se requieren de dos elementos; las observaciones conocidas representadas por y_{ti}, y de observaciones futuras obtenidas de las proyecciones definidas en el párrafo anterior, que son representadas por $\hat{y}_{ti}$. Entonces el EM viene a ser el promedio de la comparación entre las observaciones conocidas y las proyecciones considerando el signo de la diferencia, en cambio el EAM es de lectura simple que expresa la diferencia en valor absoluto, y finalmente el ECM es el promedio de las diferencias al cuadrado, por tanto incorpora la varianza del estimador y su sesgo. A continuación expresamos las ecuaciones para determinar cada uno de los errores:

$$\begin{aligned} EM_t &= \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_{t,i} - y_{t,i}) = \frac{1}{m}\sum_{i=1}^{m} e_{t,i}, \\ EAM_t &= \frac{1}{m}\sum_{i=1}^{m}|\hat{y}_{t,i} - y_{t,i}| = \frac{1}{m}\sum_{i=1}^{m}|e_{t,i}|, \\ ECM_t &= \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_{t,i} - y_{t,i})^2 = \frac{1}{m}\sum_{i=1}^{m} e_{t,i}^2. \end{aligned} \tag{5.3}$$

Para realizar el análisis de los resultados empíricos vamos a tener 2 escenarios, de acuerdo a lo definido en los párrafos anteriores y detallamos a continuación:

- En el primer escenario vamos a calcular los errores presentados en (5.3) considerando como observaciones conocidas a $y_{ti} = Z_t\alpha_{ti}$ donde α_{ti} es el vector de estados generado

en la simulación para la serie i en el tiempo t, y considerando las proyecciones $\hat{y}_{ti} = Z_t\hat{\alpha}_{ti(FK)} = \hat{Y}_{FK}$ aplicando el filtro de Kalman, y $\hat{y}_{ti} = Z_t\hat{\alpha}_{ti(PT)} = \hat{Y}_{PT}$ aplicando el algoritmo de Pfeffermann y Tiller, cuyos resultados se pueden ver en el Cuadro (5.1).

	$t = 25$		$t = 75$		$t = 150$	
	$\hat{Y}_{FK}$	$\hat{Y}_{PT}$	$\hat{Y}_{FK}$	$\hat{Y}_{PT}$	$\hat{Y}_{FK}$	$\hat{Y}_{PT}$
EM	0.014	0.038	0.002	0.002	0.014	0.018
EAM	1.048	1.087	0.711	0.763	0.643	0.689
ECM	1.742	1.874	0.793	0.918	0.648	0.748

Cuadro 5.1: Tabla de errores de predicción para las proyecciones realizadas con el filtro de Kalman $\hat{Y}_{FK} = Z_t\hat{\alpha}_{ti(FK)}$ y para las proyecciones realizadas con el algoritmo de Pfeffermann y Tiller $\hat{Y}_{PT} = Z_t\hat{\alpha}_{ti(PT)}$ para $t = 25$, $t = 75$ y $t = 150$.

- El segundo escenario es diferente del primero porque calculamos los mismos errores de (5.3) pero considerando como observaciones conocidas a las series simuladas $y_{ti} = Y_{ti(sim)}$ y las proyecciones serán las mismas que en el escenario anterior aplicando el algoritmo del filtro de Kalman, mientras que le adicionamos la predicción de los errores a las proyecciones obtenidas con el algoritmo de Pfeffermann y Tiller $\hat{y}_{ti} = Z_t\hat{\alpha}_{ti(PT)} + E(e_{ti}|e_{t-1,i}, e_{t-2,i}, e_{t-3,i}) = \hat{Y}_{PT}$, dado que el algoritmo considera los errores correlacionados, y cuyos resultados se pueden ver en el Cuadro (5.2).

	$t = 25$		$t = 75$		$t = 150$	
	$\hat{Y}_{FK}$	$\hat{Y}_{PT}$	$\hat{Y}_{FK}$	$\hat{Y}_{PT}$	$\hat{Y}_{FK}$	$\hat{Y}_{PT}$
EM	-0.041	-0.017	-0.026	-0.025	0.029	0.033
EAM	1.026	0.744	1.176	1.099	1.200	1.130
ECM	1.669	0.886	2.161	1.895	2.276	2.021

Cuadro 5.2: Tabla de errores de predicción para las proyecciones realizadas con el filtro de Kalman $\hat{Y}_{FK} = Z_t\hat{\alpha}_{ti(FK)}$ y para las proyecciones realizadas con el algoritmo de Pfeffermann y Tiller $\hat{Y}_{PT} = Z_t\hat{\alpha}_{ti(PT)} + E(e_{ti}|e_{t-1,i}, e_{t-2,i}, e_{t-3,i})$ para $t = 25$, $t = 75$ y $t = 150$.

Conforme a los escenarios planteados, se pueden ver los resultados empíricos para $t = 25$, $t = 75$ y $t = 150$ en los cuadros (5.1) y (5.2), de los cuales podemos sacar algunas conclusiones, en principio, si observamos los resultados del Cuadro (5.1) las magnitudes de los errores obtenidos para ambos métodos son bastante próximos, mostrando una ligera ventaja para filtro de Kalman. Como hemos indicado, este cuadro muestra los resultados para el primer escenario, donde comparamos la precisión obtenida mediante ambos métodos en la estimación de los componentes del vector de estados, es decir, para la tendencia, la pendiente y los efectos de estacionalidad. Así por ejemplo, en la Figura (5.2) se muestra la gráfica de la tendencia original y las estimadas con los algoritmos respectivos, para las series simuladas 25 y 75, se observa que la trayectoria de las estimaciones de las tendencias obtenida con el filtro de Kalman y el algoritmo de Pfeffermann y Tiller son muy próximos, confirmando gráficamente los resultados de Cuadro (5.1).

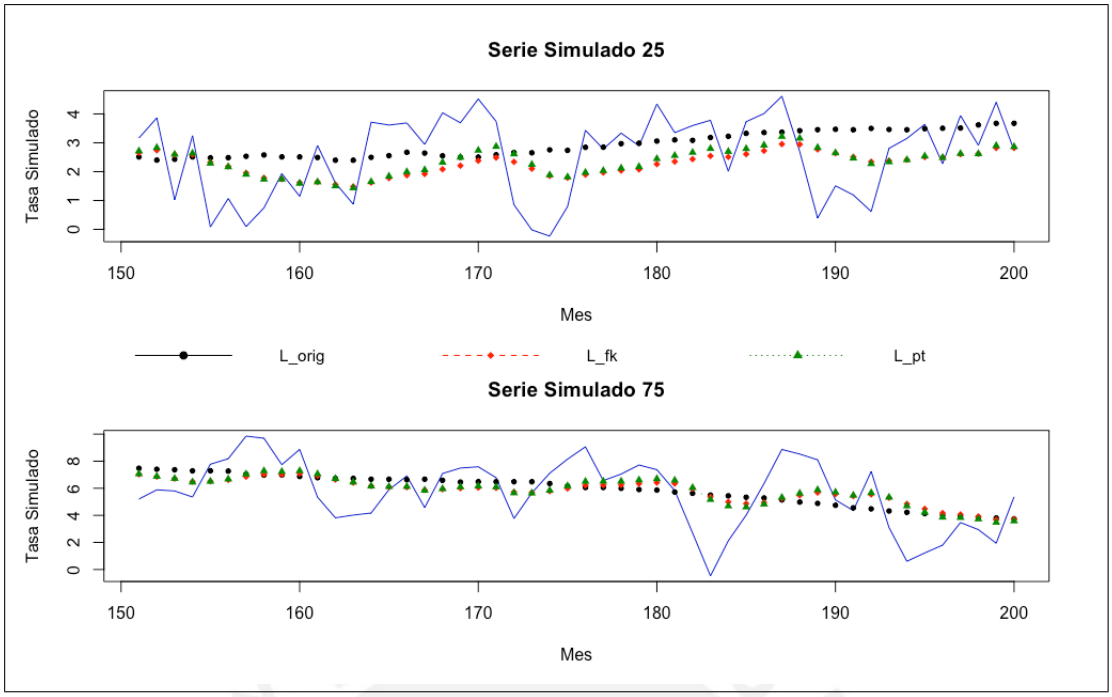


Figura 5.2: Gráfica de series originales con sus tendencias originales L_{orig} comparado con las tendencias obtenidas mediante el filtro de Kalman L_{fk} y mediante el algoritmo de Pfeffermann y Tiller L_{pt} para las 50 últimas observaciones.

Para completar el análisis de los resultados, en el Cuadro (5.2) se muestran los resultados para el segundo escenario. Como se ha detallado, en este cuadro se muestran la precisión de las predicciones alcanzadas por ambos algoritmos expresadas como errores de predicción, para obtener estos errores se comparan las observaciones conocidas que son las series de observaciones simuladas con las proyecciones obtenidas para cada algoritmo, es decir, las observaciones de las series están compuestos por los componentes del modelo estructural básico utilizado para generar las series; las tendencias, las pendientes, los efectos de estacionalidad y los términos aleatorios o irregulares.

Entonces considerando los componentes completos del modelo utilizado para generar las series, el algoritmo de Pfeffermann y Tiller tiene un desempeño más eficiente, obteniendo un menor EM, EAM y ECM para cada t, por ejemplo, en la Figura (5.3) se muestran las proyecciones obtenidas aplicando los algoritmos del filtro de Kalman y de Pfeffermann y Tiller, estos se comparan con las observaciones originales de la serie simulada número 100 para los últimos 12 periodos $t = 189, 190, ..., 200$. Se puede apreciar, que en la mayoría de puntos las proyecciones obtenidas con el algoritmo de pfeffermann y Tiller están más próximos a la observación verdadera, lo que se traduce en obtener menores errores de predicción.

Adicionalmente, en la Figura (5.4) se visualiza las observaciones proyectadas para 12 periodos con ambos algoritmos comparada con las observaciones originales de la serie simulada número 150, a diferencia del gráfico anterior, en este se muestra para el periodo $t = 171, 172, ..., 182$ a fin de variar, se aprecia un resultado similar al gráfico anterior, las proyecciones obtenidas con el algoritmo de Pfeffermann y Tiller están más próximos a las

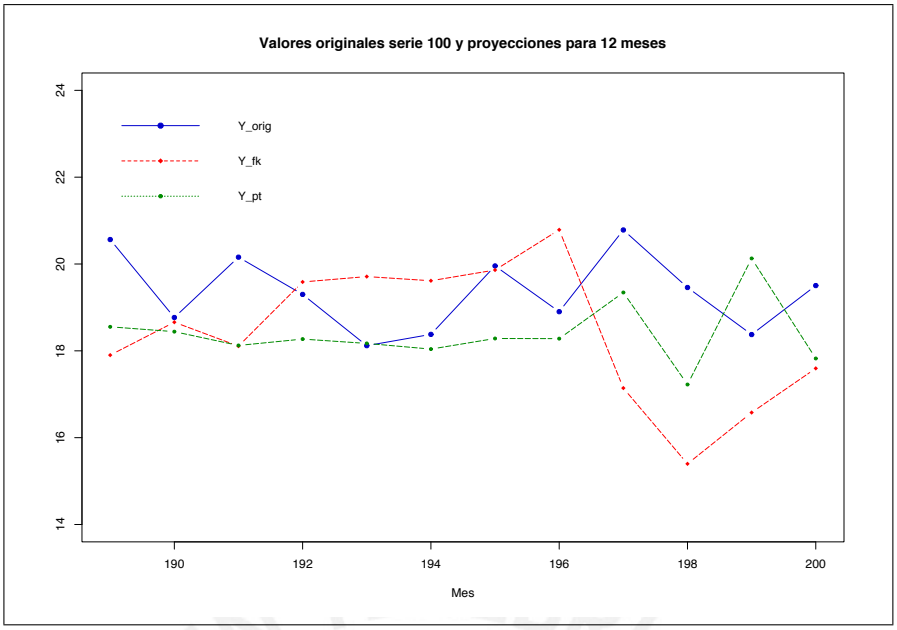


Figura 5.3: Predicciones para 12 meses obtenidas mediante el filtro de Kalman Y_{fk} y el algoritmo de Pfeffermann y Tiller Y_{pt} comparadas con las observaciones de la serie original 100 Y_{orig}.

observaciones verdaderas y por ende tienen un menor error. Lo mostrado en los gráficos confirman el resultado del Cuadro (5.2), entonces es posible decir, que cuando se tiene series de tiempo con errores correlacionados el algoritmo de Pfeffermann y Tiller ofrece mejores resultados.

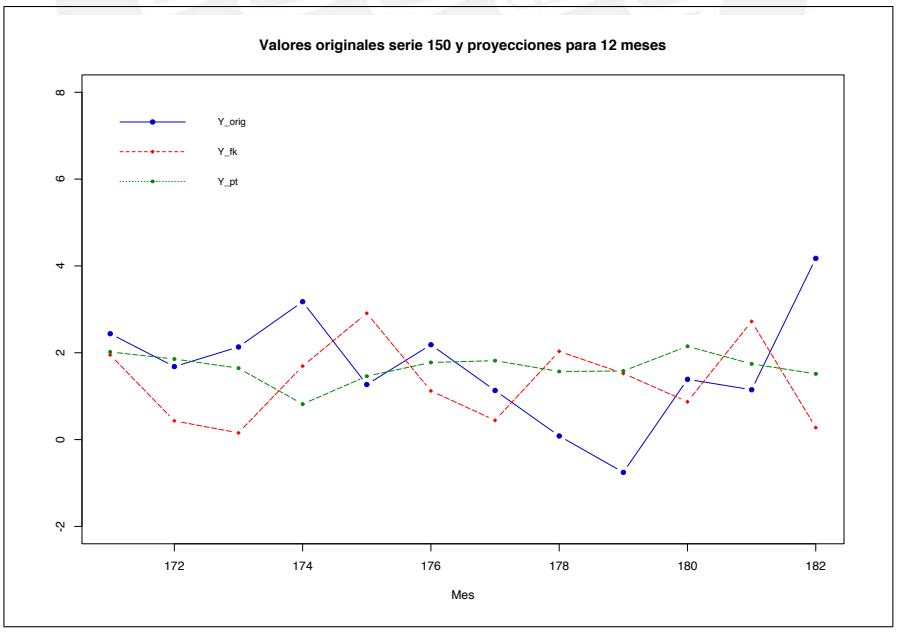


Figura 5.4: Predicciones para 12 meses obtenidas mediante el filtro de Kalman Y_{fk} y el algoritmo de Pfeffermann y Tiller Y_{pt} comparadas con las observaciones de la serie original 150 Y_{orig}.

Capítulo 6

Aplicación a la tasa de desempleo - Perú

6.1. Descripción de los datos

La actual Encuesta Permanente del Empleo - EPE, es la fuente oficial de diferentes entidades del gobierno para estimar la tasa de desempleo, un indicador fundamental del mercado laboral para medir la absorción del empleo de la economía de un país. La EPE es una encuesta representativa a nivel de Lima Metropolitana, con una muestra probabilística de 2 etapas cuya unidad primaria de muestreo son los conglomerados y la unidad secundaria las viviendas, con tamaño de muestra mensual de 1,600 viviendas. La estimación de la tasa de desempleo junto con otros indicadores del mercado laboral es publicado mensualmente por el Instituto Nacional de Estadística e Informática - INEI, cuya estimación mensual se realiza con la data acumulada de los últimos tres meses.

Los datos de la EPE son longitudinales que dependen de una muestra de panel rotativo con traslape parcial y tienen presencia de errores de muestreo correlacionados. Es importante resaltar que la presencia de errores correlacionados depende de las veces que una vivienda haya participado en el panel de la muestra, según el diseño del esquema de rotación de la EPE, para una muestra mensual de 1,600 viviendas, el 50 % (800) son de panel y el otro 50 % (800) son viviendas nuevas que se renuevan cada mes, mientras que las viviendas de panel son aquellas que fueron entrevistadas en un mes del año (Enero) como primer momento y dentro de 3 meses (Abril) vuelven a ser entrevistados como segundo momento, luego son sacados de la muestra de panel por el resto del año, pero al año siguiente vuelven a participar en el panel en los mismos meses como un tercer y cuarto momento, y se renuevan cada 2 años. En esta dinámica la participación de las viviendas se traslapa en ciertos meses (Enero-Abril-Enero-Abril) justamente es ahí donde se genera la correlación de los errores muestrales.

Explorando los datos de la tasa de desempleo, se puede ver que el diseño del esquema de rotación detallado en el párrafo anterior tiene variaciones en la composición de la muestra, por ejemplo, la muestra de Setiembre de 2017 estaba compuesto por 1,481 viviendas, de las cuales 768 eran viviendas nuevas que recién ingresaban a la muestra panel, 646 viviendas ya habían participado en la muestra en Junio del mismo año, en Setiembre y Junio de 2016 completando su cuarta participación en el panel, además se identificaron 13 viviendas en su segunda participación en el panel que ya había participado en Junio de 2017, y otras 54 viviendas en su tercera participación que ya habían sido parte de la muestra en Junio de

2017 y en Setiembre de 2016. Poniendo otro ejemplo, la muestra de Agosto de 2017 estaba integrado por 1,476 viviendas, de las cuales 763 eran viviendas nuevas, 652 viviendas estaban en su cuarta participación, es decir, que ya habían participado en Mayo de 2017, en Agosto y Mayo de 2016, otras 45 estaban en su tercera participación y otras 16 viviendas en su segunda participación en la muestra de panel. Entonces para el primer ejemplo, se introduce la correlación entre los estimadores de Setiembre y Junio de 2017, entre los estimadores de Setiembre de 2017 y Setiembre de 2016, entre los estimadores de Setiembre de 2017 y Junio de 2016, entre los estimadores de Junio de 2017 y Setiembre de 2016, entre los estimadores de Junio de 2017 y Junio de 2016, entre los estimadores de Setiembre y Junio de 2016; y de forma similar para la muestra de Agosto de 2017 cuyos estimadores estarán correlacionados con los de Mayo de 2017, con los de Agosto de 2016 y Mayo de 2016. Por lo tanto, en función a lo mostrado con estos ejemplos, la serie de observaciones de la tasa de desempleo puede presentar correlaciones de orden 3, 6, 9 y 12, dado que los ejemplos son una muestra de lo que sucede en la base datos en general, además, estas correlaciones serían positivas porque a mayor número de viviendas que se repiten en los meses relacionados mayor será el error muestral.

Para el cálculo de la tasa de desempleo, la EPE recolecta información sobre la Población Económicamente Activa - PEA definida como toda persona de 14 años a más que habita la vivienda, y dentro de la PEA se determina a la población desocupada, definida como el número de personas que forman parte de la PEA pero que al momento de la encuesta se encuentran desocupadas y en busca de empleo durante la última semana previa al día de la encuesta, entonces de la razón de la población desocupada entre la PEA sale la tasa de desempleo.

Para estimar la tasa de desempleo, se ha identificado las variables relevantes del cuestionario de la EPE consideradas para hallar la tasa de desempleo, que son; *año, mes, conglomerado, vivienda, codpanel* (código de panel), *sexo, edad, condición* (que indica la condición de ocupado o desocupado) y *factor* (factor de ponderación). Con estas variables se ha extraído un sub conjunto de datos de los microdatos publicados en la página web del INEI, por mes y por cada año desde el 2001 hasta el 2018, formando así una única base de datos para el proceso de estimación, previamente se uniformizó las variables en las bases de datos dado que no coincidían para hacer posible la unión. Adicional a estas variables, para fines del estudio se requería del *código de ubigeo* el cuál no está disponible en los microdatos pues indica el INEI que el acceso es restringido de uso interno, siendo sólo posible proporcionar el código de ubigeo a nivel de conos (los conos agrupan distritos y dividen a Lima Metropolitana en cono norte, cono sur, cono este, centro de Lima y el Callao), la idea inicial era tener a nivel de distritos para obtener la tasa de desempleo a nivel de distritos. Con toda esta información se calculo la tasa de desempleo para Lima Metropolitana y para cada una de las divisiones.

El esquema de rotación de la EPE ha sufrido variaciones desde sus inicios en el 2001 hasta la fecha, por esta razón se ha seleccionado para el análisis los datos correspondientes al periodo 2007 - 2017, porque de acuerdo a los diseños muestrales más recientes revisados, en este periodo habrían ocurrido menos variaciones, aunque no se tiene la certeza de que

modificaciones sucedieron con exactitud debido a que los antecedentes del diseño son muy generales y los documentos históricos son de acceso restringido.

Como se ha indicado en párrafos anteriores, la EPE es sólo válido para Lima Metropolitana, que está constituido por 43 distritos de la Provincia de Lima y 6 distritos de la Provincia Constitucional del Callao, para fines del estudio y de acuerdo a la disponibilidad de la información el área total se divide en cinco divisiones; Lima Norte, Lima Centro, Lima Este, Lima Sur y el Callao, que serán denotados por $i = 1, 2, 3, 4, 5$ respectivamente, sin embargo, para el estudio se considerará sólo las 3 primeras divisiones.

Con la tasa de desempleo calculado para Lima Metropolitana y sus divisiones, se puede hacer un análisis descriptivo y comparativo de las series, veamos si las series son estacionarias y si tienen efectos de estacionalidad. En la Figura (6.1) se muestran las series de la tasa de desempleo para Lima Metropolitana y El Callao, poniendo énfasis en la serie de Lima Metropolitana como un todo, vemos que es no estacionario ni en media ni en varianza, al parecer tiene efectos de estacionalidad donde los picos más elevados corresponden a las observaciones de Enero o Febrero de cada año, esto debido a que la demanda laboral de las empresas disminuye luego de las campañas navideñas y de fin de año, consecuentemente, los picos más bajos corresponden a los meses de Setiembre, Octubre, Noviembre y Diciembre dado que las campañas navideñas empiezan en esos meses y la demanda laboral de las empresas aumenta, por ende, se presentan las tasas más bajas de desempleo. La Función de Autocorrelación en los rezagos 12, 24 y 36 parece confirmar estacionalidad anual, mientras que la serie para El Callao también es no estacionario y los efectos de estacionalidad no parecen tener los mismos rezagos, por ejemplo, los picos más elevados no necesariamente corresponde a los meses de Enero o Febrero.

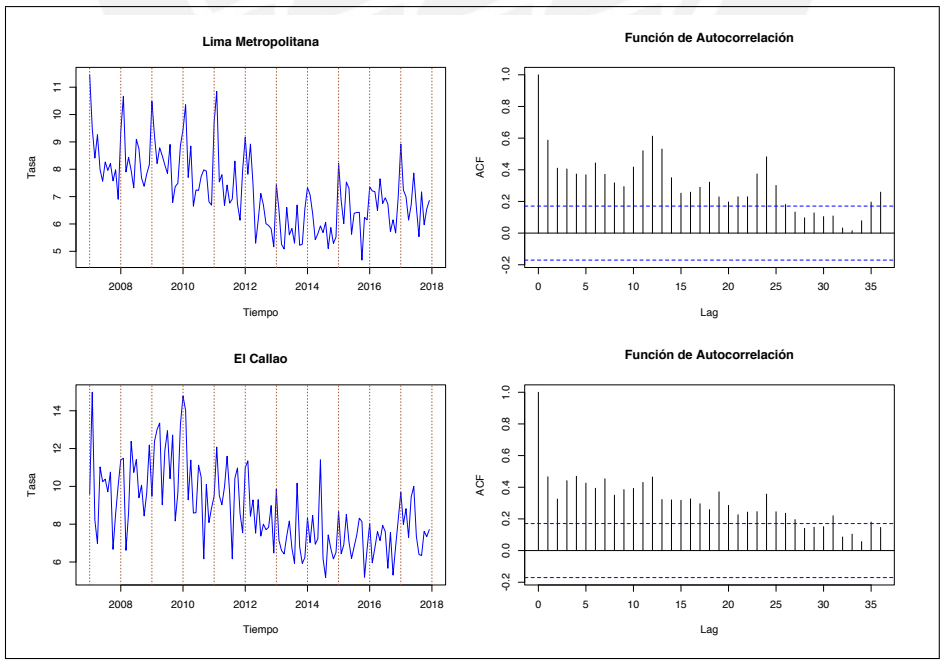


Figura 6.1: Tasa de desempleo para Lima Metropolitana y El Callao (2007 - 2017) y la Función de Autocorrelación respectiva.

En la misma lógica, en la Figura (6.2) se visualizan las series de la tasa de desempleo para Lima Centro y Lima Este ambas series son no estacionarios, sin embargo, en términos de estacionalidad la serie de Lima Este parece tener los mismos efectos que la de Lima Metropolitana aunque algo menos evidente y en Lima Centro no es muy claro si tiene los mismos efectos estacionales.

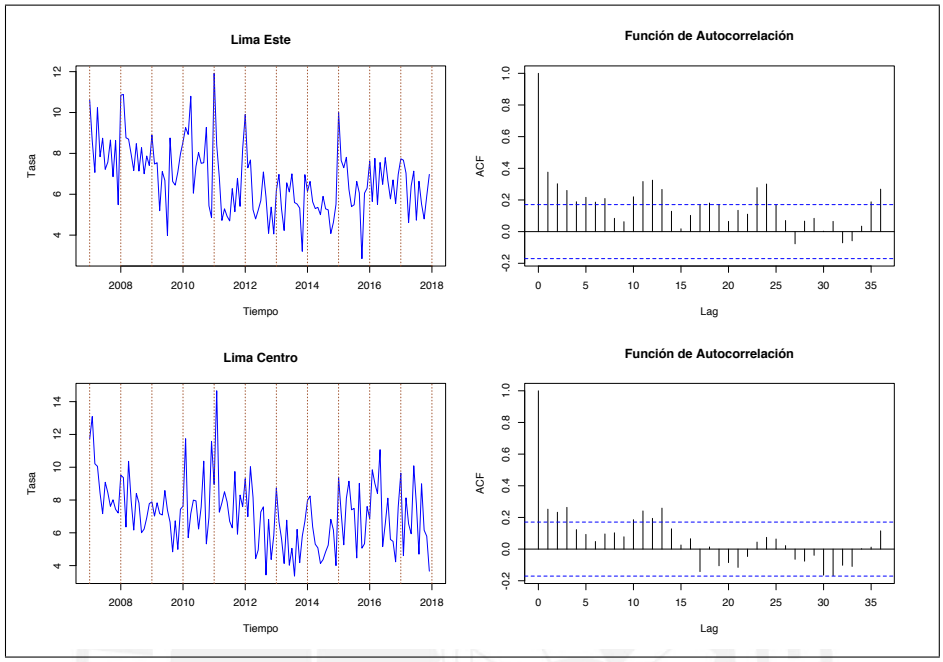


Figura 6.2: Tasa de desempleo para Lima Centro y Lima Este (2007 - 2017) y la Función de Autocorrelación respectiva.

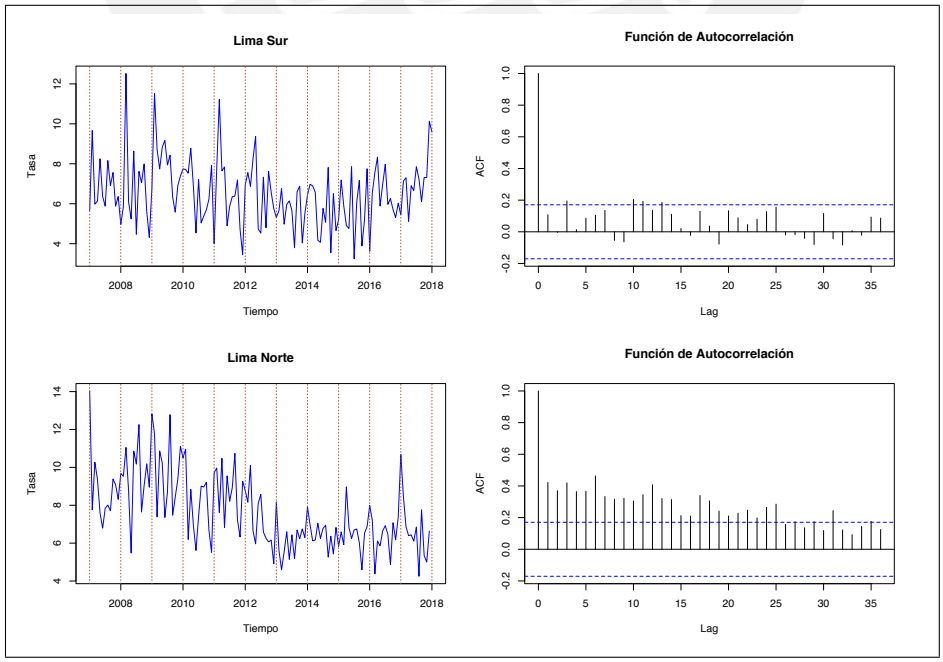


Figura 6.3: Tasa de desempleo para Lima Sur y Lima Norte (2007 - 2017) y la Función de Autocorrelación respectiva.

La dinámica laboral entre Lima Norte y Lima Sur parecen ser diferentes, al menos para la tasa de desempleo tal como de puede ver en la Figura (6.3) aunque ambas son no estacionarios. Los efectos de estacionalidad parecen ser diferentes, para Lima Norte los picos coinciden con las líneas verticales anuales similar al de Lima Metropolitana, mientras que para Lima Sur las gráficas no son muy claros para asegurar estacionalidad.

En resumen se puede decir que es necesario hacer un análisis más profundo para conocer la dinámica laboral de cada división, por ejemplo, los picos más altos coinciden en los mismos meses para Lima Metropolitana, Lima Norte y Lima Este, mientras difieren en las otras divisiones indicando que tienen una dinámica diferente. Por lo que podrían ameritar un análisis individualizado, adicionalmente existe un detalle que podría llevar a conclusiones erroneas, es que las series tienen correlación en los errores muestrales y podría confundirse con los efectos de estacionalidad, justamente los modelos de espacio de estados que consideran los errores muestrales correlacionados pueden ayudar en esto.

6.2. Modelo de espacio de estados para la tasa de desempleo

Sea y_{it} que denota la estimación directa de la tasa de desempleo (obtenida de la EPE) para la división i en el mes t, donde $i = 1, 2, 3, 4, 5$ y $t = 1, 2, 3, ..., 132$; esto debido a que en el periodo 2007-2017 se cuenta con 132 observaciones de la tasa de desempleo para cada una de las divisiones. Entonces consideremos el siguiente modelo de espacio de estados con errores correlacionados (Pfeffermann y Tiller, 2006) para las series de tiempo y_{it} sabiendo que las series tienen errores correlacionados, basado en lo que fue definido en (3.5) se tiene:

Ecuación de observación:

$$\begin{aligned} &y_{it} = Z_{it}\alpha_{it} + e_{it}, \quad E(e_{it}) = 0, \\ &E(e_{it}e'_{it}) = \Sigma_{itt}, \quad E(e_{i\tau}e'_{it}) = \Sigma_{i\tau t} \end{aligned} \tag{6.1}$$

Ecuación de estados:

$$\begin{aligned} &\alpha_{it} = T\alpha_{i,t-1} + \eta_{it}, \quad E(\eta_{it}) = 0, \\ &E(\eta_{it}\eta'_{it}) = Q_i, \quad E(\eta_{it}\eta'_{i,t-k}) = 0, \quad k > 0, \end{aligned} \tag{6.2}$$

donde se asume que $E(\eta_{it}e'_{i\tau}) = 0$ para todo t, i y τ.

Para poder construir los vectores y matrices del sistema (6.1) y (6.2), es necesario ver los elementos que contendrá el vector de estados α_{it}, para ello, es necesario definir el modelo estructural básico para las series y_{it} considerando lo definido en el marco teórico (3.2.3). Entonces, formulamos la serie y_{it} como una descomposición clásica de series de tiempo (Harvey, 1989), expresado en un modelo aditivo cuyos componentes son el *nivel de tendencia* μ_{it}, los *efectos estacionales* γ_{it} y los términos *irregulares* ε_{it} como *ruidos blancos*:

$$y_{it} = \mu_{it} + \gamma_{it} + \varepsilon_{it}; \quad \varepsilon_{it} \sim N(0, \sigma^2_{i\varepsilon}), \tag{6.3}$$

donde cada componente tiene su propia ecuación; la tendencia reflejará como cada serie se

comporta en el largo plazo, para esto la tendencia en el tiempo t dependerá del nivel de tendencia y la pendiente en el tiempo $t-1$ más un ruido blanco del mismo tiempo t, a su vez la pendiente que refleja el cambio local en el tiempo t depende de la pendiente en el tiempo $t-1$ más un ruido blanco en el tiempo t, y finalmente los efectos de estacionalidad que reflejan movimientos oscilatorios dentro del año, se formulan mediante la representación trigonométrica, de forma similar a lo desarrollado en el ejemplo práctico (3.7), cada uno de estas ecuaciones se muestra a continuación:

$$\begin{aligned}
&\mu_{it} = \mu_{i,t-1} + \nu_{i,t-1} + \xi_{it}; \quad \xi_{it} \sim N(0, \sigma^2_{i\xi}),\\
&\nu_{it} = \nu_{i,t-1} + \zeta_{it}; \quad \zeta_{it} \sim N(0, \sigma^2_{i\zeta}),\\
&\gamma_{it} = \sum_{j=1}^{6} \gamma_{jit}\\
&\gamma_{jit} = \gamma_{j,i,t-1} \cos\lambda_j + \gamma^*_{j,i,t-1} \operatorname{sen}\lambda_j + w_{jit}; \quad w_{jit} \sim N(0, \sigma^2_{iw}),\\
&\gamma^*_{jit} = -\gamma_{j,i,t-1} \operatorname{sen}\lambda_j + \gamma^*_{j,i,t-1} \cos\lambda_j + w^*_{jit}; \quad w^*_{jit} \sim N(0, \sigma^2_{iw}),\\
&\lambda_j = \frac{2\pi j}{s}, \quad j = 1, ..., s/2,
\end{aligned} \tag{6.4}$$

donde los términos ε_{it}, ξ_{it}, ζ_{it}, w_{jit}, y w^*_{jit} son series de ruidos blancos independientes y permiten que los efectos de cada componente se desarrollen de forma estocástica a lo largo del tiempo.

Con lo definido en (6.3) y sus respectivos componentes en (6.4), ya podemos determinar los elementos del vector de estados α_{it}, como ya hemos indicado la serie y_{it} de la tasa de desempleo es mensual ($s = 12$), entonces tendrá una dimensión de $((s+1)\times 1)$ y se define de la siguiente forma:

$$\alpha_{it} = (\mu_{it}, \nu_{it}, \gamma_{1it}, \gamma^*_{1it}, \gamma_{2it}, \gamma^*_{2it}, \gamma_{3it}, \gamma^*_{3it}, \gamma_{4it}, \gamma^*_{4it}, \gamma_{5it}, \gamma^*_{5it}, \gamma_{6it})'_{13\times 1},$$

en función del vector de estados definido se formulan las matrices del sistema Z_{it}, T y Q_i que tendrán la forma siguiente:

$$Z_{it} = (1,0,1,0,1,0,1,0,1,0,1,0,1)_{1\times 13};$$

$$T = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0\\ 0 & 1 & 0 & 0 & \cdots & 0\\ 0 & 0 & cos(\pi/6) & sen(\pi/6) & \cdots & 0\\ 0 & 0 & -sen(\pi/6) & cos(\pi/6) & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & 0 & \cdots & -1 \end{bmatrix}_{13\times 13}; Q_i = \begin{bmatrix} \sigma^2_{i\xi} & 0 & 0 & \cdots & 0\\ 0 & \sigma^2_{i\zeta} & 0 & \cdots & 0\\ 0 & 0 & \sigma^2_{iw} & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & \sigma^2_{iw} \end{bmatrix}_{13\times 13},$$

aquí las matrices T y Z_{it} serán invariables en el tiempo y en cada una de las divisiones, sin embargo la matriz Q_i va ser constante en el tiempo pero diferente para cada división porque depende de los siguientes parámetros $\sigma^2_{i\xi}$, $\sigma^2_{i\zeta}$ y σ^2_{iw}, a continuación vemos como se estiman.

6.2.1. Estimación de los parámetros del modelo

Los parámetros del modelo estructural básico $\sigma^2_{i\xi}$, $\sigma^2_{i\zeta}$ y σ^2_{iw} para cada serie i, son determinados maximizando el logaritmo de la función de verosimilitud $log[L(Y_{in}; \sigma^2_{i\xi}, \sigma^2_{i\zeta}, \sigma^2_{iw})]$, que depende de los errores de predicción υ_{it} y su varianza F_{it} como puede verse en (3.23). Los parámetros estimados por Máxima Verosimilitud para cada serie de las divisiones se pueden ver en el Cuadro (6.1). Como se puede ver las varianzas de los errores de la pendiente salieron igual a cero para Lima Metropolitana y Lima Norte, también la varianza del error de la tendencia resultó igual a cero para Lima Centro, por otro parte, como ya hemos indicado para Lima Sur y El Callao se muestra igual a cero porque no se consideraron para el ajuste del modelo.

	σ^2_{ξ}	σ^2_{ζ}	σ^2_{w}
Lima Metropolitana	0.000112	0	0.00000103
Lima Norte	0.000081	0	0.00002
Lima Este	0.0000047	0. 0302	0.0001
Lima Centro	0	0.0045	0.01015
Lima Sur	0	0	0
El Callao	0	0	0

Cuadro 6.1: Parámetros estimados por Máxima Verosimilitud para la tasa de desempleo de Lima Metropolitana y sus divisiones.

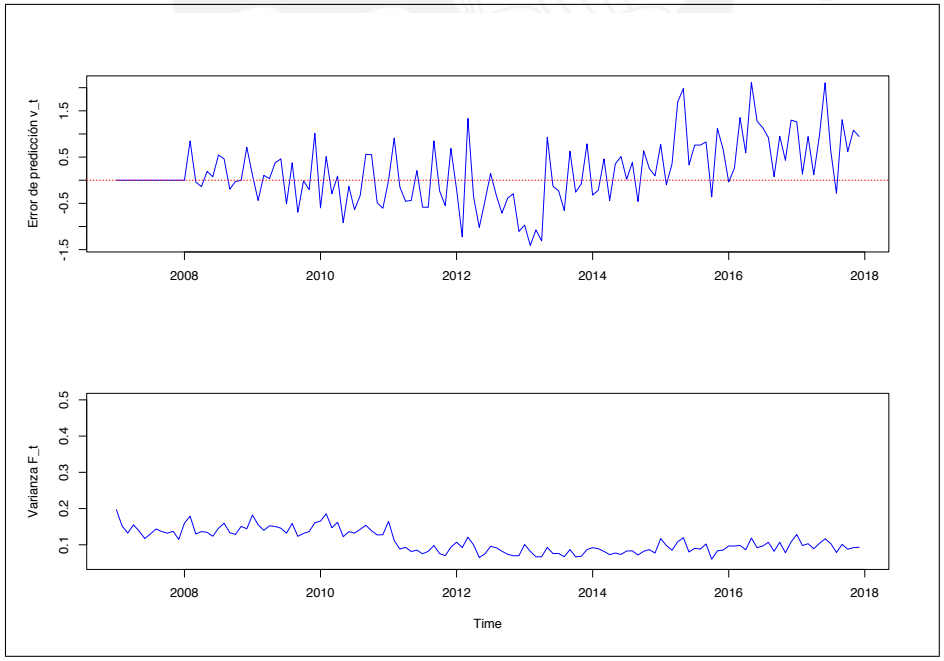


Figura 6.4: Errores de predicción (arriba) y sus varianzas (abajo) para la tasa de desempleo de Lima Metropolitana (2007-2017).

Los errores de predicción υ_{it} y sus varianzas juegan un rol importante en la maximización del logaritmo de la función de verosimilitud, entonces sean $\upsilon_{it} = y_{it} - Z_t\hat{\alpha}_{it}$ y

$F_{it} = Var(y_{it}|Y_{i,t-1})$, los errores de predicción cuantifican la falta de precisión de $\hat{\alpha}_t$ en la predicción del valor observado y_t en cada tiempo t, estos errores también se denominan innovaciones porque aportan nueva información, lo que permite que el sistema en estudio se adapte a la nueva información entrante, por ejemplo, en la Figura (6.4) se muestran todos los errores de predicción υ_t y la varianza F_t obtenidos con el filtro de Kalman para la tasa de desempleo de Lima Metropolitana, donde se observa que los errores de predicción oscilan al rededor de cero y las varianzas decrecen en el tiempo.

6.2.2. Considerando los errores correlacionados en el modelo

Hasta el punto desarrollado en la sección anterior, el modelo de espacio de estados planteado en (6.1) y (6.2) ya estaría completo para la estimación de los vectores de estados aplicando el filtro de Kalman. Sin embargo, las series de las tasas de desempleo como se ha descrito al inicio del capítulo tienen errores correlacionados, entonces para que sea posible aplicar el algoritmo de *Pfeffermann y Tiller* que supone considera los errores correlacionados como se detallo en (3.6), es necesario conocer cuales son esos errores correlacionados $E(e_{i\tau}e'_{it}) = \Sigma_{i\tau t}$. Entonces para esto, de acuerdo al algoritmo de *Pfeffermann y Tiller* los vectores de estados α_{it} y sus matrices de covarianzas P_{it} se obtienen de las siguientes expresiones:

$$\begin{aligned} \hat{\alpha}_{it} &= \left[(I, Z'_{it})V_{it}^{-1}\begin{pmatrix} I \\ Z_{it}\end{pmatrix}\right]^{-1}(I, Z'_{it})V_{it}^{-1}\begin{pmatrix} T\hat{\alpha}_{i,t-1} \\ y_{it}\end{pmatrix}, \\ P_{it} &= E[(\hat{\alpha}_{it} - \alpha_{it})(\hat{\alpha}_{it} - \alpha_{it})'] = \left[(I, Z'_{it})V_{it}^{-1}\begin{pmatrix} I \\ Z_{it}\end{pmatrix}\right]^{-1}, \end{aligned} \tag{6.5}$$

donde I es una matriz identidad ($I_{13\times 13}$). En (6.5) la matriz de covarianzas de los errores V_{it} es justamente la que considera los errores correlacionados y se expresa de la siguiente forma:

$$V_{it} = var\begin{pmatrix} u_{i,t|t-1} \\ e_{it}\end{pmatrix} = \begin{pmatrix} P_{i,t|t-1} & C_{it} \\ C'_{it} & \Sigma_{itt}\end{pmatrix}$$

donde la matriz $P_{i,t|t-1} = TP_{i,t-1}T' + Q_i$, y Σ_{itt} es la varianza en la división i en el tiempo t, en tanto que las covarianzas C_{it} se definen como $C_{it} = cov[T\hat{\alpha}_{i,t-1} - \alpha_{it}, e_{it}]$, el procedimiento para su cálculo es un poco complejo y se detalla a continuación, en base a la siguiente igualdad extraída de (6.5):

$[I, Z'_{ij}]V_{ij}^{-1} = [B_{ij1}, B_{ij2}]_{13\times 14}$, donde

B_{ij1} contiene las primeras 13 columnas de $[B_{ij1}, B_{ij2}]_{13\times 14}$ y

B_{ij2} contiene el resto de las columnas.

Entonces sean:

$A_{ij} = TP_{ij}B_{ij1}$, $\tilde{A}_{ij} = TP_{ij}B_{ij2}$, $j = 2, 3, ..., t-1$, $\tilde{A}_{i1} = TK_{i1}$, donde

$K_{i1} = P_{i,1|0}Z'_{it}F_{i1}^{-1}$, conocido como ganancia de Kalman

$P_{i,1|0} = TP_{i0}T' + Q_i$, es la matriz de covarianzas del error de predicción, y

$F_{i1} = Z_{it}P_{i,1|0}Z'_{it} + \Sigma_{i11}$, es la matriz de covarianzas de las innovaciones.

Entonces, la matriz C_{it} queda definido como:

$$C_{it} = A_{i,t-1}A_{i,t-2}...A_{i,t-11}\tilde{A}_{i.t-12}\Sigma_{i,t-12,t} + A_{i,t-1}A_{i,t-2}...A_{i,t-8}\tilde{A}_{i,t-9,t}\Sigma_{i,t-9,t} +$$
$$+ A_{i,t-1}A_{i,t-2}A_{i,t-3}A_{i,t-4}A_{i,t-5}\tilde{A}_{i,t-6}\Sigma_{i,t-6,t} + A_{i,t-1}A_{i,t-2}\tilde{A}_{i,t-3}\Sigma_{i,t-3,t}$$

donde las correlaciones de orden 3, 6, 9 y 12 estarán en; $\Sigma_{i,t-3,t} = corr(e_{i,t-3}, e_{it})\sqrt{var(e_{i,t-3})var(e_{it})}$, aquí $corr(e_{i,t-3}, e_{it})$ indica la correlación cuando la vivienda participa en la muestra en el mes t y 3 meses después vuelve a participar. $\Sigma_{i,t-6,t} = corr(e_{i,t-6}, e_{it})\sqrt{var(e_{i,t-6})var(e_{it})}$ donde $corr(e_{i,t-6}, e_{it})$ indica la correlación cuando la vivienda participa en la muestra en el mes t y 6 meses después vuelve nuevamente a participar en la muestra, de forma similar serán para $\Sigma_{i,t-9,t} = corr(e_{i,t-9}, e_{it})\sqrt{var(e_{i,t-9})var(e_{it})}$ y $\Sigma_{i,t-12,t} = corr(e_{i,t-12}, e_{it})\sqrt{var(e_{i,t-12})var(e_{it})}$.

Estas correlaciones son de acuerdo a lo encontrado en la exploración de la base de datos, donde el orden de las correlaciones obedece a la participación de las viviendas en el panel detallado al inicio de este capítulo. Las correlaciones estimadas se pueden ver en el Cuadro (6.2).

	$corr(e_{t-3}, e_t)$	$corr(e_{t-6}, e_t)$	$corr(e_{t-9}, e_t)$	$corr(e_{t-12}, e_t)$
Lima Metropolitana	0	0.50229	0	0
Lima Norte	0	0	0.43004	0.35182
Lima Este	0	0.31729	0.00888	0
Lima Centro	0.24909	0	0	0
Lima Sur	0	0	0	0
El Callao	0	0	0	0

Cuadro 6.2: Correlaciones de los errores muestrales de la tasa de desempleo de Lima Metropolitana y sus divisiones.

6.3. Análisis de las proyecciones estimadas

Con los componentes estimados del modelo estructural básico utilizado para modelar la tasa de desempleo de Lima Metropolitana, se puede realizar un análisis desagregado por cada componente. En la Figura (6.5) se muestra la gráfica de los componentes estimados de la serie: la tendencia, los efectos de estacionalidad y el término irregular estimados mediante

los algoritmos del filtro de Kalman y de Pfeffermann y Tiller.

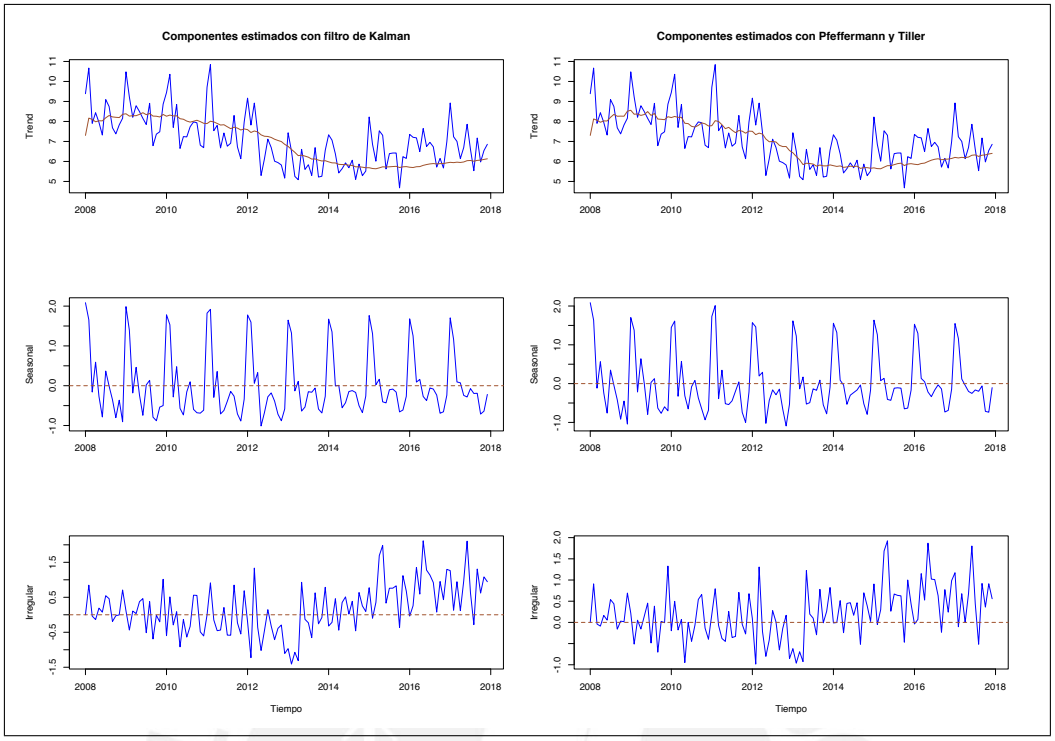


Figura 6.5: Componentes de la tasa de desempleo de Lima Metropolitana (2008 - 2017): tendencia, estacionalidad y el término irregular estimados mediante algoritmos del filtro de Kalman y de Pfeffermann y Tiller.

Las tendencias estimadas por ambos algoritmos que se muestran en la Figura (6.5) son similares, la tendencia nos muestra los movimientos subyacentes de la tasa de desempleo de Lima Metropolitana ocurridos desde el 2008 hasta el 2017. Para tener una mejor apreciación en la Figura (6.6) se muestran las tendencias estimadas, se puede ver que en general la tendencia es decreciente en el tiempo, sobre todo entre los años 2009 y 2014, esto puede estar relacionado a que la productividad laboral (medida como el cociente entre la PEA y el Producto Bruto Interno) mantuvo un ritmo de crecimiento de 3.5 % por año desde el 2009 hasta el 2014, según el informe anual del empleo para el 2014 del Ministerio de Trabajo y Promoción del Empleo, esto implicó que la demanda laboral se incremente en ese periodo y por ende la tasa de desempleo habría decrecido. Desde el 2015 en adelante al parecer hay una estabilización donde las tasas fluctúan alrededor de 6 puntos porcentuales, sin embargo, la tendencia estimada con el algoritmo de Pfeffermann y Tiller muestra un incremento más pronunciado que el estimado con el filtro de Kalman.

Respecto del componente de estacionalidad, la Figura (6.5) muestra claramente que la serie de la tasa de desempleo tiene efectos de estacionalidad, recordemos que la tasa de desempleo en estudio, mide el desempleo urbano que se genera al interior de los distritos de Lima Metropolitana, entonces es bastante lógico que la dinámica del mercado laboral sea en función a la demanda laboral de las empresas, en este sentido tal como se ha señalado en la descripción de datos, las tasas de desempleo son mayores en los primeros trimestres de cada

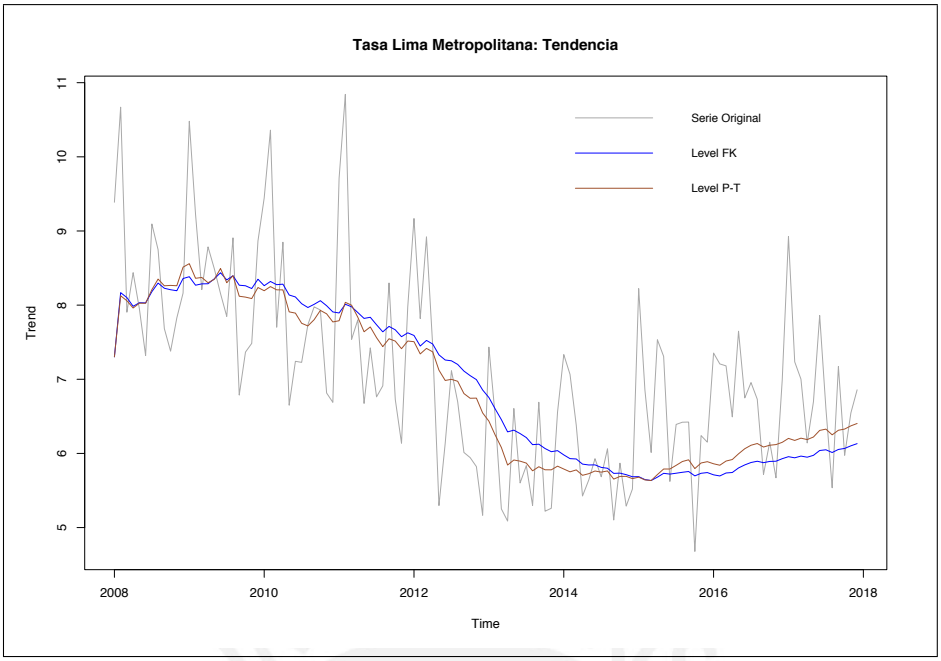


Figura 6.6: Tendencia (Level) de la tasa de desempleo de Lima Metropolitana (2008-2017), estimados mediante algoritmos del filtro de Kalman (FK) y de Pfeffermann y Tiller (P-T).

año y esto coincide con el periodo de vacaciones escolares y universitarios, por tanto una baja demanda laboral. También se ha señalado que las menores proporciones de desempleo corresponden a los últimos trimestres de cada año, coincidentemente las campañas navideñas empiezan 3 ó 4 meses antes, entonces la tasa de desempleo fluctuaría durante el año en función de los planes de producción de las empresas que también operan por estaciones o campañas, esto se estaría reflejando el los gráficos de estacionalidad.

En la Figura (6.7) se muestra las predicciones esperadas de la tasa de desempleo de Lima Metropolitana $\hat{Y}_{lima,t} = \hat{\mu}_{lima,t} + \hat{\gamma}_{lima,t}$ estimadas por ambos algoritmos, comparadas con la serie original se puede la existencia de brechas con las proyecciones esperadas, esta falta de precisión podría ser explicado por el error dado que las proyecciones esperadas no consideran estos errores.

Las predicciones de la serie original $\hat{Y}_{lima,t} = \hat{\mu}_{lima,t} + \hat{\gamma}_{lima,t} + \hat{e}_{lima,t}$ para la tasa de desempleo de Lima Metropolitana se muestran en la Figura (6.8) donde se considera los errores de muestreo estimados. Para tener una mejor visualización de las diferencias se gráfica para los 2 últimos años, en la gráfica se han considerado la serie original como punto de referencia, las predicciones esperadas tanto para los estimados con el filtro de Kalman como para los estimados con el algoritmo de Pfeffermann y Tiller, y las predicciones de la serie considerando los errores estimados con el último algoritmo en referencia. Se puede indicar que, las estimaciones más próximas al verdadero valor de la serie son aquellas estimaciones obtenidas con el algoritmo de Pfeffermann y Tiller que considera los errores estimados, entonces, cuando el modelo de espacio de estados considera los errores muestrales correlacionados para series de observaciones que cuentan con errores de medición la mejor estimación se obtiene con el algoritmo propuesto por Pfeffermann y Tiller, esto es consecuente con los resultados

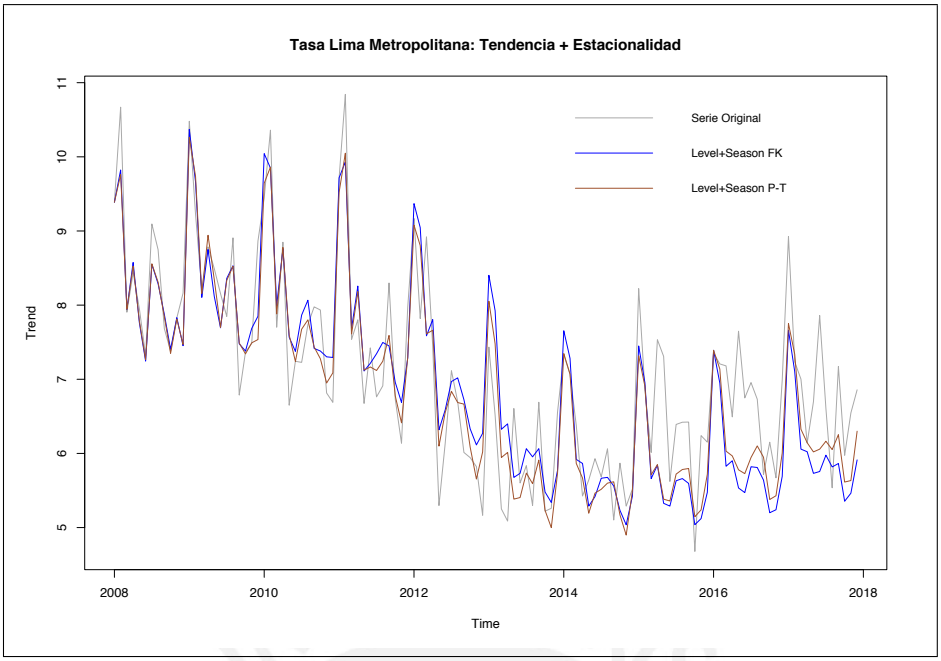


Figura 6.7: Tendencia + Estacionalidad de la tasa de desempleo de Lima Metropolitana (2008-2017), estimados mediante algoritmos del filtro de Kalman (FK) y de Pfeffermann y Tiller (P-T).

obtenidos en el capítulo de simulaciones, donde se llega a la misma conclusión.

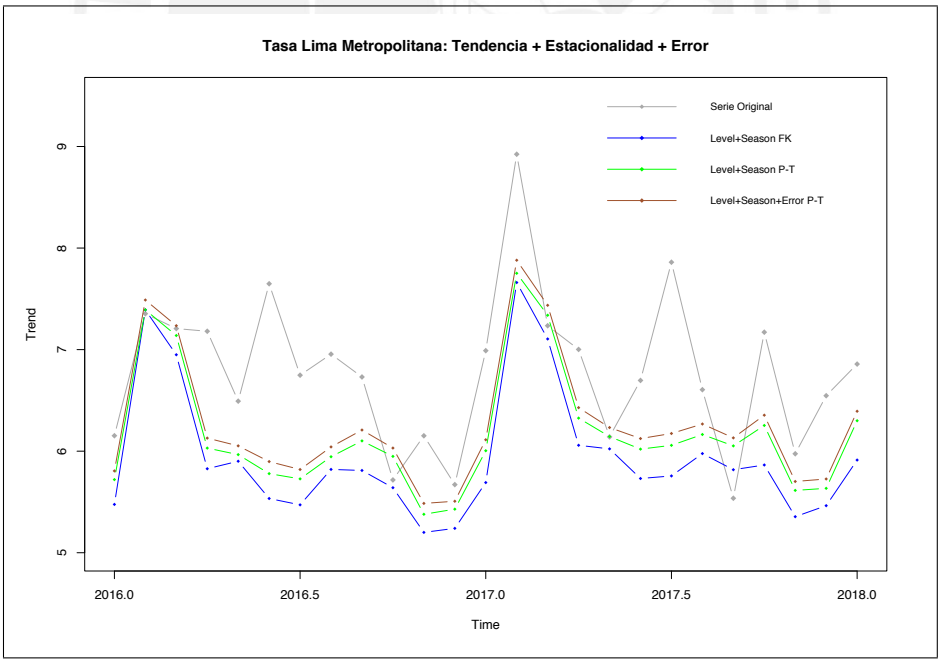


Figura 6.8: Tendencia + Estacionalidad + Error de la tasa de desempleo de Lima Metropolitana (2016-2017), estimados mediante algoritmos del filtro de Kalman (FK) y de Pfeffermann y Tiller (P-T).

Como se ha indicado al inicio del presente capítulo, adicionalmente a la tasa de desempleo de Lima Metropolitana se ha estimado también los componentes para la tasa de desempleo de Lima Norte, Lima Este y Lima Centro. En la Figura (6.9) se muestra las tendencias

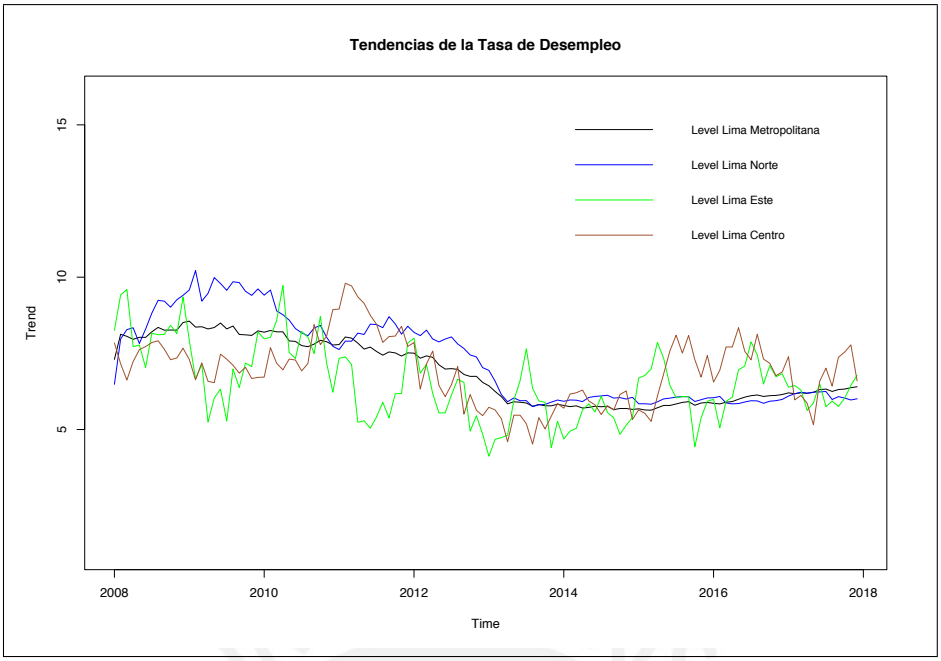


Figura 6.9: Tendencias estimadas de la tasa de desempleo de Lima Metropolitana, Lima Norte, Lima Este y Lima Centro (2008-2017).

estimadas para cada una de las divisiones indicadas, lo que resalta a simple vista es que las tendencias no son iguales, la tendencia que más se aproxima a la de Lima Metropolitana es la tendencia de Lima Norte, los otros parecen indicar señales de quiebre estructural o ausencia de información.

Para completar la comparación de la dinámica laboral, en la Figura (6.10) se tiene la gráfica de los efectos de estacionalidad estimadas para cada división, si bien los efectos estacionales parecen tener picos elevados anualmente, la dinámica laboral es diferente lo que ameritaría estudios focalizados.

6.4. Modelo ARIMA para la tasa de desempleo

Con base a la teoría desarrollada en el Capítulo (2), en esta sección la idea es encontrar el modelo ARIMA subyacente de la tasa de desempleo de Lima Metropolitana desde el enfoque de Box and Jenkins, considerando los pasos de identificación, estimación y diagnóstico, con la finalidad de poder comparar con lo obtenido mediante el modelo de espacio de estados. En este enfoque, en lugar de modelar separadamente los componentes del modelo estructural básico, el objetivo es eliminar la tendencia y los efectos de estacionalidad mediante la diferenciación de la serie, la serie diferenciada es tratada como una serie de tiempo estacionaria con media, varianza y covarianza constante en el tiempo, tal como fue definido en (2.4.1).

Como hemos visto en la sección de descripción de datos, la serie de observaciones de la tasa de desempleo no tiene características de una serie estacionaria, presenta tendencia por lo que el primer objetivo es eliminar la tendencia buscando el orden correcto de p, q y d del modelo ARIMA(p,d,q), adicionalmente también, al parecer la serie presenta estacionalidad,

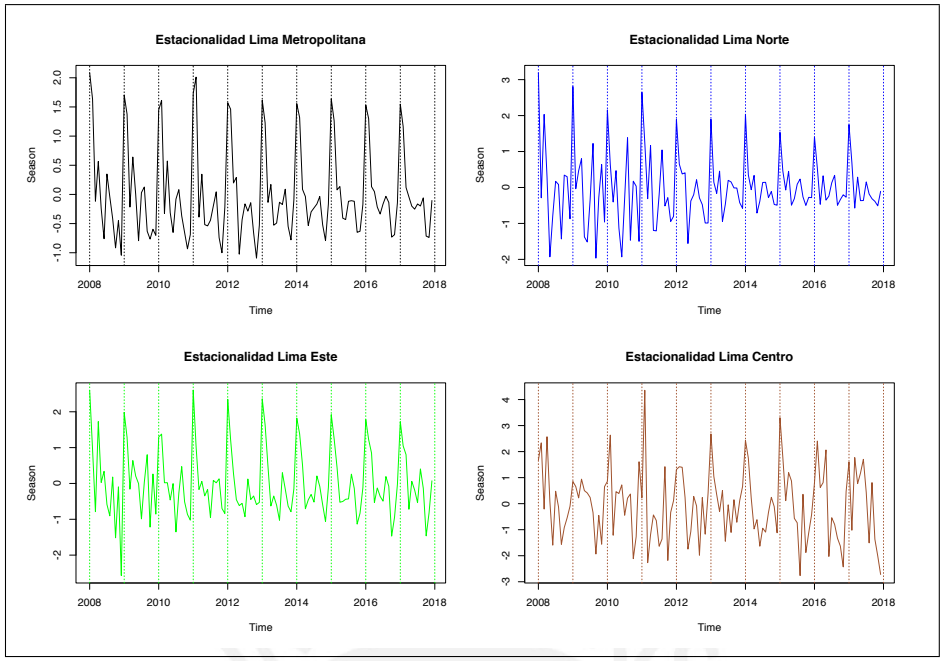


Figura 6.10: Efectos de estacionalidad estimados de la tasa de desempleo de Lima Metropolitana, Lima Norte, Lima Este y Lima Centro (2008-2017).

por tanto el otro objetivo será eliminar los efectos de estacionalidad en caso corresponda, identificando los valores correctos de P, Q y D del modelo SARIMA(P,D,Q), por tanto, para modelar adecuadamente la serie de la tasa de desempleo que tiene tendencia y estacionalidad será adecuado usar la combinación de ambos modelos, como fue expuesto en (2.6) y (2.7). Este modelo es denotado por $SARIMA(p,d,q)(P,D,Q)_s$ y se expresa como:

$$(1-B^s)^D(1-B)^d\phi_p(B)\Phi_P(B^s)y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t,$$

donde la serie de la tasa de desempleo y_t es mensual (s=12), p y P son el orden de la parte autorregresiva de tendencia y estacionalidad respectivamente, de forma similar, q y Q son el orden de la parte de medias móviles, d será el número de diferencias tomadas para eliminar la tendencia y D será el número de diferencias tomadas para eliminar los efectos de estacionalidad. Entonces, aplicando una diferencia a la tendencia $d = 1$ de la serie original de la tasa de desempleo, al parecer es suficiente para conseguir la estacionariedad de la serie, en la Figura (6.11) se visualiza la gráfica de la serie transformada y su función de autocorrelación simple (ACF) y parcial (PACF). A partir de la visualización de las autocorrelaciones podemos buscar el mejor ajuste del modelo para la tasa de desempleo, probando diferentes valores para q y Q, también se ha probado para diferentes combinaciones de p, P y D. En este sentido, se ha iniciado la busqueda de los modelos que mejor se ajustan a los datos, partiendo del modelo más simple a lo más complejo, de este procedimiento los candidatos seleccionados se muestran en el Cuadro (6.3).

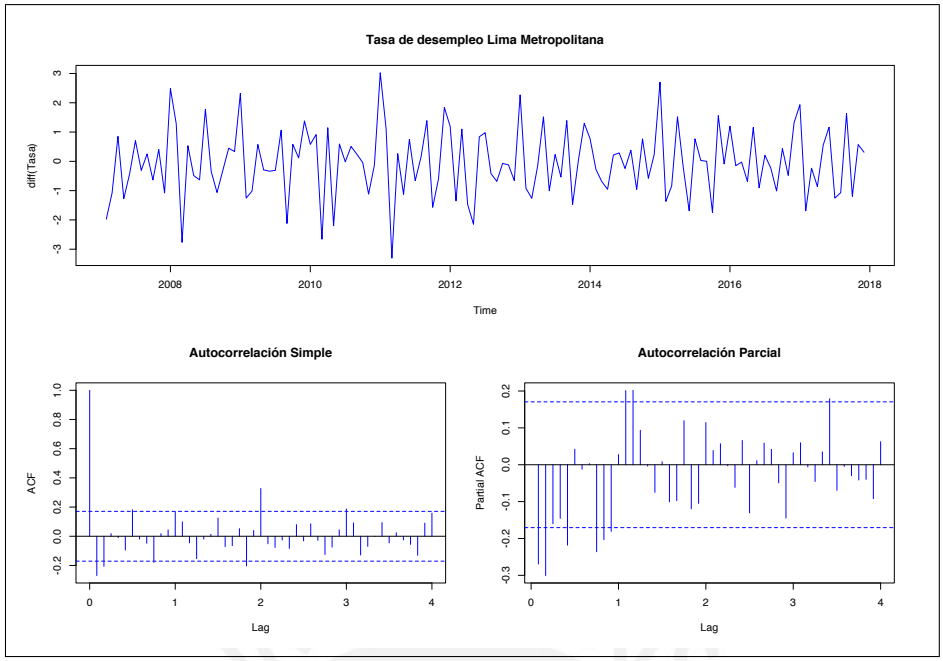


Figura 6.11: Serie de la tasa de desempleo de Lima Metropolitana (2007-2017) con una diferencia en tendencia (arriba) y la Función de Autocorrelación Simple y Parcial (abajo).

Modelo	Dickey-Fuller Test	Box-Pierce Test	AIC
SARIMA(0,1,1)(0,0,2)	0.01	0.43	355.1975
SARIMA(0,1,2)(0,0,2)	0.01	0.70	354.0131
SARIMA(0,1,2)(0,1,0)	0.01	0.0000002	354.9794
SARIMA(0,1,1)(0,1,0)	0.04	0.000000008	355.3524

Cuadro 6.3: Modelos SARIMA(p,d,q)(P,D,Q) ajustados a la tasa de desempleo de Lima Metropolitana, prueba de estacionariedad de residuales (Dickey-Fuller Test), prueba de independencia de residuales (Box-Pierce Test) y Criterio de Información de Akaike (AIC).

Los resultados del Cuadro (6.3) indican que todos los modelos son estacionarios, en la prueba de Dickey-Fuller en todos lo casos el $p-valor < 0.05$, entonces se rechaza la H_0 (H_0 : Los residuales del modelo son estacionarios), esto indica que haciendo una diferencia en la tendencia la serie es estacionaria, pero también si se agrega una diferencia en estacionalidad es estacionaria. Luego se realizó la prueba de independencia de residuales con la prueba de Box-Pierce donde el $p - valor \geq 0.05$ para los 2 primeros modelos, entonces no se rechaza la H_0 (H_0 : Los residuales del modelo son independientes) por tanto los 2 últimos modelos indicarían falta de ajuste, adicionalmente mirando el Criterio de Información de Akaike (AIC) indican que el segundo y el tercer modelo tienen menores valores, complementando el análisis también se determinado la significancia de los parámetros de los modelos que se pueden ver en el Cuadro (6.4).

Modelo	ma1	ma2	sma1	sma2
SARIMA(0,1,1)(0,0,2)	16.98	–	2.88	4.20
SARIMA(0,1,2)(0,0,2)	6.97	1.87	2.39	4.10
SARIMA(0,1,2)(0,1,0)	10.76	1.92	–	–
SARIMA(0,1,1)(0,1,0)	20.70	–	–	–

Cuadro 6.4: Modelos SARIMA(p,d,q)(P,D,Q) ajustados a la tasa de desempleo de Lima Metropolitana, con prueba de significancia de los términos estimados de Medias Móviles (ma1, ma2, sma1, sma2).

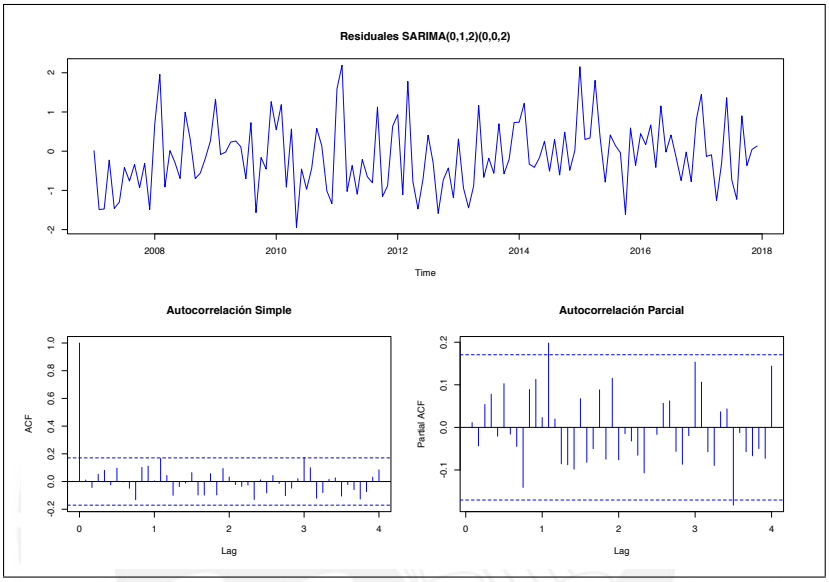


Figura 6.12: Residuales del modelo seleccionado $SARIMA(0,1,2)(0,0,2)_{12}$ y sus respectivas autocorrelaciones.

En resumen del Cuadro (6.3) el modelo seleccionado sería $SARIMA(0,1,2)(0,0,2)_{12}$ bajo los siguientes criterios, porque es el modelo con residuales de procedencia aleatoria y que tiene el menor valor de AIC que los otros modelos. Sin embargo, en el Cuadro (6.4) uno de sus parámetros $ma2$ es no significativo (para $\alpha = 0.05$) ($|\frac{\hat{\theta}}{DE(\hat{\theta})}| = 1.87 < 1.96$), esto implicaría, no considerar en el modelado el parámetro $ma2$, pero el primer modelo $SARIMA(0,1,1)(0,0,2)_{12}$ considera esta opción y su AIC resulta mayor, por tanto, el modelo que mejor se ajustaría a los datos de la tasa de desempleo en este análisis sería $SARIMA(0,1,2)(0,0,2)_{12}$, cuya gráfica de los residuales podemos observar en la Figura (6.12).

6.5. Comparación de modelos de la tasa de desempleo

En esta última sección vamos a realizar la comparación entre los modelos de espacio de estados y los modelos ARIMA con los cuales se ha ajustado la tasa de desempleo de Lima Metropolitana. Partiremos indicando que ambos enfoques considera que una serie está compuesto por componentes de tendencia, efectos estacionales y el término irregular, sin embargo, son 2 enfoques totalmente distintos, la metodología propuesta por Box and Jenkins (1970) en lugar de modelar por separado cada componente, su idea es eliminar la tendencia

y estacionalidad haciendo diferencias al inicio del análisis, las series diferenciadas se tratan como series estacionarias cuya propiedad es mantener la media, la varianza y la covarianza invariables en el tiempo, en este proceso la información de la tendencia y estacionalidad se pierde, no es necesario tener un modelo detrás de los datos ya que justamente con los datos, usando los pasos de diagnóstico y ajuste de modelos ARIMA se aproxima al modelo deseado el cual obtenga los mejores pronósticos.

Mientras que para el uso de los modelos de espacio de estados se requiere tener un modelo razonable detrás de los datos, ya que cada componente se modela por separado. La ventaja de esto, es que puedes hacer un análisis por separado de los componentes manteniendo la estructura original de la serie sin necesidad de que sean estacionarios, es decir, puedes observar la tendencia y ver los cambios generados en el tiempo o ver si los datos presentan efectos de estacionalidad. Conocer el modelo de cada componente permite usar conocimientos previos para mejorar el proceso y obtener mejores estimaciones, pero también puede volverse complejo ya que introduce estados no observables en el proceso, en todo caso, cualquier modelo ARIMA puede ser representado en la forma de espacio de estados pero sólo algunos modelos de espacio de estados pueden ser representados en la forma de un modelo ARIMA.

Para poder comparar los modelos ajustados a la tasa de desempleo de Lima Metropolitana, se ha obtenido algunos errores de predicción, como el error medio (EM), el error absoluto medio (EAM) y el error cuadrático medio (ECM) utilizados en el capítulo de simulación, con los resultados se ha elaborado el Cuadro (6.5) donde se aprecia los errores de predicción para 12 meses (2017) obtenidos para cada modelo, de parte de los modelo ARIMA están los 2 mejores modelos ajustados en la sección anterior, SARIMA(0,1,2)(0,0,2) y SARIMA(0,1,1)(0,0,2), para los modelos de espacio de estados están los estimados con; el filtro de Kalman (FK) el cual no considera los errores correlacionados de muestreo de la tasa de desempleo y con el algoritmo de Pfeffermann y Tiller (P-T) el cual sí considera los errores correlacionados.

	FK	SARIMA(0,1,2)(0,0,2)	SARIMA(0,1,1)(0,0,2)	P-T
EM	-0.82	-0.002	-0.03	-0.48
EAM	0.87	0.61	0.63	0.62
ECM	1.04	0.63	0.67	0.56

Cuadro 6.5: Tabla de errores de predicción de la tasa de desempleo de Lima Metropolitana (2017), para modelos estimados con el algoritmo del Filtro de Kalman (FK), con el algoritmo de Pfeffermann y Tiller (P-T) y con modelos ARIMA.

La lectura de los resultados indica que entre los modelos ARIMA, el que mejor se ajusta a los datos de la tasa de desempleo es el modelo SARIMA(0,1,2)(0,0,2) que presenta menores errores y confirma lo concluido en la sección anterior. Mientras que para los modelos de espacio de estados, como ya hemos visto anteriormente, los menores errores presenta el modelo estimado con el algoritmo de Pfeffermann y Tiller que considera los errores correlacionados, ya que las observaciones de la tasa de desempleo cuentan con errores de medición como fue descrito al inicio de este capítulo. Ahora comparando el modelo ARIMA que mejor se ajusta a los datos con el modelo de espacio de estados con errores correlacionados hay discrepancias,

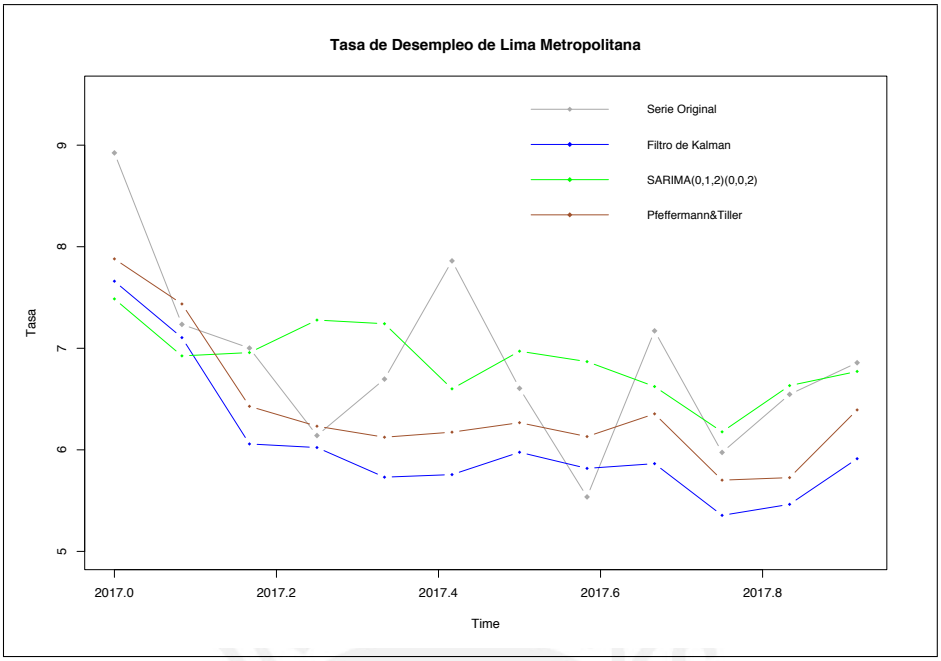


Figura 6.13: Predicciones de la tasa de desempleo de Lima Metropolitana para 2017, obtenidas con los algoritmos del filtro de Kalman, de Pfeffermann y Tiller y con un modelo ARIMA

el modelo ARIMA presenta menor error medio y menor error absoluto medio, sin embargo, el modelo que tiene menor error cuadrático medio es el estimado con el algoritmo de Pfeffermann y Tiller. Basándonos estrictamente en los resultados del cuadro, se podría indicar que el modelo que mejor se ajusta a los datos de la tasa de desempleo de Lima Metropolitana es el modelo estimado con el algoritmo de Pfeffermann y Tiller, ya que en error absoluto medio son similares (0.61 y 0.62) y al tener menor error cuadrático medio (0.56) sería el modelo más eficiente para estimar los componentes del modelo porque este error incluye la varianza y el sesgo.

En la Figura (6.13) se puede observar las predicciones realizadas para el 2017 de la tasa de desempleo de Lima Metropolitana comparadas con la serie original. Efectivamente las predicciones hechas con el modelo SARIMA(0,1,2)(0,0,2) y el modelo de espacio de estados estimado con el algoritmo de Pfeffermann y Tiller son las que más se aproximan a los datos reales. Sin embargo, es importante resaltar que el modelo de espacio de estados con errores correlacionados se podría aún mejorar superando las limitaciones que se tuvo en el estudio, es decir, teniendo acceso al código de ubigeo de las encuestas y conociendo más a profundidad los cambios realizados en el plan de rotación de la muestra de la Encuesta Permanente del Empleo durante el periodo de referencia del estudio. Esto podría mejorar el ajuste al modelo ya que se encuentra directamente relacionado con el orden de las correlaciones.

Capítulo 7

Conclusiones

El modelo de espacio de estados de este trabajo está basado en un modelo estructural básico que está formado por componentes según la descomposición clásica de series de tiempo (Harvey, 1989). Estos son: la tendencia que es el componente que refleja los cambios que la serie a tenido en el tiempo y está expresado en función de la pendiente más un ruido blanco, a su vez la pendiente es expresado como un paseo aleatorio que cambia en el tiempo; mientras que el componente de estacionalidad normalmente requiere $(s-1)$ ecuaciones donde s es la periodicidad de la serie, este componente extrae el patrón recurrente de la serie sea cual sea la periodicidad de los datos; semanales, mensuales, trimestrales, anuales, etc. por lo tanto la suma de la tendencia, la estacionalidad más el componente irregular forman el modelo estructural básico. Este modelo de espacio de estados puede tener dos enfoques, una cuando los errores del modelo se consideran independientes y otra cuando estos errores están autocorrelacionados, para el primer caso lo más frecuente es aplicar el algoritmo recursivo del filtro de Kalman para el proceso de estimación, mientras que para el segundo caso se requiere un tratamiento diferenciado donde es necesario considerar los errores correlacionados en el proceso de estimación, en estos casos se aplica el algoritmo recursivo de Pfeffermann y Tiller (Pfeffermann and Tiller, 2006).

El objetivo del presente trabajo fue fundamentar y aplicar el modelo de espacio de estados con errores correlacionados a la tasa de desempleo de Perú. Este indicador es publicado por el Instituto Nacional de Estadística e Informática - INEI de forma mensual, cuya información es recolectada por la Encuesta Permanente del Empleo - EPE y que únicamente es representativa para Lima Metropolitana. Se consideró este indicador por la relevancia que tiene tanto en el ámbito laboral como en lo económico, dado que simplifica en una cifra lo que ocurre en la oferta y demanda del mercado laboral y refleja la capacidad de la economía de un país para absorber el empleo. Tal es así, que es incluido por la Organización Internacional del Trabajo - OIT en su informe anual del Panorama Laboral para América Latina. La EPE inicio su funcionamiento en el 2001, desde entonces hasta la fecha viene recolectando la información de manera continua y es en función a esta información que se buscó aplicar el modelo mencionado.

Para considerar los errores correlacionados en el modelo se requieria conocer el orden de las correlaciones, es decir, saber cuantas veces una vivienda participó en la muestra de panel, por cuanto tiempo, cuantas viviendas eran nuevas y cuantas antiguas que ya venían participando

en el panel en cada recojo de información. Para esto se realizó la exploración documentaría con la finalidad de saber cual fue el diseño muestral utilizado por la EPE, específicamente para conocer el esquema de rotación utilizado para la rotación de las muestras en el panel, se encontró que el esquema de rotación, el tamaño muestral y otros componentes del diseño muestral habían tenido cambios en diferentes puntos del tiempo desde sus inicios hasta la fecha, por esta razón se trabajó con información comprendida entre los años de 2007 al 2017, debido a que fue el periodo identificado donde ocurrió menos cambios.

Previamente al ajuste del modelo, se realizó un estudio de simulación con la finalidad de comparar la precisión de las estimaciones entre los dos algoritmos utilizados en este trabajo, para lo cual se generaron 5,000 series de tiempo con errores correlacionados. Se encontró que para las predicciones de las observaciones originales ofrece una mayor precisión las realizadas con el algoritmo de Pfeffermann y Tiller que las obtenidas con el algoritmo del filtro de Kalman, por ejemplo, para $t = 75$ los errores de predicción fueron ($EM = -0.026$, $EAM = 1.176$, $ECM = 2.161$) para filtro de Kalman y ($EM = -0.025$, $EAM = 1.099$, $ECM = 1.895$) para Pfefferman y Tiller, en todos los casos los menores errores corresponden a este último. Entonces se coincide con lo indicado por Pfeffermann y Tiller (2006) que el algoritmo es aplicable para cualquier modelo de espacio de estados con errores de medición correlacionados y ofrece estimaciones más precisas en estos escenarios.

Para la aplicación del modelo se dividió el área total de Lima Metropolitana en 5 divisiones, Lima Norte, Lima Este, Lima Centro, Lima Sur y El Callao de las cuales para el proceso de comparación sólo se uso las 3 primeras divisiones. Para cada uno de estas divisiones y para el área total se calculo sus respectivos parámetros y correlaciones, con estos insumos se estimó los componentes de la serie de la tasa de desempleo en cada división y para Lima Metropolitana aplicando los algoritmos del filtro de Kalman y de Pfeffermann y Tiller. El resultado coincidió con los del estudio de simulación, las predicciones más cercanas al valor real de la tasa de desempleo pertenecieron a los obtenidos con el algoritmo de Pfeffermann y Tiller.

Finalmente, se ajustó los datos de la tasa de desempleo de Lima Metropolitana a un modelo ARIMA mediante el enfoque propuesto por Box and Jenkins, cuyo resultado fue que los datos se ajustaron mejor al modelo $SARIMA(0,1,2)(0,0,2)$, con este modelo se realizó las predicciones y se comparo con las predicciones obtenidas aplicando el algoritmo de Pfeffermann y Tiller. Se encontró que la precisión de ambas predicciones son similares comparadas con el verdadero valor de la serie, sin embargo, se observo un menor $ECM = 0.56$ para las predicciones realizadas con el algoritmo de Pfeffermann y Tiller que para las predicciones hechas con el modelo ARIMA cuyo $ECM = 0.67$, resaltando que las predicciones realizadas con el algoritmo de Pfeffermann y Tiller son mejorables aún, conociendo en mayor detalle el diseño muestral de la EPE. El contraste entre ambos enfoques ha sido interesante, dado que, para los modelos de espacio de estados es necesario partir teniendo una idea del modelo detrás de los datos, mientras que para los modelos ARIMA no es necesario partir de un modelo debido a que se enfoca en el modelo que mejor se ajuste a los datos. Los modelos ARIMA necesitan que las series sean estacionarias, mientras que para los modelos de espacio

de estados esto no es necesario, mantienen la estructura original de la serie lo que permite hacer un análisis por separado de cada componente, en cambio esto no es posible con los modelos ARIMA. Por último, todo modelo ARIMA puede ser expresado como un modelo de espacio de estados mientras que sólo algunos modelos de espacio de estados pueden ser expresados como un modelo ARIMA.

7.1. Sugerencias para investigaciones futuras

- Un aprendizaje importante que se presentó en el desarrollo de este trabajo y puede servir para futuros estudios, es que si se piensa trabajar con datos de las encuestas del INEI, antes de iniciar el estudio verificar que los datos esenciales para su estudio estén incluidos en la base de datos, por ejemplo, para el propósito de este trabajo era esencial que los datos contengan el código de ubigeo, sin embargo, esto no fue así y dificultó el alcance de los objetivos iniciales de obtener indicadores de la tasa de desempleo a nivel distrital, este estudio puede ser replicado si se liberan estos datos.

- Para aplicar específicamente los modelos de espacio de estados con errores correlacionados a datos longitudinales, se requiere el conocimiento del orden de las correlaciones cuya información es obtenida del esquema de rotación de la muestra de panel, este esquema en el transcurso del tiempo puede haber tenido cambios por diferentes razones, se recomienda identificar el periodo de tiempo donde el esquema no haya tenido cambios para aplicar el modelo de manera uniforme.

- Dentro del estudio de series de tiempo, si el objetivo es únicamente realizar pronósticos los modelos ARIMA son una gran alternativa, y si adicionalmente a realizar pronósticos el objetivo fuese analizar la tendencia histórica e identificar los patrones de estacionalidad, los modelos de espacio de estados son la mejor opción, además estos cuentan con una amplia variedad de modelos que permiten el tratamiento de diferentes tipos de datos.

Apéndice A

Códigos de R

A.1. Simulación

```
#############################################################################
## Simulación de series de tiempo con errores correlacionados
#############################################################################

set.seed(779)
n = 300 # número de observaciones a simular
N = 5000 # número de simulaciones
mu_e = c(0,0,0,0,0,0,0,0,0,0,0,0,0)
Be_sim = array(0, c(13,n,N))
Ye_sim = matrix(0,n,N) # Matriz que acumula las series simuladas

## Errores correlacionados
E11 = 2.8
E12 = matrix(c(1.62, 0.87, 0.14), 1,3)
E21 = matrix(c(1.62, 0.87, 0.14), 3,1)
E22 = matrix(c(2.8, 1.62, 0.87, 1.62, 2.8, 1.62, 0.87, 1.62, 2.8), 3,3)
y = matrix(c("e1", "e2", "e3"), 3,1)
ux = 0
uy = matrix(c(0,0,0), 3,1)
E_xy = E11 - E12 %*% solve(E22) %*% E21
u_xy = ux + E12 %*% solve(E22) %*% (y - uy)

e_t = numeric(n)
## Proceso de simulación mediante modelo de espacio de estados
for (w in 1:N) {
  for(h in 2:n) {
    if (h == 2) {
      Be_sim[,h,w] = Tt %*% Be_sim[,h-1,w] + mvrnorm(1,mu_e,Qt)
      Ye_sim[h,w] = Zt %*% Be_sim[,h,w] + 0
      next;
```

```
    }
    if (h == 3) {
      Be_sim[,h,w] = Tt %*% Be_sim[,h-1,w] + mvrnorm(1,mu_e,Qt)
      Ye_sim[h,w] = Zt %*% Be_sim[,h,w] + 0
      next;
    }
    y = matrix(c(e_t[h-1], e_t[h-2], e_t[h-3]), 3,1)
    u_xy = ux + E12 %*% solve(E22) %*% (y-uy)
    e_t[h] = rnorm(1,u_xy,sqrt(sigma2_xy))
    Be_sim[,h,w] = Tt %*% Be_sim[,h-1,w] + mvrnorm(1,mu_e,Qt)
    Ye_sim[h,w] = Zt %*% Be_sim[,h,w] + e_t[h]
  }
}

## Eliminando las primeras 100 observaciones simulados
Ye_sim = Ye_sim[101:300,]
Be_sim = Be_sim[,101:300,]

## Estimación de los vectores de estados
# Matrices del sistema
Zt = matrix(c(1,0,1,0,1,0,1,0,1,0,1,0,1),1,13)

# Matriz de transición
Tt = matrix(c(1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,
              0,0,cos(pi/6),-sin(pi/6),0,0,0,0,0,0,0,0,0,
              0,0,sin(pi/6),cos(pi/6),0,0,0,0,0,0,0,0,0,
              0,0,0,0,cos(pi/3),-sin(pi/3),0,0,0,0,0,0,0,
              0,0,0,0,sin(pi/3),cos(pi/3),0,0,0,0,0,0,0,
              0,0,0,0,0,0,cos(pi/2),-sin(pi/2),0,0,0,0,0,
              0,0,0,0,0,0,sin(pi/2),cos(pi/2),0,0,0,0,0,
              0,0,0,0,0,0,0,0,cos(2*pi/3),-sin(2*pi/3),0,0,0,
              0,0,0,0,0,0,0,0,sin(2*pi/3),cos(2*pi/3),0,0,0,
              0,0,0,0,0,0,0,0,0,0,cos(5*pi/6),-sin(5*pi/6),0,
              0,0,0,0,0,0,0,0,0,0,sin(5*pi/6),cos(5*pi/6),0,
              0,0,0,0,0,0,0,0,0,0,0,0,-1), nrow = 13, ncol = 13)

GL = 0.0024 # varianza del error de la tendencia
GR = 0.0004 # varianza del error de la pendiente
GS = 0.0000001 # varianza de efectos de estacionalidad
Ge = 2.8 # varianza del error de la serie original

# Matriz de varianzas
```

```
Qt = diag(c(GL,GR,GS,GS,GS,GS,GS,GS,GS,GS,GS,GS,GS))

# Vector de estados
# Bt = c(Lt, Rt, S1t, SS1t, S2t, SS2t, S3t, SS3t, S4t, SS4t, S5t, SS5t, S6t)

Bhat = array(0,c(13,300,5000)) # Arreglo que acumula los betas estimados
Phat0 = diag(10000, nrow = 13, ncol = 13) # Acumula las covarianzas
Phat = array(Phat0, c(13,13,1))

## Aplicando el filtro de Kalman

for (t in 1:5000) {
  for (l in 2:200) {
    Bhat10 = Tt %*% %Bhat[ ,l-1,t]
    Phat10 = Tt %*% %Phat[ , ,l-1] %*% %t(Tt) + Qt
    Fhat = Zt %*% %Phat10 %*% %t(Zt) + Ge
    Bhat[,l,t] = Bhat10 + Phat10 %*% %t(Zt) %*% %(1/Fhat) %*% %
    (Ysim[l,t] - Zt %*% %Bhat10)
    Ph = Phat10 - Phat10 %*% %t(Zt) %*% %(1/Fhat) %*% %Zt %*% %Phat10
    Phat = array(c(Phat, Ph), c(13,13,l))
  }
}

## Aplicando el algoritmo de Pfeffermann y Tiller
# Correlaciones
rho_1 = 0.23 # corr(e_{t},e_{t-1})
rho_2 = 0.17 # corr(e_{t},e_{t-2})
rho_3 = 0.11 # corr(e_{t},e_{t-3})

# Varianzas gamma_0 = Ge = 1.8, var(e_{t}) = var(e_{t-1}) = var(e_{t-2})
gamma_0 = 1.8

# Covarianzas
gamma_1 = rho_1*gamma_0
gamma_2 = rho_2*gamma_0
gamma_3 = rho_3*gamma_0

# Matriz de covarianzas
Pe0 = diag(10000, nrow = 13, ncol = 13)
P_10 = Tt %*% %Pe0 %*% %t(Tt) + Qt # matriz de predicción del error
F1 = Zt %*% %P_10 %*% %t(Zt) + gamma_0 # matriz de cov de la innovación
K1 = P_10 %*% %t(Zt) %*% %(1/F1) # ganancia de Kalman
```

```
#Ah1 = Tt %* %K1

C2 = Ah1 %* %gamma_1 # covarianzas de los errores del muestreo
C1 = matrix(0,13,1)
Ct = matrix(0,13,0)
Ct = cbind(C1,C2)

H1 = 1/(gamma_0 - t(Ct[,1]) %* %solve(P_10) %* %Ct[ ,1])
Pe1  = solve(solve(P_10) + (t(Zt) - solve(P_10) %* %Ct[ ,1]) %* %H1
%* %(Zt - t(Ct[ ,1]) %* %solve(P_10)))
P_21 = Tt %* %Pe1 %* %t(Tt) + Qt
Pe = array(Pe1, c(13,13,1))

sigma_tt = matrix(gamma_0,1,1) # transformando a matriz de covarianzas
v1 = cbind(P_21,Ct[,2]) # formando una matriz
v2 = cbind(t(Ct[,2]),sigma_tt)
Vt2 = rbind(v1,v2) # matriz de covarianzas de los errores

I = diag(1,13,13)
IZ = cbind(I,t(Zt))
IZV1 = IZ %* %solve(Vt2)
IZV = array(0,c(13,14,1))
IZV = array(c(IZV,IZV1),c(13,14,2))

Behat = array(0,c(13,200,5000))
B0e = matrix(0,13,1)

# beta hat pata t = 1
for (z in 1:5000) {
  Behat[,1,z] = (I-K1 %* %Zt) %* %Tt %* %B0e + K1 %* %Ye_sim[1,z]
}

Phe = array(Pe1, c(13,13,1))
C2e = Tt %* %K1 %* %gamma_1
Cte = array(0, c(13,202,5000))
Cte[,2,] = C2e
IZVh = array(0,c(13,14,1))
IZVh = array(c(IZVh,IZV1),c(13,14,2))

## Algoritmo de Pfefferman & Tiller

for (p in 1:5000) {
```

```
for (q in 2:200) {
  Phe_21 = Tt %* %Phe[, ,q-1] %* %t(Tt) + Qt
  Hth  = 1/(Ge - t(Cte[ ,q,p]) %* %solve(Phe_21) %* %Cte[ ,q,p])
  Phe2  = solve(solve(Phe_21) + (t(Zt) - solve(Phe_21) %* %
  Cte[ ,q,p]) %* %Hth %* %(Zt - t(Cte[ ,q,p]) %* %solve(Phe_21)))
  Behat[ ,q,p] = Tt %* %Behat[ ,q-1,p] + Phe2 %* %(t(Zt) -
  solve(Phe_21) %* %Cte[ ,q,p]) %* %Hth %* %(Ye_sim[q,p] -
  Zt %* %Tt %* %Behat[ ,q-1,p])
  Phe = array(c(Phe, Phe2), c(13,13,q))

  if (q == 2) {
    At_1h = Tt %* %Phe[ , ,q] %* %IZVh[ ,1:13,q]
    Ahat_t_1h = Tt %* %Phe[ , ,q] %* %IZVh[,14,q]
    C3h = At_1h %* %Ah1 %* %gamma_2 + Ahat_t_1h %* %gamma_1
    Cte[ ,q+1,p] = C3h

    Phe_32 = Tt %* %Phe[ , ,q] %* %t(Tt) + Qt
    Cch = cbind(Phe_32,Cte[ ,q+1,p])
    Cfh = cbind(t(Cte[ ,q+1,p]),sigma_tt)
    V3h = rbind(Cch,Cfh)
    IZV3h = IZ %* %solve(V3h)
    IZVh = array(c(IZVh,IZV3h),c(13,14,q+1))
    next;
  }

  if (r == 3) {
    At_1h = Tt %* %Phe[ , ,q] %* %IZVh[ ,1:13,q]
    Ahat_t_1h = Tt %* %Phe[ , ,q] %* %IZVh[ ,14,q]
    At_2h = Tt %* %Phe[ , ,q-1] %* %IZVh[ ,1:13,q-1]
    Ahat_t_2h = Tt %* %Phe[ , ,q-1] %* %IZVh[ ,14,q-1]
    C4h = At_1h %* %At_2h %* %Ah1 %* %gamma_3 +
    At_1h %* %Ahat_t_2h %* %gamma_2 + Ahat_t_1h %* %gamma_1
    Cte[ ,q+1,p] = C4h

    Phe_43 = Tt %* %Phe[ , ,q] %* %t(Tt) + Qt
    Cch = cbind(Phe_43,Cte[ ,q+1,p])
    Cfh = cbind(t(Cte[ ,q+1,p]),sigma_tt)
    V4h = rbind(Cch,Cfh)
    IZV4h = IZ %* %solve(V4h)
    IZVh = array(c(IZVh,IZV4h),c(13,14,q+1))
    next;
  }
```

```
    At_1h = Tt %*% Phe[ , ,q] %*% IZVh[ ,1:13,q]
    At_2h = Tt %*% Phe[ , ,q-1] %*% IZVh[ ,1:13,q-1]
    Ahat_t_1h = Tt %*% Phe[ , ,q] %*% IZVh[ ,14,q]
    Ahat_t_2h = Tt %*% Phe[ , ,q-1] %*% IZVh[ ,14,q-1]
    Ahat_t_3h = Tt %*% Phe[ , ,q-2] %*% IZVh[ ,14,q-2]
    Ctemph = At_1h %*% %At_2h %*% %Ahat_t_3h %*% %gamma_3 +
    At_1h %*% %Ahat_t_2h %*% %gamma_2 + Ahat_t_1h %*% %gamma_1
    Cte[,q+1,p] = Ctemph

    Phe_54 = Tt %*% Phe[ , ,q] %*% t(Tt) + Qt
    Cch = cbind(Phe_54,Cte[ ,q+1,p])
    Cfh = cbind(t(Cte[,q+1,p]),sigma_tt)
    Vth = rbind(Cch,Cfh)
    IZVth = IZ %*% solve(Vth)
    IZVh = array(c(IZVh,IZVth),c(13,14,q+1))
  }
}
```

Bibliografía

Box, G. E. P. y Jenkins, G. G. (1976). *Time Series Analysis: Forecasting and Control*, Oakland, California, U.S.A.: HOLDEN-DAY INC.

Commandeur, J. J. F. y Koopman, S. J. (2007). *An Introduction to Space State Time Series Analysis*, Oxford University Press Inc., New York.

Durbin, J. y Koopman, S. (2012). *Time Series Analysis by State Space Methods*, London, U.K.: Oxford University Press.

Durbin, J. y Quenneville, B. (1997). Benchmarking by state-space models, *International Statistical Review* **65**: 23–48.

George E. P. Box, Gwilym G. Jenkins, G. C. R. y Ljung, G. M. (2016). *Time Series Analysis: Forecasting and Control*, Hoboken, New Jersey: John and Sons, Inc.

Harvey, A. C. (1989). *Forecasting Structural Time Series With the Kalman Filter*, Cambridge, U.K.: Cambridge University Press.

Hillmer, S. C. y Trabelsi, A. (1987). Benchmarking of economic time series, *Journal of the American Statistical Association* **82**: 1064–1071.

ILO (2019). *International Labour Organization: Quick guide on interpreting the unemployment rate*, International Labour Organization, First published.

INEI (2020). *Instituto Nacional de Estadística e Informática: Encuesta Permanente del Empleo - EPE.*
URL: *https://webinei.inei.gob.pe/anda_inei/index.php/catalog/704*

NU (2008). *Naciones Unidas: Diseño de muestras para encuestas de hogares: directrices prácticas.*
URL: *https://unstats.un.org/unsd/publication/seriesf/Seriesf_98s.pdf*

Pfeffermann, D. (2013). New important developments in small area estimation, *Statistical Science* **28**: 40–68.

Pfeffermann, D., S. A. y Tiller, R. (2014). Single and two stage cross-sectional and time series benchmarking procedures for small area estimation, *TEST* **23**: 631–666.

Pfeffermann, D. y Tiller, R. B. (2005). Bootstrap approximation to prediction mse for state-space models with estimated parameters, *Journal of Time Series Analysis* **26**: 893–916.

Pfeffermann, D. y Tiller, R. B. (2006). Small-area estimation with state-space models: Subject to benchmark constraints, *Journal of the American Statistical Association* **101**: 1387–1397.

R., C. (1974). *Stochastic Processes*, L. Marder, University of Southampton.

Rao, J. y Molina, I. (2015). *Small Area Estimation, Second Edition*, Wiley.