CIMAT

Centro de Investigación en Matemáticas, A.C.

Indicador de alta frecuencia de la actividad económica de México

T E S I S

Que para obtener el grado de

Maestro en Cómputo Estadístico

Presenta

Orlando de Jesus Uc Kantun

Director de Tesis:

Dra. Graciela María de los Dolores González Farías

Co-director de Tesis:

Dr. Francisco de Jesús Corona Villavicencio.

Autorización de la versión final

Monterrey, Nuevo León, Noviembre de 2021

Dedicado a todas aquellas personas que me quieren

Resumen

En la presente tesis de maestría se crea un indicador económico diario a partir de un Modelo de Factores Dinámicos ajustado a un conjunto de variables tradicionales y a otro conjunto de variables no tradicionales, con el cual se estima un factor dinámico de alta frecuencia, que, restringido a los valores mensuales del Indicador Global de la Actividad Económica, permite analizar la actividad económica de México. El factor dinámico se estima con el método del suavizamiento de Kalman y se elige el número de factores a partir de tres criterios diferentes. El indicador estimado lleva por nombre Indicador Diario de la Actividad Económica.

Palabras clave: actividad económica, alta frecuencia, modelo de factores dinámicos, regla de combinación, variables no tradicionales.

Abstract

In this master's thesis, a daily economic indicator is created based on a Dynamic Factor Model adjusted to a set of traditional variables and another set of non-traditional variables, with which a high-frequency dynamic factor is estimated, which, restricted to the monthly values of the Global Indicator of Economic Activity, it allows to analyze the economic activity of Mexico. The dynamic factor is estimated with the Kalman smoothing method and the number of factors is chosen from three different criteria. The estimated indicator is called Daily Indicator of Economic Activity.

Keywords: combination rule, dynamic factor model, economic activity, high frequency, non-traditional variables.

Agradecimientos

Agradezco infinitamente a mis asesores, la Dra. Graciela González y el Dr. Francisco Corona, por haberme dirigido en el desarrollo de la presente tesis. Les agradezco muchísimo el tiempo, las observaciones, las correcciones, los consejos y las instrucciones, pues son personas a quienes respeto y admiro profundamente, en gran parte por sus conocimientos y su experiencia, pero más aún por su gran calidad humana. Son un gran modelo a seguir para todos nosotros, sus alumnos. Ha sido, y será siempre, un gran honor.

Le agradezco enormemente al Centro de Investigación en Matemáticas A.C., pero de manera muy particular a la Unidad Monterrey, por haberme formado y enseñado más que matemáticas, pues siendo estudiante de la Maestría en Cómputo Estadístico crecí y madure muchísimo, y me convertí en una persona más completa.

Agradezco de la manera más cariñosa a mi mamá, Enf. Silvia del Socorro Kantun Chan, por haberme querido siempre, y por haberme enseñado tantas grandes lecciones que me permitieron llegar hasta donde estoy. ¡Te amo mamá!

Le agradezco de la manera más efusiva a mi papá, C.D. Lino Orlando Uc LLanes, por haberme dado las herramientas necesarias para seguir creciento, y por haberme enseñado la importancia de la disciplina y el trabajo para alcanzar el éxito. ¡Te quiero papá!

Agradezco de la manera más amorosa a mi hermano Daniel, por haber luchado siempre a mi lado, por ser mi amigo, por ser mi gran apoyo, y por ser una motivación más para ser siempre un buen ejemplo. ¡Te quiero hermano!

Agradezco atentamente a toda mi familia, por haber estado siempre cerca de mí, por quererme y enseñarme mucho, por aprenciarme y estar al pendiente de mi bienestar.

Agradezco profundamente a todos mis amigos y amigas, por ser parte de mi vida, por estar conmigo, por poder compartir con ustedes risas y momentos felices, pero sobre todo, por haber estado conmigo cuando más los he necesitado. Donde quiera que estén, porque son tantos y de lugares tan distintos, que ni todas las páginas de éste documento me permitirían enlistarlos, gracias, muchas gracias.

Por último, le agradezco al Consejo Nacional de Ciencia y Tecnología, por haberme apoyado con una beca de maestría, en el periodo agosto 2019 - julio 2021 con CVU 962019, para poder concluir mi posgrado a tiempo y de manera exitosa.

Índice general

Índice de figuras

Capítulo 1

Introducción

Es sabido que en la actualidad disponemos de una gran cantidad de información para comprender fenómenos de diversa índole. En lo que respecta a fenómenos macroeconómicos, la gran cantidad de series de tiempo económicas y financieras colectadas en las últimas décadas por los bancos de información de cada país ha permitido construir modelos estadísticos y econométricos que, entre diversas cosas, coadyuvaron a entender las causas de crisis económicas e incluso, poder anticiparse en el corto plazo a éstas (Bai y Ng, 2008; Stock y Watson, 2012).

En este sentido, no obstante los avances en técnicas de estimación, principalmente los llamados Modelos de Factores Dinámicos (MFD), los cuales permiten resumir las dinámicas de las variables en un menor número de conjunto de series de tiempo, llamados factores. En la literatura son pocos los trabajos que combinan el uso de información tradicional, no tradicional y de alta frecuencia para estimar las componentes del MFD y que permitan desentrañar de mejor forma los fenómenos económicos en el corto plazo.

La pandemia de la COVID-19 ha generado retos en términos de modelación econométrica y estadística dado que la sociedad actual requiere conocer la magnitud de lo que implica fenómenos de esta índole y mejor, si esto se hace en tiempo real. Por ejemplo, el comité de fechados de ciclos económicos del Banco Nacional de Investigación Económica de los Estados Unidos (*NBER* por sus siglas en inglés) reconoció que la economía estadounidense estuvo en recesión durante marzo y abril de 2020, pero

para ello tuvo que pasar más de un año para que se realizara dicho anuncio [1] . En México, el Instituto Nacional de Estadística y Geografía (INEGI) generó el Indicador Oportuno de la Actividad Económica (IOAE) con el fin de generar *nowcasts* del Indicador Global de la Actividad Económica (IGAE) hasta 8 semanas antes que el dato oficial publicado. Estos dos ejemplos dan argumento de que es necesario generar los mecanismos apropiados para brindar información sobre el estado de la economía.

En este sentido, se observa que hay una brecha por cubrir, en relación con generar un indicador económico que brinde información del estado de la economía en tiempo real. Aunque la metodología propuesta en este trabajo puede ser utilizada para cualquier economía, nosotros enfatizaremos la aplicación para el caso de México, a fin de dar conocer a los tomadores de decisión y la sociedad en general, un nuevo indicador económico para el análisis de la coyuntura macroeconómica mexicana.

Por lo tanto, **el objetivo general de esta tesis es construir un indicador económico de alta frecuencia para la economía mexicana**. Para cumplir lo anterior, se propone estimar un factor dinámico de alta frecuencia mediante el método de Componentes Principales (PC) y con el método de estimación en dos pasos por suavizamiento de Kalman, el cual esté altamente correlacionado con la actividad económica, restringiendo el comportamiento mensual del factor a los valores mensuales del IGAE. Lo anterior, a través de la Regla de Combinación (RC) de Guerrero y Nieto (1999) que permite restringir contemporánea y temporalmente el comportamiento de las estimaciones a relaciones contables oficiales; por ejemplo, que series mensuales satisfagan el comportamiento de una serie trimestral o que la suma de desagregados correspondan a nivel de agregación mayor. El indicador propuesto se denomina Indicador Diario de la Actividad Económica (IDAE).

En este orden de ideas, para lograr el objetivo general de la tesis, proponemos lo siguiente: i) Uso de información de búsquedas de internet, principalmente variables extraídas de Google Trends (GT) que estén relacionados con la actividad económica. ii) Variables económicas y financieras de alta frecuencia. iii) Estimación de factores dinámicos. iv) Restricción del comportamiento de los factores dinámicos al compor-

[1] https://www.nber.org/news/business-cycle-dating-committee-announcement-july-19-2021

tamiento de una variable macroeconómica, en este caso, el IGAE mediante el uso RC.

En consecuencia, y transversalmente, los objetivos específicos son usar información tradicional, no tradicional y de alta frecuencia altamente correlacionada con el IGAE; emplear MFD posiblemente no estacionarios para estimar el factor a través de CP; usar la RC para restringir los valores del factor dinámico a los valores agregados y mensuales del IGAE; usar un indicador de alta frecuencia restringido para analizar la coyuntura económica de México en el corto plazo.

Literatura relacionada con nuestro trabajo la podemos ver en Bai y Ng (2008), donde se presentan los principales resultados teóricos del modelo de factores dinámicos, por lo que se considera un apropiado punto de partida para la revisión de las especificaciones del modelo.

En el marco del estado del arte de los indicadores de alta frecuencia Aprigliano, Foroni, Marcellino, Mazzi, y Venditti (2017) y Lourenço y Rua (2021) propusieron indicadores económicos de frecuencia diaria. Recientemente Lewis, Mertens, y Stock (2020) propusieron la creación de un índice económico semanal para cuantificar el impacto económico del COVID-19 en Estados Unidos y Eraslan y Götz (2021) desarrollaron un indicador económico semanal para la economía alemana basado en nueve indicadores de alta frecuencia.

En otros trabajos relacionados con el nuestro, recientemente, Corona, González-Farías, y López-Pérez (2021) ajustaron un MFD para realizar estimaciones oportunas del IGAE.

La organización de esta tesis es como sigue, en la Sección 2 describimos las metodologías empleadas para cumplir con el objetivo. En la Sección 3 presentamos a detalle la estimación empírica del IDAE y finalmente, en la Sección 4 concluimos.

Capítulo 2

Metodología

En este capítulo se introduce la definición del MFD y sus componentes, así como la estimación por el método de PC y la estimación por suavizamiento de Kalman. Se presentan también tres criterios para la selección del número de factores y las pruebas "PANIC" (*Panel Analysis of Nonstationarity in the Idiosyncratic and Common components*, en inglés), además de la RC. Se hace un breve comentario sobre la regresión por Mínimos Cuadrados Parciales (PLS) que se utilizará en la aplicación para la selección de variables. Al final se da una introducción a los métodos de clústering para series de tiempo, particularmente el método Soft DTW k-means (*Soft Dynamic Time Warping* k-*means*, en inglés), que se utilizaron con fines exploratorios.

2.1. Modelo de Factores Dinámicos

Sea N un conjunto de series de tiempo, observadas de $t = 1, \ldots, T$, el MFD está definido mediante las siguientes expresiones Corona, González-Farías, y López-Pérez (2021):

$$Y_t = PF_t + \epsilon_t, \tag{2.1}$$

$$\Phi(L)F_t = \eta_t, \tag{2.2}$$

$$\Gamma(L)\epsilon_t = a_t, \tag{2.3}$$

donde $Y_t = (y_{1t}, \ldots, y_{Nt})'$ es el vector de observaciones dimensión $N \times 1$, $P = (p_1, \ldots, p_N)'$ es la matriz de cargas de dimensión $N \times r$, donde cada $p_i = (p_{i1}, \ldots, p_{ir})'$ es un vector de dimensión $r \times 1$, $F_t = (F_{1t}, \ldots, F_{rt})'$ es un vector de dimensión r-dimensional de factores comunes ($r < N$) y $\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{Nt})'$ es un vector de errores idiosincráticos de dimensión $N \times 1$. Además $\Phi(L) = I - \sum_{i=1}^{k} \Phi L^i$ y $\Gamma(L) = I - \sum_{j=1}^{s} \Gamma L^j$ contienen a las matrices de los coeficientes autorregresivos de órdenes k y s para los factores y errores idiosincráticos, Φ y Γ respectivamente, y donde L es el operador retardo. Por otro lado, $\eta_t = (\eta_{1t}, \ldots, \eta_{rt})'$ y $a_t = (a_{1t}, \ldots, a_{Nt})'$ son los disturbios de los factores y los errores idiosincráticos, respectivamente, que se suponen con media 0 y matriz de covarianza positiva-definida.

Existen otras maneras de definir al MFD, que pueden utilizarse según sea conveniente, ver por ejemplo Bai y Ng (2007) y Barigozzi, Lippi, y Luciani (2016), donde se presentan generalizaciones del MFD para el caso estacionario y no estacionario respectivamente.

2.2. Estimación por componentes principales

Uno de los métodos más populares para la estimación de los parámetros del MFD es la estimación por PC. Dicho procedimiento tiene la ventaja de que no es necesario suponer una distribución paramétrica para los errores idiosincráticos ϵ_t, ni para los factores comunes F_t.

Considérese el siguiente supuesto propuesto por Chamberlain y Rothschild (1983)

$$\lambda_{\text{máx}}(\Sigma_\epsilon) \leq c < \infty, \text{para todo } N, \tag{2.4}$$

donde $\lambda_{\text{máx}}(\Sigma_\epsilon)$ representa al valor propio más grande de la matriz $\Sigma_\epsilon = \mathbb{E}[\epsilon_t' \epsilon_t]$. La condición (2.4) implica que el valor propio $\lambda_{\text{máx}}(\Sigma_\epsilon)$ es finito y acotado por una constante c, por tanto, la variabilidad del valor esperado de los errores idiosincráticos cuadrados es finita, entonces, los errores idisioncráticos tienen una correlación finita entre las series analizadas.

Supongamos que los factores dinámicos F_t pueden ser estimados a partir de una matriz de pesos W, de tal forma que:

$$\hat{F}_t\left(\frac{1}{N}W\right) = \frac{1}{N}W'Y_t, \tag{2.5}$$

y supongamos también:

$$\lim_{N\to\infty}\frac{1}{N}W'P = H, \tag{2.6}$$

donde H es una matriz de dimensión $r \times r$ de rango completo.

Dado que suponemos que los errores idisioncráticos ϵ_t siguen un proceso ruido blanco, entonces tienen media 0 y varianza constante, y por la ley débil de los grandes números $\epsilon_t \xrightarrow{p} 0$, más aún, por propiedades de la convergencia en probabilidad:

$$\frac{1}{N}W'\epsilon_t \xrightarrow{p} 0. \tag{2.7}$$

Partiendo del MFD:

$$\begin{aligned}\hat{F}_t\left(\frac{1}{N}W\right) &= \frac{1}{N}W'\left(PF_t + \epsilon_t\right)\\ &= \frac{1}{N}W'PF_t + \frac{1}{N}W'\epsilon_t,\end{aligned}$$

luego, dados los límites (2.6) y (2.7):

$$\lim_{N\to\infty}\hat{F}_t\left(\frac{1}{N}W\right) = \lim_{N\to\infty}\frac{1}{N}W'PF_t + \lim_{N\to\infty}\frac{1}{N}W'\epsilon_t = HF_t. \tag{2.8}$$

El resultado de la ecuación (2.8) permite hacer observaciones importantes del MFD. En primer lugar, si el número de series analizadas es suficientemente grande ($N \to \infty$), entonces el efecto de los errores idiosincráticos ponderados converge a 0, lo cual implica que con una estimación correcta de los factores comunes el efecto de los errores idiosincráticos es despreciable, además, solamente se mantienen los efectos de las combinaciones lineales de los factores. Por otro lado, no es necesario imponer una distribución paramétrica sobre los errores idiosincráticos para obtener las convergen-

cias, es decir, no es necesario suponer normalidad en la distribución de los errores idiosincráticos.

El estimador por el método de componentes principales $\hat{F}_t$ puede ser obtenido a partir de las soluciones del problema de mínimos cuadrados:

$$\hat{F}_t = \min_{F_1,\ldots,F_t,P} S(F_t, P),$$

donde la función objetivo S está dada por:

$$S(F_t, P) = \frac{1}{NT}\sum_{t=1}^{T}(Y_t - PF_t)'(Y_t - PF_t). \tag{2.9}$$

Obsérvese que si se tiene, por ejemplo, una matriz arbitraria A ortogonal de dimensión $r \times r$, entonces:

$$PF_t = P(AA^{-1})F_t = (PA)(A^{-1}F_t) = F^*P^*,$$

y por tanto, el MFD también se puede expresar como $Y_t = F^*P^* + \epsilon_t$, es decir, el MFD puede ser reexpresado en términos de alguna rotación de la matriz de coeficientes P y de los factores F_t. Luego, es necesario aplicar restricciones para poder fijar a P y a F_t, de otra manera, se podrían obtener infinitas soluciones al problema de mínimos cuadrados. Una opción es imponer las condiciones $\frac{1}{N}P'P = I_r$ en conjunto con la condición de que $F'F$ sea diagonal, con $F = (F_1, \ldots, F_T)$ la matriz de factores comunes de dimensión $r \times T$. Otra alternativa es imponer la condición $\frac{1}{T}F_t'F_t = I_r$, en conjunto con la condición de que $P'P$ sea diagonal Bai y Ng (2008).

Por tanto, para obtener una solución única se expresa al problema de mínimos cuadrados cómo:

$$\hat{F}_t = \min_{F_1,\ldots,F_t,P} S(F_t, P) \text{ s.a. } \begin{cases} \frac{1}{N}P'P = I_r \\ F'F \text{ es una matriz diagonal,} \end{cases} \tag{2.10}$$

o equivalentemente:

$$\hat{F}_t = \min_{F_1,\ldots,F_t,P} S(F_t, P) \text{ s.a. } \begin{cases} \frac{1}{T} F_t' F_t = I_r \\ P'P \text{ es una matriz diagonal,} \end{cases} \tag{2.11}$$

donde la función objetivo S está dada por:

$$S(F_t, P) = \frac{1}{NT} \sum_{t=1}^{T} (Y_t - PF_t)'(Y_t - PF_t). \tag{2.12}$$

De acuerdo con Stock y Watson (2011), para resolver el problema de optimización (2.10), primero se minimiza con respecto a F_t considerando a P fijo, para obtener $\hat{F}_t\,(P(P'P)^{-1}) = (P'P)^{-1}P'Y_t$, por lo que la función a optimizar se convierte en:

$$\min_P \frac{1}{T} \sum_{t=1}^{T} Y_t'[I - P(P'P)^{-1}P]Y_t \quad \text{s.a. } \begin{cases} \frac{1}{N} P'P = I_r \\ F'F \text{ es una matriz diagonal.} \end{cases}$$

Este problema de minimización es equivalente a:

$$\max_P \operatorname{tr}\left[(P'P)^{-\frac{1}{2}\prime} P' \left(\frac{1}{T} \sum_{t=1}^{T} Y_t Y_t' \right) P(P'P)^{-\frac{1}{2}} \right] \quad \text{s.a. } \begin{cases} \frac{1}{N} P'P = I_r \\ F'F \text{ es una matriz diagonal,} \end{cases}$$

que se puede reexpresar como:

$$\max_P P'\hat{\Sigma}_Y P \quad \text{s.a. } \begin{cases} \frac{1}{N} P'P = I_r \\ F'F \text{ es una matriz diagonal,} \end{cases} \tag{2.13}$$

donde $\hat{\Sigma}_Y = \frac{1}{T} \sum_{t=1}^{T} Y_t Y_t'$ es la matriz de covarianza muestral.

La solución del problema de optimización (2.13) es tomar a $\hat{P}$ como la matriz formada por los vectores propios ordenados de $\hat{\Sigma}_Y$ correspondientes a sus r valores propios más grandes, multiplicada por $\sqrt{N}$. Finalmente, el estimador por componentes

principales de F_t es:

$$\hat{F}_t = \frac{1}{N}\hat{P}'Y_t, \quad t = 1, \ldots, T. \tag{2.14}$$

Cabe señalar que, al seleccionar a los r valores propios más grandes, con la estimación PC solamente se conserva un porcentaje de la variabilidad de las series, por lo que la determinación de un valor adecuado para r es un aspecto fundamental que se abordará en la Sección 2.4.

Es importante notar que, a pesar de que la estimación por máxima verosimilitud de los factores dinámicos F_t es más robusta, supone un problema computacional, por el gran número de parámetros. Por otro lado, Bai y Ng (2002) probaron que la estimación por PC es consistente, y tiene la ventaja de no requerir tanto poder de cómputo, por lo que usualmente se prefiere la estimación por PC.

2.3. Estimación por suavizamiento de Kalman

El método de estimación en dos pasos por suavizamiento de Kalman (2SKS, por sus siglas en inglés), fue propuesto por Giannone, Reichlin, y Small (2008), y desarrollado también por Doz, Giannone, y Reichlin (2011). Consiste en primero estimar a los factores comunes a través del método de PC y después actualizar y reajustar a las matrices que describen la dinámica del MFD por medio del suavizamiento de Kalman.

El suavizamiento de Kalman es un algoritmo que permite hacer estimaciones de una variable F_t a partir de un conjunto de observaciones previas $\mathcal{F}_t = \{Y_i | i = 1, \ldots, T\}$. El algoritmo comienza con la observación inmediata anterior y va en orden recursivo hacia atrás.

Un aspecto importante de la estimación 2SKS es que permite hacer estimaciones a factores que pueden ser no estacionarios, siempre y cuando los errores idiosincráticos sean estacionarios Corona, Poncela, y Ruiz (2020).

El algoritmo 2SKS adaptado de Corona, Poncela, y Ruiz (2020) se presenta a continuación:

1. Estime P por el método de PC. Calcule $\hat{F}_t = \frac{1}{N}\hat{P}'Y_t$, para $t = 1, \ldots, T$, defina

$\hat{\epsilon} = Y - \hat{P}\hat{F}$ y calcule la matriz de covarianza muestral de los errores idiosincráticos $\hat{\Psi} = \text{diag}(\hat{\Sigma}_\epsilon)$, donde la diagonal de $\hat{\Psi}$ incluye las varianzas específicas σ_i^2 $(i = 1, \ldots, N)$ de cada serie de las variables que componen a Y_t.

2. Aplique la prueba ADF a cada factor estimado $\hat{F}_j$, con $j = 1, \ldots, r$.

   - Si se rechaza la hipótesis nula, calcule el estimador por mínimos cuadrados del coeficiente autorregresivo $\hat{\phi}_j$, los residuales $\hat{u}_{jt} = F_{jt} - \hat{\phi}_i F_{jt-1}$ y la varianza muestral de los disturbios de los factores $\hat{\sigma}_{\eta t} = \frac{1}{T}\sum_{t=1}^{T} u_{jt}$. Asuma que el estado inicial del factor tiene media 0 y varianza estimada $\hat{\sigma}^2_{F_j} = \frac{\hat{\sigma}^2_{\eta j}}{1-\hat{\phi}_j^2}$.
   - Si no se rechaza la hipótesis nula, entonces $\hat{\phi}_j = 1$, y los residuales se calculan como $\hat{u}_{jt} = \Delta\hat{F}_{jt}$. Calcule la varianza de los disturbios de los factores $\hat{\sigma}_{\eta t} = \frac{1}{T-1}\sum_{t=2}^{T} \Delta\hat{F}^2_{jt}$. Asuma que el estado inicial del factor es difuso con media 0 y varianza $\hat{\sigma}^2_{F_j} = k$, con k una constante grande.

3. Exprese al MFD definido en las ecuaciones (2.1), (2.2) y (2.3), y con las matrices estimadas $\hat{P}$, $\hat{\Psi}$, $\hat{\Phi}$, $\hat{\Sigma}_\eta$ y $\hat{\Sigma}_f$ use el suavizamiento de Kalman para actualizar y reajustar las estimaciones de los factores $\hat{F}$.

Como se puede observar, el método 2SKS tiene la ventaja de que refina las estimaciones obtenidas por el método de PC, ya que tiene una menor varianza.

2.4. Selección del número de factores

Los criterios para la selección del número de factores utilizados en éste trabajo toman como parámetro a un valor máximo de parámetros a considerar denotado como $r_{\text{máx}}$. Corona, Muriel, y González-Farías (2021) sugieren, de manera empírica, elegir $r_{\text{máx}} = \lfloor 1.55 \min\{T^{\frac{2}{5}}, N^{\frac{2}{5}}\} \rfloor$.

2.4.1. Criterios de información de Bai y Ng

Los criterios de información de Bai y Ng (2002) son un conjunto de funciones a minimizar que toman como argumentos el número máximo de factores a considerar $r_{\text{máx}}$, con $r \leq r_{\text{máx}}$, el valor $m = \text{mín}\{N, T\}$ y la función objetivo $\{S_k(\hat{F}_t, \hat{P})\}$ definida como en la ecuación (2.12) pero con F_t y P estimados por PC.

Las funciones a minimizar con respecto a k son:

$$IC_1(k) = \ln S_k(\hat{F}_t, \hat{P}) + k\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right), \tag{2.15a}$$

$$IC_2(k) = \ln S_k(\hat{F}_t, \hat{P}) + k\left(\frac{N+T}{NT}\right)\ln m, \tag{2.15b}$$

$$IC_4(k) = \ln S_k(\hat{F}_t, \hat{P}) + k\left(\frac{\ln m}{m}\right), \tag{2.15c}$$

para $k = 0, \ldots, r_{\text{máx}}$.

2.4.2. Procedimiento de Onatski

El procedimiento de Onatski (2010) consiste en comparar a pares los valores propios λ_j de la matriz de covarianza muestral $\hat{\Sigma}_Y = \frac{1}{T}\sum_{t=1}^{T} Y_t Y_t'$ para $j = 1, \ldots, r_{max}$. A partir de ciertos resultados de matrices aleatorias, particularmente de la distribución Tracy-Widom,se puede suponer que a valores grandes del número de series N y del periodo de observación T que la diferencia entre $\lambda_j - \lambda_{j+1}$ tiende a 0, mientras que la diferencia $\lambda_r - \lambda_{r+1}$ diverge. Entonces el objetivo del procedimiento de Onastki es encontrar un umbral δ que distinga a los valores propios convergentes de los divergentes, de tal forma que se tiene una familia de estimadores:

$$\hat{r}(\delta) = \text{máx}\{i \leq r_{\text{máx}} : \lambda_i - \lambda_{i+1} \geq \delta\} \tag{2.16}$$

El algoritmo iterativo propuesto por Onatski (2010) se presenta a continuación:

1. Calcular los valores propios $\lambda_1, \ldots, \lambda_n$ de la matriz de covarianza muestral $\hat{\Sigma}_Y = \frac{1}{T}\sum_{t=1}^{T} Y_t Y_t'$. Sea $j = r_{\text{máx}} + 1$.

2. Calcular $\hat{\beta}$, la pendiente de la regresión por mínimos cuadrados con una constante, siendo $\lambda_j, \ldots, \lambda_{j+4}$ las variables explicadas y con $(j-1)^{\frac{2}{3}}, \ldots, (j+3)^{\frac{2}{3}}$ las variables explicativas. Sea $\delta = 2|\hat{\beta}|$.

3. Calcular $\hat{r}(\delta) = \text{máx}\{i \leq r_{\text{máx}} : \lambda_i - \lambda_{i+1} \geq \delta\}$, o bien, si $\lambda_i - \lambda_{i+1} < \delta$ para todo $i \leq r_{\text{máx}}$ sea $\hat{r}(\delta) = 0$.

4. Sea $j = \hat{r}(\delta) + 1$. Repita los pasos 2 y 3 hasta converger.

2.4.3. Razones de valores propios de Ahn y Horenstein

El criterio de Ahn y Horenstein (2013) consiste en encontrar el valor de k que maximice un par de razones de valores propios. Se basa en el hecho de que los r valores propios más grandes de la matriz de covarianzas muestral $\hat{\Sigma}_Y = \frac{1}{T}\sum_{t=1}^{T} Y_t Y_t'$ divergen conforme N incrementa, mientras que el resto de los valores propios crecen asintóticamente.

Sea $k = 0, \ldots, r_{\text{máx}}$, los coeficientes a maximizar son:

$$
\begin{aligned}
ER(k) &= \frac{\hat{\lambda}_k}{\hat{\lambda}_{k+1}}, \\
GR(k) &= \frac{\ln\left[\frac{S_{k-1}(\hat{F}_t,\hat{P})}{S_k(\hat{F}_t,\hat{P})}\right]}{\ln\left[\frac{S_k(\hat{F}_t,\hat{P})}{S_{k+1}(\hat{F}_t,\hat{P})}\right]} = \frac{\ln\left(1+\hat{\lambda}_k^*\right)}{\ln\left(1+\hat{\lambda}_{k+1}^*\right)},
\end{aligned}
$$

con:

$$
\begin{aligned}
\hat{\lambda}_0 &= \frac{1}{m}\sum_{k=1}^{m}\frac{\hat{\lambda}_k}{\ln(m)}, \\
\hat{\lambda}_k^* &= \frac{\hat{\lambda}_k}{\sum_{j=k+1}^{m}\hat{\lambda}_j},
\end{aligned}
$$

y donde *"ER"* son las siglas de "Razón de Eigenvalores" (*Eigenvalue Ratio*, en inglés) y *"GR"* son las siglas de "Razón de Crecimiento" (*Growth Ratio*, en inglés), y con el valor de $\hat{\lambda}_0$ elegido de acuerdo con Ahn y Horenstein (2013).

2.5. Pruebas PANIC

En la definición del MFD en las ecuaciones (2.1), (2.2) y (2.3) se supone que los disturbios de los errores idiosincráticos ϵ_t tienen media cero y matriz de covarianza positiva-definida, entonces uno de los supuestos implícitos del MFD es que los errores idiosincráticos son estacionarios. Por otro lado, existe la posibilidad de que alguno o todos los componentes de F_t sean no estacionarios.

Bai y Ng (2004) propusieron aplicar pruebas de estacionariedad a los errores idiosincráticos y a los factores de manera separada, y se conocen como pruebas PANIC.

Dado que se supone que el MFD está correctamente estimado por los métodos de PC y 2SKS, las pruebas PANIC son de particular interés para verificar la estacionariedad de los errores idiosincráticos.

La idea principal de las pruebas PANIC consiste en que si solamente se tiene un factor ($r = 1$), entonces se realiza una prueba ADF tradicional a F_t. Si $r > 1$, entonces la prueba PANIC determinará el número de tendencias estocásticas independientes r_1 detrás de los r factores comunes y, además verificará si existe una raíz unitaria en cada uno de los errores idiosincráticos Bai y Ng (2004).

Para el caso $r > 1$ se consideran dos estadísticos. El primer estadístico filtra a los factores bajo el supuesto de que tienen una representación VAR(k), y se denota como MQ_f. El segundo corrige la correlación serial de forma arbitraria mediante la estimación no paramétrica de los parámetros de ruido relevantes, y se denota como MQ_c. Los estadísticos son de la forma:

- $MQ_c(m) = T[v_c^c(m) - 1]$.
- $MQ_f(m) = T[\hat{v}_f^c(m) - 1]$.

Donde cada $v_c^c(m)$ y $\hat{v}_f^c(m)$ son los valores propios más pequeños de cierto estadístico calculado a partir de una representación VAR(p), teniendo dos variaciones, una representación con intercepto y otra representación con intercepto y tendencia. Los detalles se pueden encontrar en Bai y Ng (2004).

Los estadísticos $MQ_c^c(m)$ y $MQ_f^c(m)$ no tienen una forma cerrada, por lo que los valores críticos se obtienen mediante simulación Monte Carlo.

2.5.1. Pruebas PANIC para errores idiosincráticos

Para probar si los errores idiosincráticos tienen al menos una raíz unitaria, Bai y Ng (2004) desarrollaron un par de pruebas ponderadas conocidas como *pruebas pooled*, o pruebas PANIC para errores idiosincráticos.

Sean ϵ_{it} los errores idiosincráticos del MFD. Supóngase que los ϵ_{it} son independientes con respecto al índice i, y considérense las hipótesis $H_0 : \rho_i = 1$ vs. $H_1 : \rho_i < 1$ para algún i. Sean $P_{\hat{\epsilon}}^c(i)$ y $P_{\hat{\epsilon}}^\tau(i)$ los p-valores asociados con las pruebas $ADF_{\hat{\epsilon}}^c(i)$ y $ADF_{\hat{\epsilon}}^\tau(i)$, respectivamente. Entonces:

$$P_{\hat{\epsilon}}^c = \frac{-2\sum_{i=1}^N \log P_{\hat{\epsilon}}^c(i) - 2N}{\sqrt{4N}} \xrightarrow{d} N(0,1),$$
$$P_{\hat{\epsilon}}^\tau = \frac{-2\sum_{i=1}^N \log P_{\hat{\epsilon}}^\tau(i) - 2N}{\sqrt{4N}} \xrightarrow{d} N(0,1).$$

Las pruebas PANIC para errores idiosincráticos se utilizan para probar la hipótesis nula de que $\rho_i = 1$, para todo i, por tanto, también se pueden interpretar como pruebas para contrastar la hipótesis nula de no-cointegración en los errores idiosincráticos, pues bajo la hipótesis nula no se pueden formar combinaciones estacionarias de Y_t.

2.6. Regla de combinación

La RC, propuesta por Guerrero y Nieto (1999), es un un método óptimo de estimación que permite, a través de estimaciones preliminares, obtener ajustes finales que satisfacen restricciones contemporáneas y/o temporales según las condiciones impuestas. Particularmente, Guerrero y Nieto (1999) usan la RC en el contexto de desagregación, es decir, se busca que una serie de tiempo de frecuencia mayor tenga óptimamente, desde el punto de vista estadístico, un nivel de frecuencia menor.

El fundamento teórico del algoritmo se basa en la minimización del Error Cuadrático Medio (ECM) a partir de la obtención del Mejor Estimador Lineal Insesgado (MELI).

Sea el $IDAE$ observado de $1, \ldots, T$ el indicador económico a estimar donde T es el número total de días, sea F una estimación preliminar del $IDAE$ observado de $1, \ldots, T$ y sea el IGAE la variable auxiliar observada de $1, \ldots, n$, donde n es el número total de meses, tal que n está incluída en T ($n \in T$). El IDAE y el IGAE guardan la relación:

$$IGAE_i = \sum_{j=1}^{T} c_{ij} IDAE_{ij}, \quad i = 1, \ldots, n, \tag{2.17}$$

donde las c_{ij} son constantes conocidas [1] no todas iguales a 0, que se guardan en la matriz C, de tal forma que las relaciones de la ecuación (2.17) se pueden expresar a través de la igualdad:

$$IGAE = C \times IDAE.$$

Supóngase que $Z_t = F + S_t$, donde S_t es un proceso VAR(p) estacionario, de tal forma que:

$$\Pi(L) S_t = a_t,$$

donde $\Pi(L)$ contiene a las matrices de los coeficientes autorregresivos de a_t, donde a_t es un proceso ruido blanco vectorial tal que $\mathbb{E}[a_t] = 0$ y $\Sigma_a = \mathbb{E}\left(a_t a_t'\right)'$.

Guerrero y Nieto (1999) probaron un resultado que justifica que el MELI del $IDAE$ dados F y el $IGAE$ es:

$$\widehat{IDAE} = F + A(IGAE - CF),$$

con:

$$Cov(\widehat{IDAE} - IDAE|F) = (I_T - AC)\Pi^{-1}(P \otimes \Sigma_a)\Pi^{-1'},$$

y:

$$A = \Pi^{-1}(P \otimes \Sigma_a)\Pi^{-1'} C'[C\Pi^{-1}(P \otimes \Sigma_a)\Pi^{-1'} C']^{+},$$

[1] En la aplicación las constantes c_{jl} se calculan como $\frac{1}{\text{número de días}}$ de cada mes

donde el símbolo $\otimes$ representa al producto de Kronecker y el superíndice $+$ denota a la pseudoinversa de Moore-Penrose.

El algoritmo de la RC de Guerrero y Nieto (1999) para obtener $\widehat{IDAE}$ es el siguiente:

1. Sea el $IDAE$ la serie a estimar observado de $1, \ldots, T$ donde T es el número de días, y sea el $IGAE$ la serie auxiliar observada de $1, \ldots, n$, donde n es el número total de meses. Ajuste y valide un modelo AR(p) a la estimación preliminar F, escogiendo el valor de p ya sea por criterios como AIC, BIC o el de Hannan-Quinn.

2. Use el modelo AR(p) de la serie preliminar F para calcular $\hat{A}$. Luego, a partir de mínimos cuadrados, calcule $\widehat{IDAE}$ y la matriz de covarianzas $\hat{\Sigma}$.

3. Calcule la diferencia $D_t^* = \widehat{IDAE} - F$.

4. Mediante pruebas Portmanteu y Breusch-Godfrey, pruebe si la diferencias D_t^* es no autocorrelacionada.

5. Si la hipótesis nula del paso 4 no se rechaza, termine el algoritmo. Si la hipótesis nula del paso 4 se rechaza suponga que $\Lambda D^* = (Q \otimes I)\, \Pi D^* = u$, donde Q es una matriz no singular tal que $QPQ' = I$,y note que:

$$\mathbb{E} = [uu'|W] = (Q \otimes I)\,(P \otimes \Sigma)\,(Q' \otimes I) = I \otimes \Sigma,$$

   y:

$$\Psi = \Pi^{-1}\,(P \otimes \Sigma)\,\Pi^{-1'} = \Lambda^{-1}\,(I \otimes \Sigma)\,\Lambda^{-1'}$$

   Repita los pasos 2 y 3 para D_t^* usando Mínimos Cuadrados Generalizados (GLS, por sus siglas en inglés) para obtener $\widetilde{IDAE}$ y $\hat{\Sigma}$ con una estructura de autocorrelación adecuada.

6. Verifique empíricamente si D_t^* es insesgado y estacionario. Puede usar el método propuesto por Guerrero y Corona (2018).

2.7. Mínimos Cuadrados Parciales

El modelo de regresión PLS es un modelo que, a partir de un conjunto de variables explicativas $X = (X_1, \ldots, X_N)'$ y una variable respuesta Y, construye un conjunto de componentes $Z_1, \ldots, Z_M$ formados por combinaciones lineales de las X_i. PLS tiene la particularidad de ser un modelo de *aprendizaje supervisado*, y además de identificar a las variables explicativas X_i que tienen una mayor variabilidad, le da un mayor peso a las X_i que tienen una correlación más alta con la variable respuesta Y, entonces PLS tiene la ventaja de que permite seleccionar a las X_i fuertemente correlacionadas con Y.

El algoritmo que estima PLS comienza con la descompsición en valores propios del producto matricial:

$$Q = YX'XY'$$

donde Z_1 es el vector propio asociado al valor propio más grande de Q. El segundo componente Z_2 se calcula a partir de los residuales de $e = Y - WZ_1$, donde e se puede interpretar como la información restante que no fue explicada Z_1. El algoritmo iterativo se repite M veces para identificar a los $Z_1, \ldots, Z_M$ componentes PLS.

2.8. Clústering para series de tiempo

En el análisis de series de tiempo puede resultar de interés agrupar series con características similares. Una técnica para ello es el clústering de series de tiempo, que consiste en calcular una distancia adecuada entre series a pares y asignar cada serie a alguno de los k clústers, para un valor k elegido por el analista.

2.8.1. Dynamic Time Warping

Sean $X_t = \{x_1, x_2, \cdots, x_n\}$ y $Y_t = \{y_1, y_2, \cdots, y_n\}$ dos series de tiempo. Para poder hacer comparaciones entre ellas primero es necesario definir una medida de

distancia adecuada. La primera opción es aplicar la distancia euclidiana definida como:

$$d(X_t, Y_t) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

Pero esta distancia no permite contemplar la dependencia temporal entre las series. Además, es una distancia rígida, en el sentido de que las comparaciones son entre las observaciones en el mismo tiempo, lo cual solo permite calcular distancias entre series de la misma longitud.

Dynamic Time Warping (Deformación Dinámica del Tiempo) es un algoritmo que busca la alineación no-lineal óptima entre dos series. Fue introducido por Sakoe (1971) con aplicaciones en reconocimiento de voz.

La alineación óptima puede ser calculada recursivamente por:

$$D(X_i, Y_i) = \delta(x_i, y_i) + \text{mín} \left\{ \begin{array}{c} D(x_{i-1}, y_{j-1}) \\ D(x_i, y_{j-1}) \\ D(x_{i-1}, y_j) \end{array} \right\} \tag{2.18}$$

donde X_i es la subsecuencia $\{x_1, \cdots, x_i\}$ y Y_i es la subsecuencia $\{y_1, \cdots, y_i\}$. La distancia completa está dada por $D(X_t, Y_t) = D(X_n, Y_n)$.

Puede calcular distancias entre series de diferente longitud, lo cual es una ventaja con respecto a la distancia euclidiana.

2.8.2. Soft DTW k-means

El algoritmo Soft DTW k-means es un método para clústering de series de tiempo propuesto por Kaufman y Rousseeuw (2009). El objetivo es, dados K centroides elegidos al azar, asignar cada serie al clúster cuya distancia DTW sea la mínima, pero la distancia es ponderada por la plausibilidad de que la i-ésima serie pertenezca al k-ésimo clúster. El cálculo de los centroides también es ponderado.

El algoritmo Soft DTW k-means es el siguiente:

1. Sea K el número de clústers. Escoja aleatoriamente K series para ser los m_k

centroides iniciales.

2. Sea el conjunto de series $Y = \{y_1, y_2, \cdots, y_n\}$. Asigne cada serie y_i, al centroide más cercano, en términos de la distancia DTW:

$$\min_{1 \leq k \leq K} \gamma_{ik}^q DTW(y_i, m_k)^2,$$

con:

$$\gamma_{ik}^q = \frac{1}{\sum_{j=1}^{K} \left(\frac{DTW(y_i, m_k)}{DTW(y_i, m_j)} \right)^{\frac{1}{2}}},$$

$$m_k = \frac{\sum_{i=1}^{n} \gamma_{ik}^q y_i}{\sum_{i=1}^{n} \gamma_{ik}^q},$$

$$\sum_k \gamma_{ik} = 1.$$

3. Para cada clúster, recalcule su centroide basado en las series de ése clúster.

4. Regrese al Paso 2 mientras no se cumpla el criterio de convergencia, en otro caso se termina el algoritmo.

Referencias con una importante colección de métodos de clústering para series de tiempo son Liao (2005) y Maharaj, D'Urso, y Caiado (2019).

Capítulo 3

Aplicación

En este capítulo se ajustará un MFD estimado a partir de un conjunto de variables tradicionales y un conjuto de variables no tradicionales, siendo seleccionadas mediante PLS. Se realiza la estimación por el método PC y posteriormente por le método 2KSK. Se aplican las pruebas PANIC para verificación de supuestos. Finalmente, a partir del factor estimado, y tomando como variable auxiliar al IGAE, se construye al indicador económico diario $\widehat{IDAE}$.

3.1. Variables consideradas

Las variables consideradas se dividen en 3 grupos, en el primer grupo se encuentra el IGAE, en el segundo grupo se tienen variables tradicionales y en el tercer grupo variables no tradicionales.

3.1.1. IGAE

La primer variable considerada es el IGAE, que es un indicador mensual que permite cuantificar la actividad económica en México, cuyas componentes son similares a las del Producto Interno Bruto (PIB). El IGAE es un estadístico oficial calculado por el Instituto Nacional de Estadística y Geografía (INEGI) y está conformado por un 94.7 % de la información del Valor Agregado Bruto (VAB).

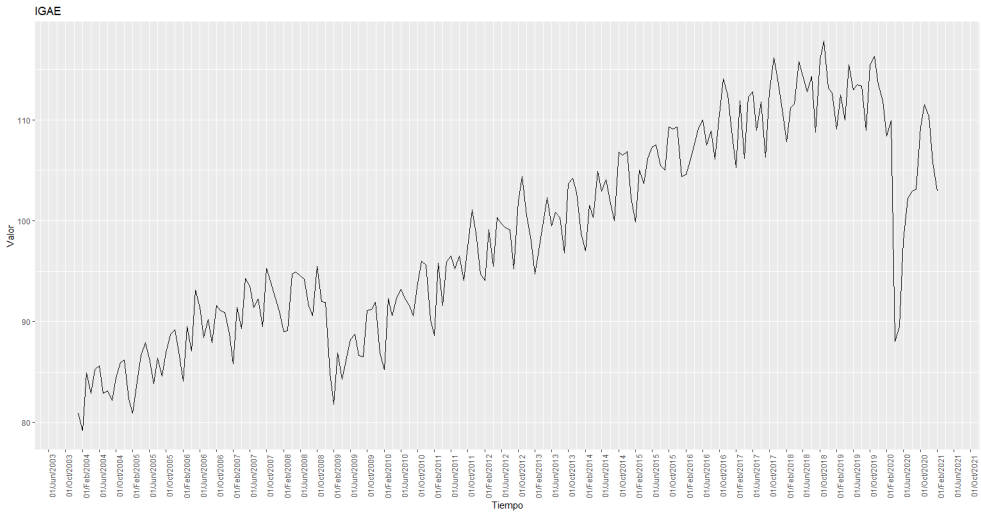

Figura 3.1: IGAE.

Se obtuvo la serie mensual del IGAE del periodo del 1/enero/2004 al 1/febrero/2021, como se puede observar en la Figura 3.1. En general el IGAE tiene una tendencia creciente, pero tuvo un descenso fuerte en la segunda mitad del 2008 y los primeros meses de 2009, coincidiendo con la crisis económica mundial del 2008, y además, tuvo un descenso muy fuerte en el mes de abril de 2020, coincidiendo con la pandemia del COVID-19. A partir de junio de 2020 se tiene un incremento en la tendencia, pero comienza a descender a partir de enero de 2021.

3.1.2. Variables tradicionales

Las variables tradicionales fueron propuestas a partir de fundamentos financieros, macroeconómicos y de conocimiento empírico. Por ejemplo, se propusieron a la Tasa de rendimiento de CETES a 28 días, la Tasa de interés interbancaria de equilibrio a 28 días y el Tipo de cambio dólar-peso (USD/MXN) por ser variables económicas clásicas. Por otro lado, se propusieron a los índices Bovespa, Dow Jones, IPC y S&P500 por que la teoría financiera dicta que son representativos del comportamiento de las series que los conforman. Además, se proponen los precios al cierre de las acciones de 9 empresas que cotizan en la Bolsa Mexicana de Valores (BMV), como posibles representates de la situación económica de México. El periodo de observación de las variables es del 1/enero/2004 al 28/feb/2021.

En la Tabla 3.1 se encuentran las abreviciones de las variables tradicionales, una breve descripción y la fuente de información.

Abreviación	Descripción	Fuente
ALFAA.MX	Grupo Industrial Alfa	Yahoo! Finance
AMXL.MX	América Móvil	Yahoo! Finance
IBOV	Índice Bovespa de la Bolsa de Valores de São Paulo	Yahoo! Finance
CETES_28	Tasa de rendimiento de CETES a 28 días	Banco de México
DJI	Índice bursátil Dow Jones	Yahoo! Finance
ELEKTRA.MX	Grupo Elektra	Yahoo! Finance
FEMSAUBD.MX	Fomento Económico Mexicano (FEMSA)	Yahoo! Finance
GFINBURO.MX	Grupo Financiero Inbursa	Yahoo! Finance
GFNORTEO.MX	Grupo Financiero Banorte	Yahoo! Finance
IPC	Índice de Precios y Cotizaciones de la BMV	Yahoo! Finance
KIMBERA.MX	Kimberly-Clark de México	Yahoo! Finance
S&P 500	Índice bursátil Standard & Poor's 500	Yahoo! Finance
TIIE_28	Tasa de interés interbancaria de equilibrio a 28 días	Banco de México
USD_MXN	Tipo de cambio dólar-peso (USD/MXN)	Banco de México
TLEVISACPO.MX	Grupo Televisa	Yahoo! Finance
WALMEX.MX	Walmart de México y Centroamérica	Yahoo! Finance

Tabla 3.1: Variables tradicionales

Se realizó una implementación en python a partir del módulo *yfinance* que permite descargar series financieros desde Yahoo! Finance de manera automática.

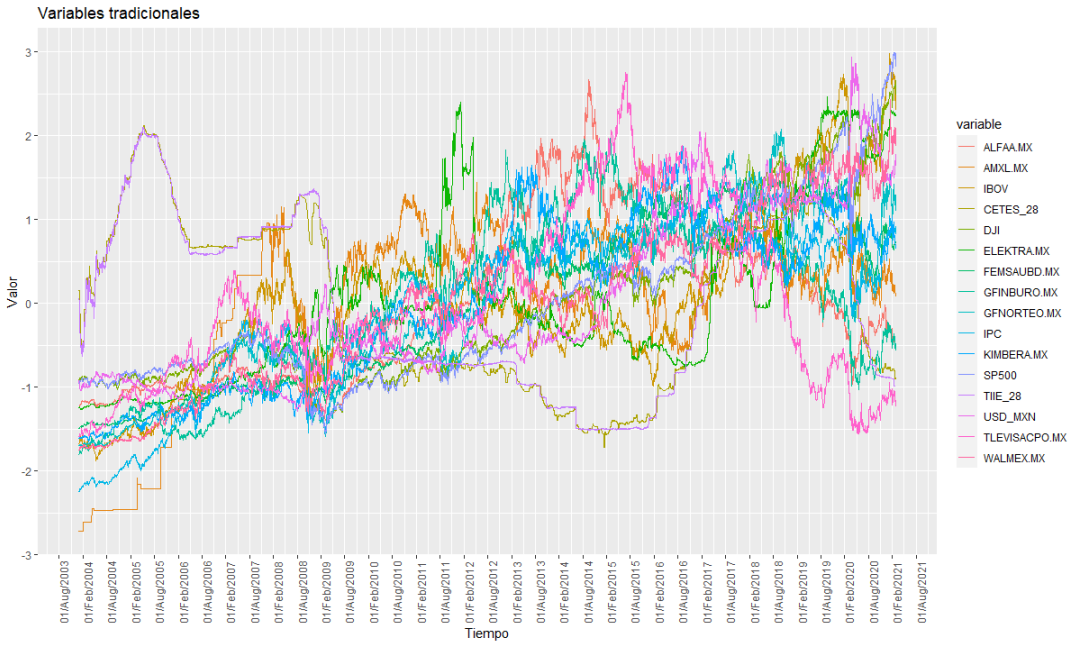

Figura 3.2: Variables tradicionales.

En la Figura 3.2 se presentan las series de las 16 variables tradicionales, se puede

ver que tienen diferentes comportamientos en media y varianza, pues reciben el impacto de las crisis económicas de manera diferente, algunas series incrementan tendencia como el tipo de cambio USD-MXN, pero otras caen como el IPC.

3.1.3. Variables no tradicionales

Los GT son un conjunto de datos no tradicionales conformados por solicitudes de búsqueda realizadas al buscador de Google. Los GT están definidos en el intervalo del 0 al 100, y están escalados de tal forma que 100 representa la unidad de tiempo en la que hubo más búsquedas, y 0 en los días en los que hubo menos búsquedas. La unidad de tiempo pueden ser horas, días, meses o años. Además, las búsquedas se pueden focalizar a una localización geográfica

En el presente trabajo se consideraron 65 tópicos de GT, con búsquedas en México, y en una ventana de observación diaria del 1/enero/2004 al 28/febrero/2021. Los tópicos se propusieron basados en temas de interés popular, como son la economía mexicana, la pandemia de COVID-19, efectos de la vacunación, relación con Estados Unidos, inseguridad en México, etc.

Los tópicos propuestos son los siguientes:

Aeropuerto	Afores	AH1N1	AMLO	AstraZeneca
Ayotzinapa	Biden	BioNTech	Brasil	Calderon
Cansino	Cartel	Casa Blanca	Chapo	China
Contagios	Coronavirus	Corrupcion	Crisis economica	Cuarentena
Cubrebocas	Desempleo	Diputado	Dolar	Elecciones
EPN	Estados Unidos	Gasolina	Gobernador	Homicidios
Inflacion	Inoculacion	Inseguridad	Israel	Mascarilla
Mascarilla N95	Medidas economicas	Migracion	Migrantes	Morena
Muertos	Muro	Narcotrafico	Outsourcing	Pacto
PAN	Pandemia	PEMEX	Peso	Petroleo
Pfizer	PRI	Reactivacion	Recesion	Recuperacion
Reformas	Reforma fiscal	Salario	Sismo	Sputnik V
Tipo de cambio	Trump	Vacuna	Vacunacion	Variante covid

Tabla 3.2: Tópicos propuestos de GT.

Se realizó una implementación en python a partir del módulo *pytrends* que permite descargar de manera automática tópicos de GT.

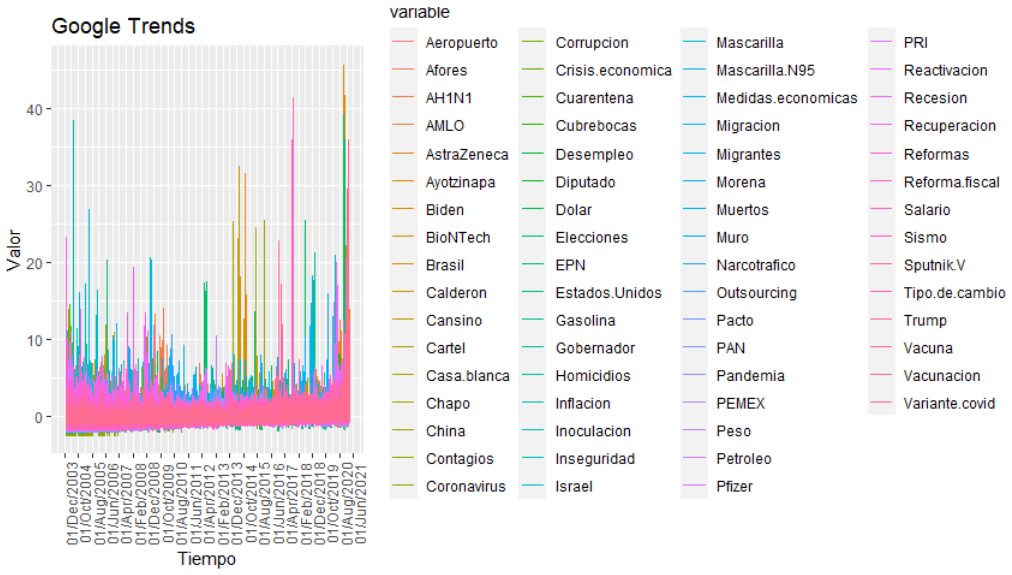

Figura 3.3: Tópicos de GT.

En la Figura 3.3 se presentan las series de los 65 tópicos de GT propuestos, se encuentran estandarizados y se puede observar que tienen diferentes picos dependiendo del momento en el que fue más buscado. La media se encuentra alrededor del cero, pero la varianza no es constante.

3.2. Preprocesamiento de las series

Con el objetivo de poder identificar las variables que están correlacionadas con el IGAE se realizó una implementación en R para agregar las series de manera mensual, es decir, se realizó una implementación que crea nuevas series con los promedios mensuales de las variables tradicionales y de las variables no tradicionales.

3.2.1. Preprocesamiento de las variables tradicionales

En la Figura 3.4 se encuentran las variables tradicionales agregadas de manera mensual, lo cual permite observar de manera suavizada el comportamiento de la media.

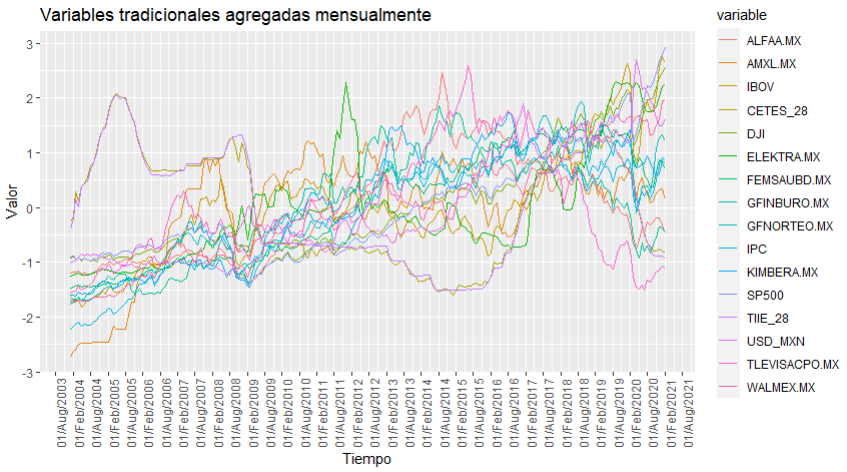

Figura 3.4: Variables tradicionales agregadas mensualmente.

La agregación mensual de las series permite analizar con mayor detenimiento la tendencia, se observa que el impacto de la pandemia fue diferente para cada series, pues en algunas la tendencia fue a la alza, mientras que en otras fue a la baja.

3.2.2. Preprocesamiento de las variables no tradicionales

En la Figura 3.5 se encuentran las variables no tradicionales agregadas de manera mensual, con el objetivo de apliar un suavizamiento sobre la media.

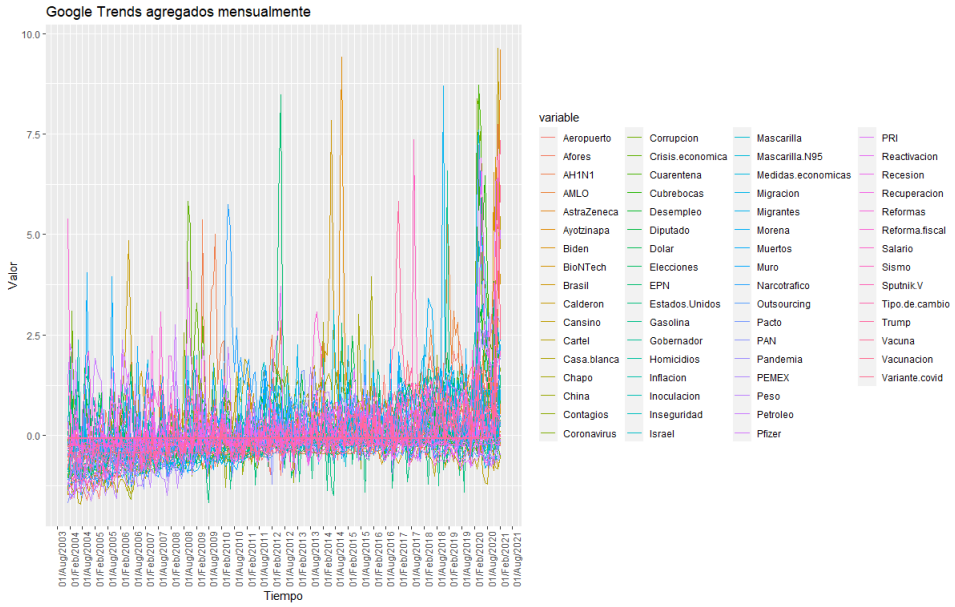

Figura 3.5: GT agregados mensualmente.

Nótese a partir de la gráfica es difícil distinguir a cada una de las variables, además de que no se observan tendencias significativas, y aún se mantienen muchos picos.

3.3. Selección de las variables significativas

Habiendo agregado a las variables tradicionales y a las variables no tradicionales se ajustaron modelos PLS para poder identificar a las variables que son altamente correlacionadas con el IGAE.

3.3.1. Selección de las variables tradicionales

Sea X el conjunto de datos formado por las variables tradicionales agregadas mensualmente y estandarizadas, y sea la variable respuesta Y el IGAE.

Se ajustaron $S = 1000$ modelos PLS con muestras dependientes de los datos y se calcularon intervalos de confianza bootstrap a los pesos del primer componente. Se seleccionaron como variables relevantes aquellas cuyo intervalo de confianza bootstrap al 99 % no contuvieran al 0, es decir, aquellos pesos que fueron distintos de 0 a una significancia del 1 %.

Las variables seleccionadas son ALFAA.MX, DJI, FEMSAUBD.MX, GFINBURO.MX, GFNORTEO.MX, IPC, KIMBERA.MX, SP500, USD_MXN, TLEVISACPO.MX y WALMEX.MX y se pueden observar en la Figura 3.6.

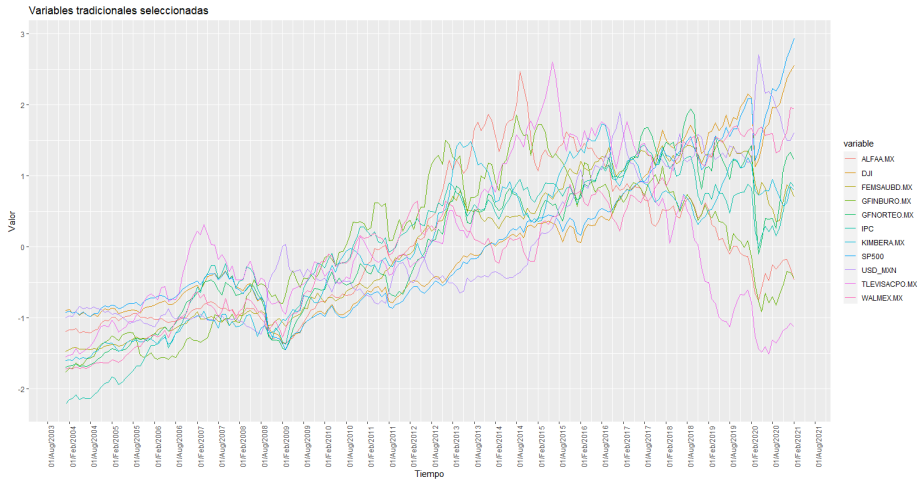

Figura 3.6: Variables tradicionales seleccionadas.

Nótese que, en general, durante gran parte del periodo de observación las series tienen una tendencia creciente, pero con una caída en la crisis del 2008. La mayoría de las series tienen una caída significativa en marzo de 2020, consecuencia de la pandemia del COVID-19, sin embargo, en los últimos meses de observación las series tienen comportamientos muy diferentes, pues mientras unas no logran recuperarse, otras crecen conisderablemente.

Clústering de las variables tradicionales

Figura 3.7: Clústering aplicado a las variables tradicionales con $k = 3$.

Con el objetivo de analizar con mayor detenimiento el comportamiento de las variables tradicionales seleccionadas se ajustó el algoritmo Soft DTW k-means para clústering de series de tiempo. Se eligió el valor $k = 3$ ya que es el valor de k que permitió una configuración más interpretable. Los clústers ajustados se encuentran la Figura 3.7, donde todos los subplots tienen en línea sólida y color rojo el centroide del clúster.

Se puede observar que el clúster 1 está conformado por las variables ALFA.MX, GINBURO.MX y TLEVISACPO.MX; estas series tenían una tendencia creciente, llegaron a cierto pico en el periodo de observación 150 y comenzaron a decrecer, de tal forma que no se lograron recuperar. En el clúster 2 se encuentran las variables FEMSAUBD.MX, GFNORTEO.MX, IPC y KIMBERA.MX, que son el conjunto de series que tenían una tendencia creciente, pero que resintieron el impacto de la pandemia y que apenas están retomando los niveles que tenían antes de la contingencia. En el clúster 3 se encuentran las variables DJI, S&P500, USD_MXN y WALMEX.MX, que son las series que tienen una tendencia creciente y que son lo suficientemente robustas como para resistir los efectores de la pandemia y continúan creciendo.

Resulta de particular interés notar que la serie del tipo de cambio USD_MXN guarda una relación inversamente proporcional a la serie de WALMEX.MX, es decir, cuando el tipo de cambio dólar-peso sube las acciones de WALMEX.MX caen, y viceversa.

Una de las conclusiones exploratorias más importantes fue que, incluso para las variables tradicionales que fueron seleccionadas a través de PLS por estar altamente correlacionadas con el IGAE, existen conjuntos de series que tienen tendencias diferentes, ver por ejemplo en la Figura 3.7 el centroide del clúster 1 comparado con el centroide del clúster 3, es evidente que el centroide del clúster 1 refleja que para ése conjunto de variables tradicionales seleccionadas el impacto de la pandemia fue mayúsculo, mientras que para el conjunto de variables tradicionales seleccionadas del clúster 3 la pandemia llevó a un incremento en la tendencia. La idea anterior se puede desarrollar en el sentido de que el hecho de que una serie esté altamente correlacionada con el IGAE no necesariamente quiere decir que sigue un comportamiento similar

en todo el periodo de observación, incluso puede darse el caso de que hayan intervalos de tiempo en los que la relación sea inversamente proporcional.

Resulta notorio el hecho de que aplicar clústering de series de tiempo permite identificar características de conjuntos de series que no hubiesen sido posible observar a través de gráficas conjuntas.

Después, se ajustó un modelo de PCA con las variables tradicionales seleccionadas agregadas mensualmente (Figura 3.8) y se calculó la correlación entre el primer componente principal y el IGAE, obteniéndose una correlación de 0.9266 con un intervalo de confianza de (0.9045,0.9438) al 95 %, lo cual es un indicador de que las variables tradicionales seleccionadas y el IGAE están fuertemente correlacionados.

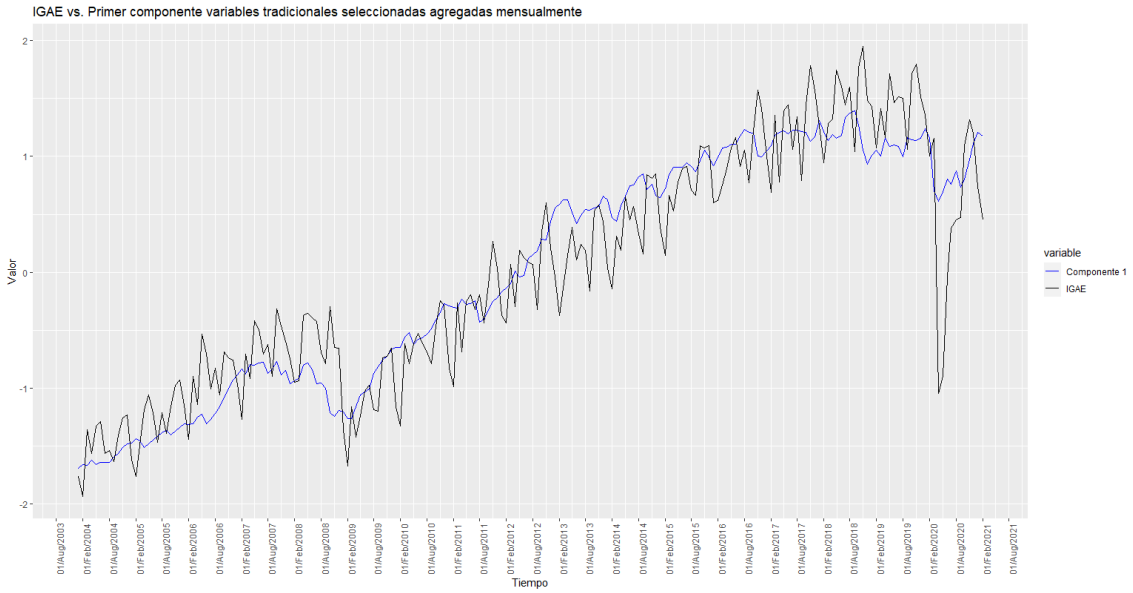

Figura 3.8: IGAE (en color negro) y primer componente de PCA con las variables tradicionales seleccionadas agregadas mensualmente (en color azul).

Nótese que ambas series muestran un comportamiento muy similar, lo cual es un indicador de que las variables tradicionales seleccionadas y el IGAE están altamente correlacionadas.

Se ajustó una regresión siendo el primer componente principal la variable explicativa y el IGAE la variable explicada para verificar que la relación no es espuria. Se aplicaron pruebas ADF con tendencia y sin tendencia a los factores residuales e_t de la regresión.

Hipótesis: H_0: Los residuales e_t tienen una raíz unitaria vs. H_1: Los residuales e_t son estacionarios.

Nivel de significancia: $\alpha = 0.05$.

p-valores obtenidos: 0.0110 y 0.01.

Resultado: A un nivel de significancia $\alpha = 0.05$ existe suficiente evidencia estadística para rechazar la hipótesis nula a favor de la hipótesis alternativa, es decir, existe suficiente evidencia para afirmar que los residuales e_t son estacionarios.

Se concluye que la relación entre el IGAE y el primer componente principal no es espuria.

Luego se ajustó un modelo de PCA con las variables tradicionales seleccionadas con observaciones diarias para poder analizar su comportamiento, véase la Figura 3.9.

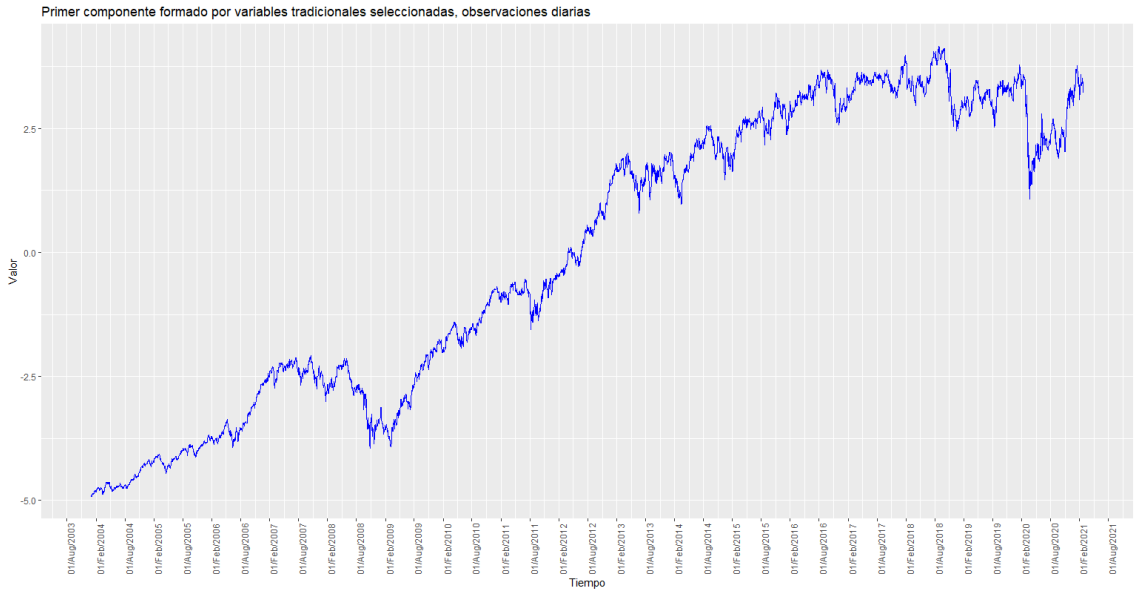

Figura 3.9: Primer componente de PCA con las variables tradicionales seleccionadas, observaciones diarias.

Nótese que sigue en general un comportamiento similar al del IGAE, pues se tiene una tendencia creciente hasta antes de la caída en 2008, para luego recuperarse y volver a caer en marzo de 2020, después se da paso a una recuperación pero en los últimos días la tendencia es decreciente.

3.3.2. Selección de las variables no tradicionales

Sea X el conjunto de datos formado por las variables no tradicionales agregadas mensualmente y estandarizadas, y sea la variable respuesta Y el IGAE.

Se ajustaron $S = 1,000$ modelos PLS con muestras dependientes de los datos y se calcularon intervalos de confianza bootstrap a los pesos del primer componente. Se seleccionaron como variables relevantes aquellas cuyo intervalo de confianza bootstrap al 99 % no contuvieran al 0, es decir, aquellos pesos que fueron distintos de 0 a una significancia del 1 %.

Los tópicos de GT seleccionados son "AstraZeneca", "Ayotzinapa", "Brasil", "Cansino", "Casa.blanca", "Chapo", "Cuarentena", "Diputado", "Dolar", "EPN", "Gasolina", "Gobernador", "Homicidios", "Inoculacion", "Israel", "Mascarilla", "Migrantes", "Morena", "Muro", "Outsourcing", "Pacto", "PAN", "PEMEX", "Peso", "Petroleo", "Reactivacion", "Recuperacion" y "Trump" y se encuentran en la Figura 3.10.

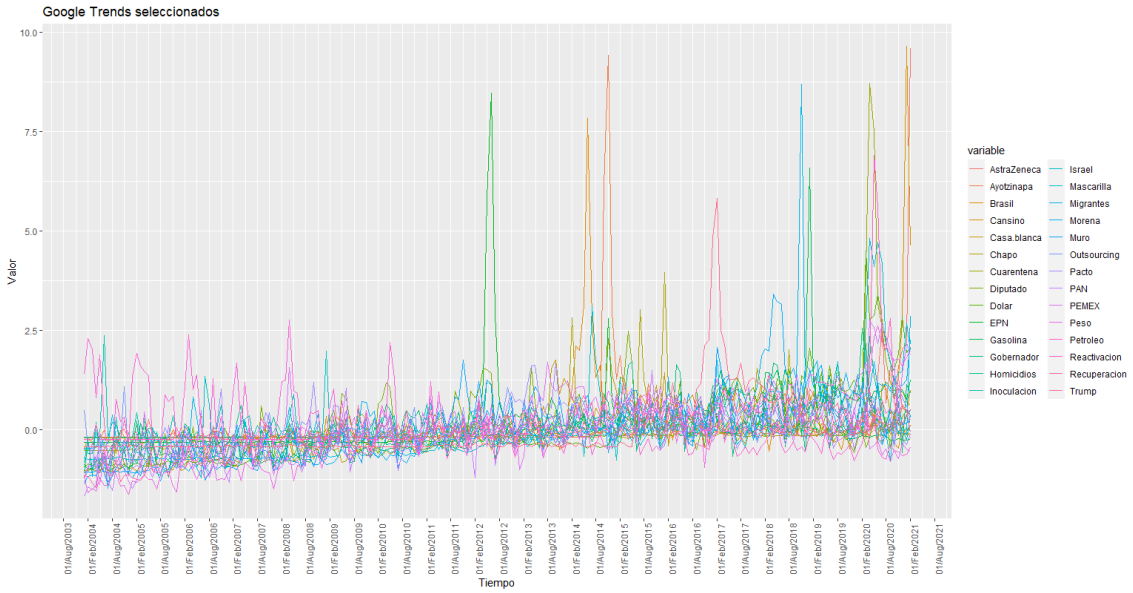

Figura 3.10: GT seleccionados.

En general, los GT seleccionados tienen una tendencia creciente, y a pesar de que están agregados mensualmente se observan muchos picos. La media crece con respecto al tiempo y la varianza también, por tanto las series no son estacionarias.

Se ajustó un modelo de PCA con los tópicos seleccionados agregados mensual-

mente (Figura 3.11) y se calculó la correlación entre el primer componente principal y el IGAE, obteniéndose una correlación de 0.8097 con el intervalo de confianza de (0.7567, 0.8520) al 95 %, lo cual es un indicador de que los tópicos seleccionados y el IGAE están fuertemente correlacionados.

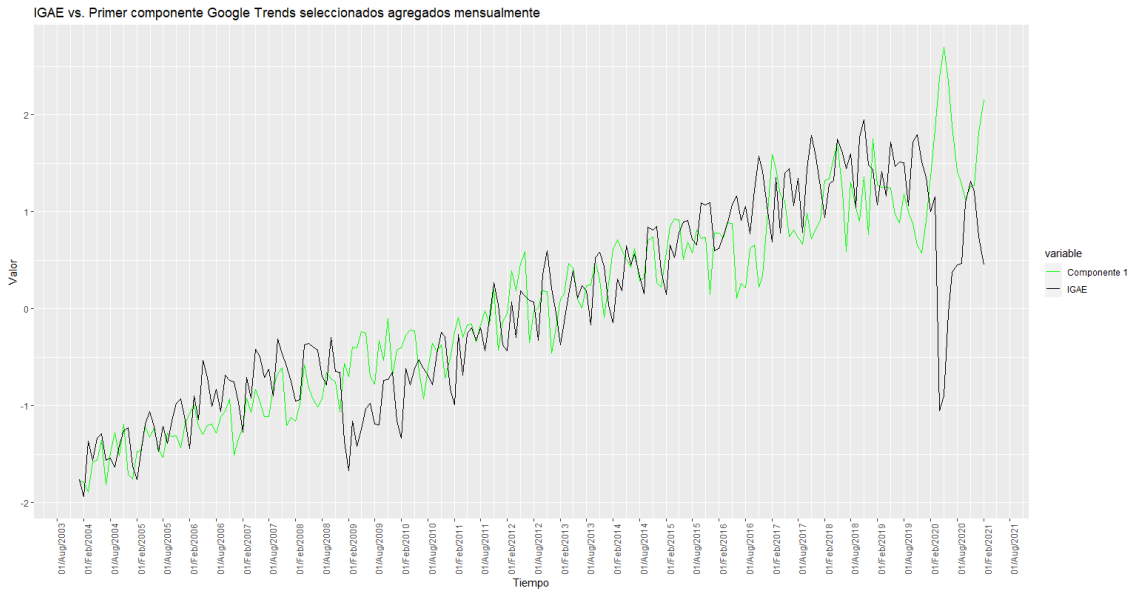

Figura 3.11: IGAE (en color negro) y primer componente de PCA con las variables no tradicionales seleccionadas agregadas mensualmente (en color verde).

Nótese que ambas series muestran un comportamiento similar, sin embargo, durante la crisis del 2008 cuando el IGAE cae los GT suben, y de manera similar, el IGAE tiene una caída abrupta en abril de 2020, mientras que los GT tienen un pico muy grande, y después, conforme el IGAE crece los GT disminuyen. Lo anterior es un indicador de que los tópicos de GT están relacionados con el IGAE, pero que la relación parece ser inversamente proporcional, es decir, conforme el IGAE disminuye los GT aumentan.

Se ajustó una regresión siendo el primer componente principal la variable explicativa y el IGAE la variable explicada para verificar que la relación no es espuria. Se aplicaron pruebas ADF con tendencia y sin tendencia a los factores residuales e_t de la regresión.

Hipótesis: H_0: Los residuales e_t tienen una raíz unitaria vs. H_1: Los residuales e_t

son estacionarios.

Nivel de significancia: $\alpha = 0.05$.

p-valores obtenidos: 0.01 y 0.01.

Resultado: A un nivel de significancia $\alpha = 0.05$ existe suficiente evidencia estadística para rechazar la hipótesis nula a favor de la hipótesis alternativa, es decir, existe suficiente evidencia para afirmar que los residuales e_t tienen son estacionarios.

Se concluye que la relación entre el IGAE y el primer componente principal no es espuria.

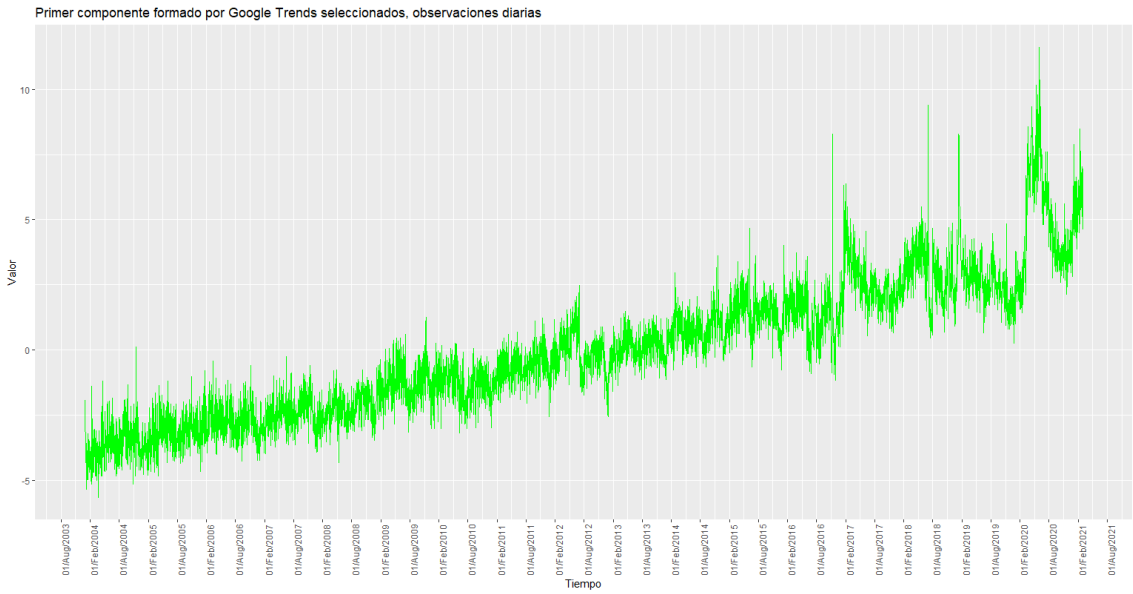

Figura 3.12: Primer componente de PCA con las variables no tradicionales seleccionadas, observaciones diarias.

Después se ajustó un modelo de PCA con los tópicos seleccionados diarios para poder analizar su comportamiento, véase la Figura 3.12. Se observa que la tendencia del componente es creciente, y que tiene un incremento en la varianza en la última tercera parte de las observaciones. Por otro lado, la tendencia del componente no sigue de manera adecuada al IGAE, pues crece cuando el IGAE disminuye, lo cual se nota durante el periodo de la pandemia del COVID-19.

El análisis anterior llevó a la conclusión de que los tópicos propuestos de GT no eran adecuados, pues no reflejaban adecuadamente el comportamiento de la actividad

económica en México, pues a pesar de que tópicos como “Cuarentena”, “Mascarilla” y “Reactivacion” están correlacionadas con el IGAE, en realidad lo que se tiene es una relación inversamente proporcional.

Lo anterior tiene aún más sentido si se toma en cuenta, por ejemplo, lo observado en el clúster 3 de la Figura 3.7, pues se encontró de manera exploratoria que la serie del tipo de cambio USD_MXN guarda una relación inversamente proporcional a la serie de WALMEX.MX, pues cuando el peso se deprecia con respecto al dólar el tipo de cambio sube, y por tanto los precios de los artículos de WALMEX.MX suben, lo que impacta negativamente en sus ventas, y por tanto los precios de sus acciones caen.

El objetivo de la presente tesis es construir un indicador económico de alta frecuencia que permita monitorear la economía mexicana, por tanto, se espera que el indicador refleje adecuadamente el movimiento económico de México durante el periodo de la pandemia, pero no es de particular interés modelar el comportamiento de la pandemia.

Por lo anterior, se proponen nuevos tópicos de GT, siendo temas que se espera que hayan tenido un descenso importante durante el periodo de la pandemia, y por tanto, que reflejen de mejor manera el comportamiento de la economía mexicana, es decir, se proponen tópicos de lo que se espera que las personas en México no hayan buscado en Google durante el periodo de la pandemia, pero que sí reflejen adecuadamente el consumo económico de las familias en México. La idea es identificar los bienes o servicios que los mexicanos dejaron de consumir durante la pandemia, y que por tanto llevaron a una caída económica, y por tanto, son los tópicos que mejor reflejan el consumo económico en México.

Los nuevos tópicos de GT propuestos son los siguientes:

Aeromexico	Aeropuerto	Antro	Auto	Avion	Banquetes
Bar	Beisbol	Buffete	Casa en venta	Cine	Circo
Hotel	Liverpool	Llantas	Maquillaje	Parque de diversiones	Playa
Rentadora de autos	Restaurante	Ropa	Salon de belleza	Tenis	Ticketmaster
Trafico	Traje	Vestido	Viajes	Vuelos	Zapatos

Tabla 3.3: Nuevos tópicos propuestos de GT.

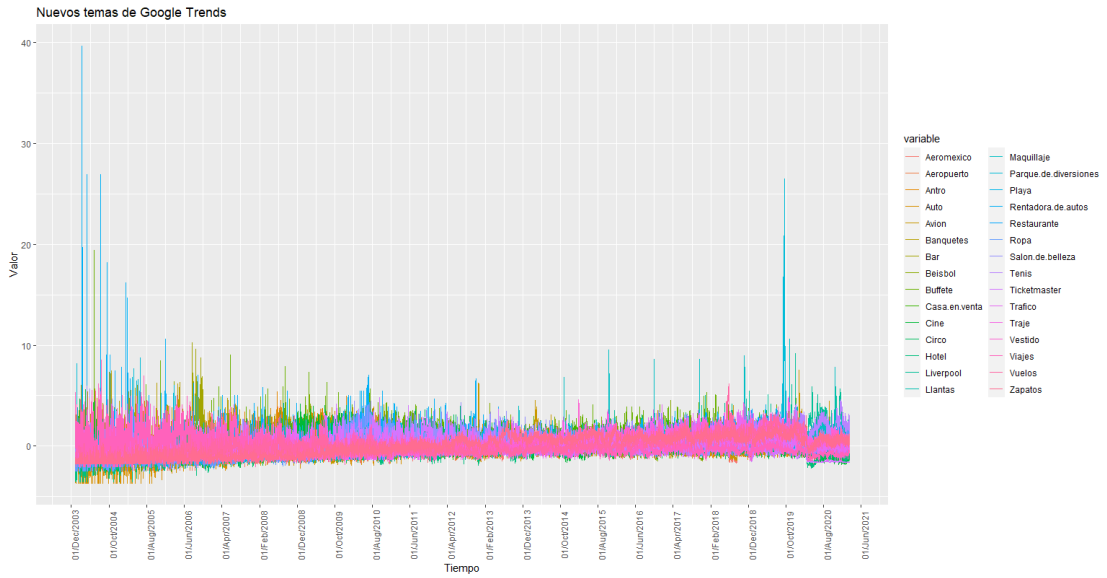

Figura 3.13: Nuevos tópicos de GT.

En la Figura 3.13 se presentan las series de los 25 nuevos tópicos de GT propuestos, donde se puede ver que las series tienen una media constante alrededor del cero, la varianza es constante para la mayoría de las series, aunque se pueden apreciar varios picos que resaltan.

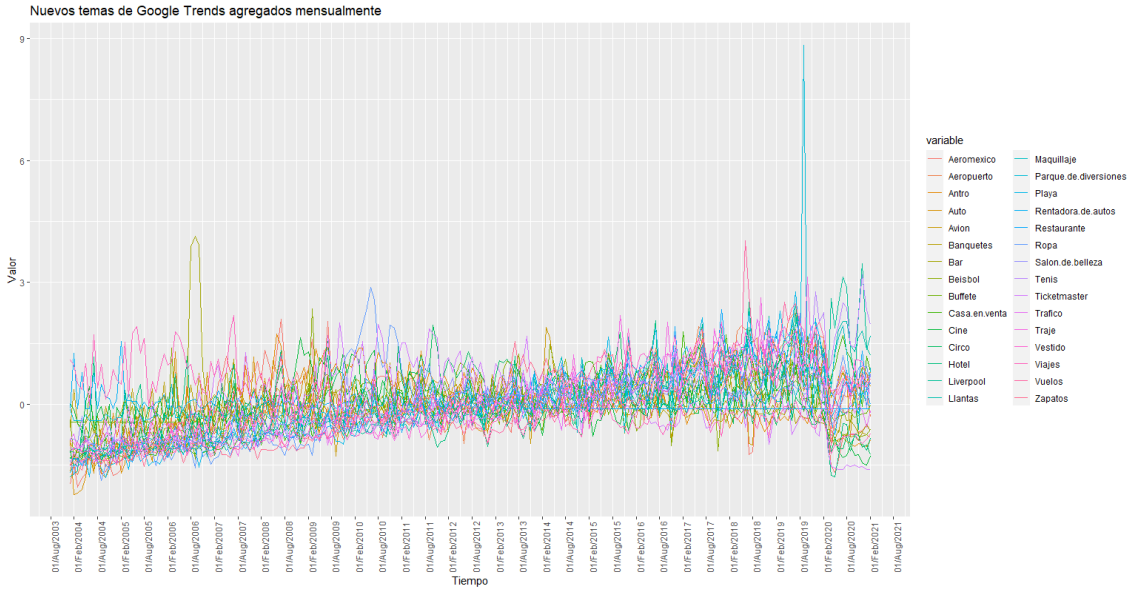

Figura 3.14: Nuevos tópicos de GT agregados mensualmente.

En la Figura 3.14 se presentan las series de los 25 nuevos tópicos de GT pro-

puestos agregados mensualmente, se tiene una media creciente pero que se encuentra alrededor del cero, también la varianza crece en la última cuarta parte del periodo de observación.

Sea X el conjunto de datos formado por los nuevos GT agregados mensualmente y estandarizados, y sea la variable respuesta Y el IGAE.

Se ajustaron $S = 1,000$ modelos PLS con muestras dependientes de los datos y se calcularon intervalos de confianza bootstrap a los pesos del primer componente. Se seleccionaron como variables relevantes aquellas cuyo intervalo de confianza bootstrap al 99 % no contuvieran al 0, es decir, aquellos pesos que fueron distintos de 0 a una significancia del 1 %.

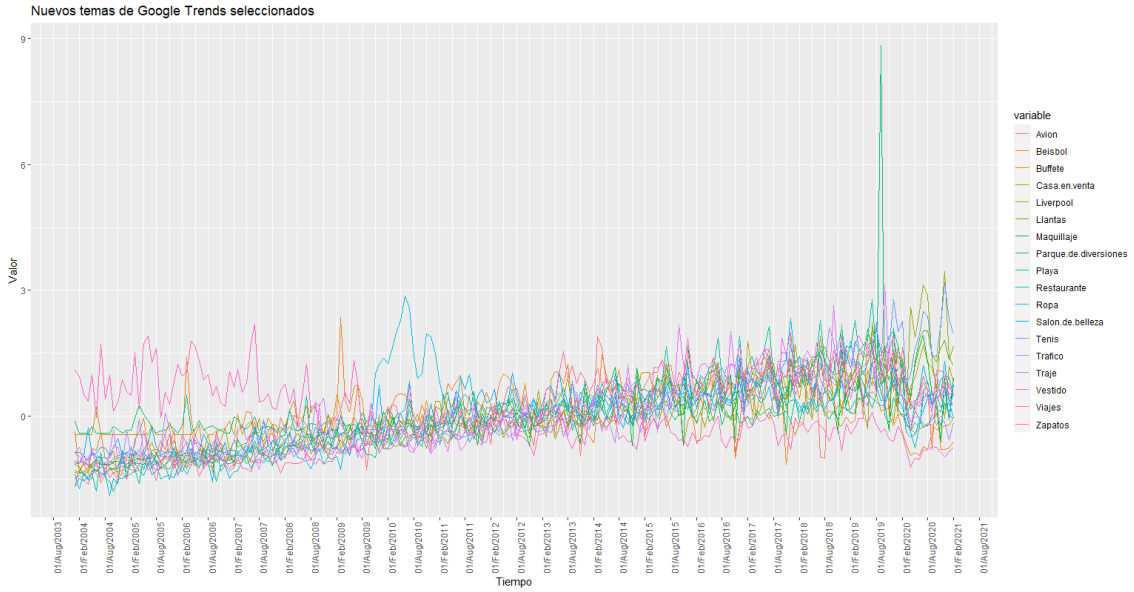

Figura 3.15: Nuevos tópicos de GT seleccionados.

Los nuevos tópicos de GT seleccionados son “Avion”, “Beisbol”, “Buffete”, “Casa.en.venta”, “Liverpool”, “Llantas”, “Maquillaje”, “Parque.de.diversiones”, “Playa”, “Restaurante”, “Ropa” “Salon.de.belleza”, “Tenis”, “Trafico”, “Traje”, “Vestido”, “Viajes” y “Zapatos”, y se pueden ver en la Figura 3.15; obsérvese que, por un lado, la media es creciente con respecto al tiempo,pero la varianza no es constante, y hay ciertos picos que resaltan muy por encima del promedio de las series.

Resulta de particular interés que en la Figura 3.15 existen menos picos que en los

primeros GT seleccionados, que se encuentran en la Figura 3.10, lo cual quiere decir que la varianza de los nuevos GT seleccionados es menor, lo cual puede llevar a un mejor ajuste del modelo.

Después, se ajustó un modelo de PCA con los nuevos GT seleccionados agregados mensualmente (Figura 3.16) y se calculó la correlación entre el primer componente principal y el IGAE, obteniéndose una correlación de 0.9201 con un intervalo de confianza de (0.8961, 0.9387) al 95 %, lo cual es un indicador de que los nuevos GT seleccionados y el IGAE están fuertemente correlacionados.

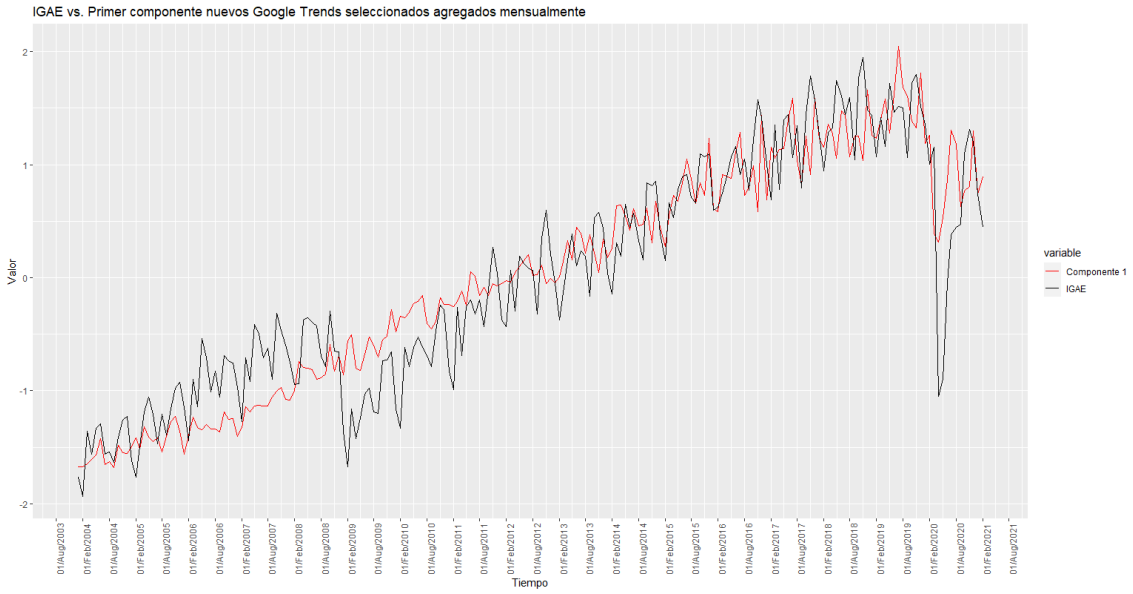

Figura 3.16: IGAE (en color negro) y primer componente de PCA con los nuevos GT seleccionados agregados mensualmente (en color rojo).

En la Figura 3.16 se encuentran la serie del IGAE (en color negro) y la serie del primer componente principal (en color rojo). Nótese que ambas series muestran un comportamiento muy similar, lo cual es un indicador de que los nuevos GT y el IGAE están altamente correlacionados, pues el componente refleja adecuadamente la caída del IGAE en abril de 2020, sin embargo, los GT no caen durante la crisis económica del 2008.

Luego se ajustó una regresión siendo el primer componente principal la variable explicativa y el IGAE la variable explicada para verificar que la relación no es espuria. Se aplicaron pruebas ADF con tendencia y sin tendencia a los factores residuales e_t

de la regresión.

Hipótesis: H_0: Los residuales e_t tienen una raíz unitaria vs. H_1: Los residuales e_t son estacionarios.

Nivel de significancia: $\alpha = 0.05$.

p-valores obtenidos: 0.05 y 0.01.

Resultado: A un nivel de significancia $\alpha = 0.05$ existe suficiente evidencia estadística para rechazar la hipótesis nula a favor de la hipótesis alternativa, es decir, existe suficiente evidencia para afirmar que los residuales e_t son estacionarios.

Se concluye que la relación entre el IGAE y el primer componente principal no es espuria.

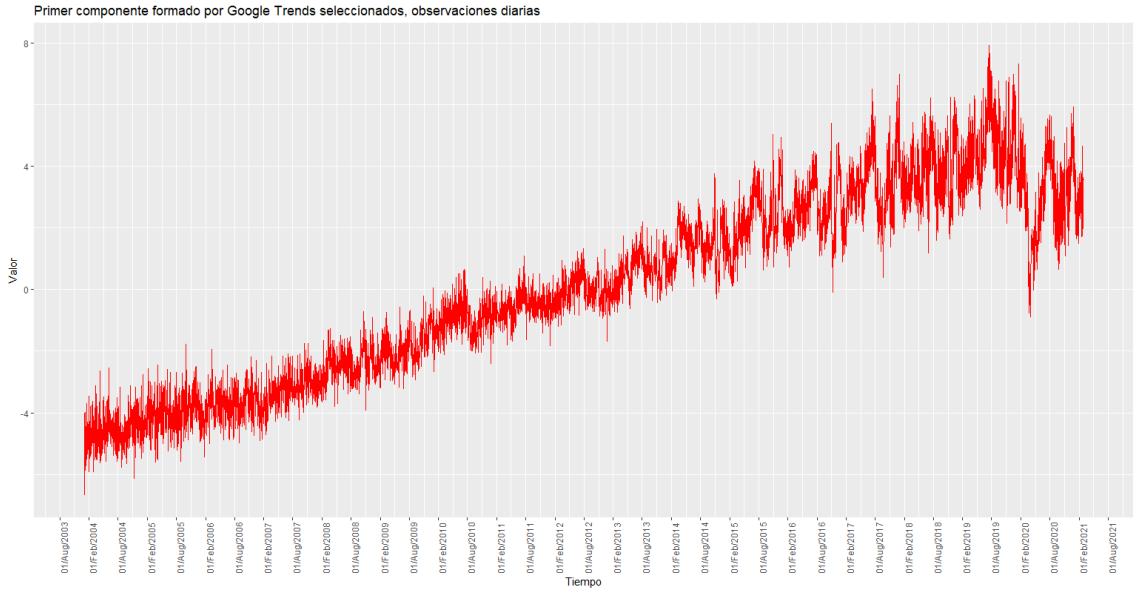

Figura 3.17: Primer componente de PCA con los nuevos GT seleccionados, observaciones diarias.

Luego se ajustó un modelo de PCA con los nuevos GT seleccionados con observaciones diarias para poder analizar su comportamiento, véase la Figura 3.16. En general el componente sigue un comportamiento similar al del IGAE, pues tiene una tendencia creciente pero tiene una caída importante en abril de 2020, aunque no refleja la caída económica de la crisis de 2008.

3.4. Ajuste del MFD

Se creó una base compuesta por 29 variables con observaciones diarias entre el 1/enero/2004 y el 28/febrero/2021, siendo 11 variables tradicionales y 18 variables no tradicionales (Tabla 3.4), a partir de la cual se ajustó el MFD.

ALFAA.MX	Avión	Beisbol	Buffete	Casa en venta
DJI	FEMSAUBD.MX	GFINBURO.MX	GFNORTEO.MX	IPC
KIMBERA.MX	Liverpool	Llantas	Maquillaje	Parque de diversiones
Playa	Restaurante	Ropa	Salon de belleza	S & P 500
Tenis	TELVISACPO.MX	Trafico	Traje	USD_ MXN
Vestido	Viajes	WALMEX.MX	Zapatos	

Tabla 3.4: Variables seleccionadas finales.

3.4.1. Número de factores ajustados.

Para determinar el número de factores óptimo para ajustar el modelo se utilizaron el Criterio de Bai y Ng, el Criterio de Onatski y el Criterio de Ahn y Horenstein.

Primero fue necesario calcular $r_{\text{máx}} = \lfloor 1.55\,\text{mín}\{T^{\frac{2}{5}}, N^{\frac{2}{5}}\}\rfloor$, y dado que $T = 6269$ y $N = 29$, se obtuvo que $r_{\text{máx}} = 5$. En la Tabla 3.5 se encuentran los $\hat{r}$ sugeridos por los diferentes criterios aplicados. Los criterios de información de Bai y Ng sugiere un $\hat{r} = 5$, pero serían demasiados factores para la adecuada interpretación de un indicador, además de que es el criterio menos robusto de los tres, por lo que se descarta. El procedimiento de Onatski sugiere un $\hat{r} = 2$, mientras que las razones de valores propios de Ahn y Horenstein sugiere un $\hat{r} = 1$, no obstante, la estimación de solamente un factor dinámico daría pie a una interpretación directa de ése factor como el que compila la información de todas las variables analizadas en un solo indicador económico, por lo que por cuestiones de interpretabilidad se elige $\hat{r} = 1$.

Criterio	**$\hat{r}$**
Bai y Ng	5
Onatski	2
Ahn y Horenstein	1

Tabla 3.5: Criterios para determinar el número de factores.

3.4.2. Estimación por el método PC

Dado que se eligió $\hat{r} = 1$, solamente se tiene que estimar un factor. Primero se realiza la estimación del factor dinámico único por el método PC, como una estimación preliminar, tal y como se puede ver en la Figura 3.18.

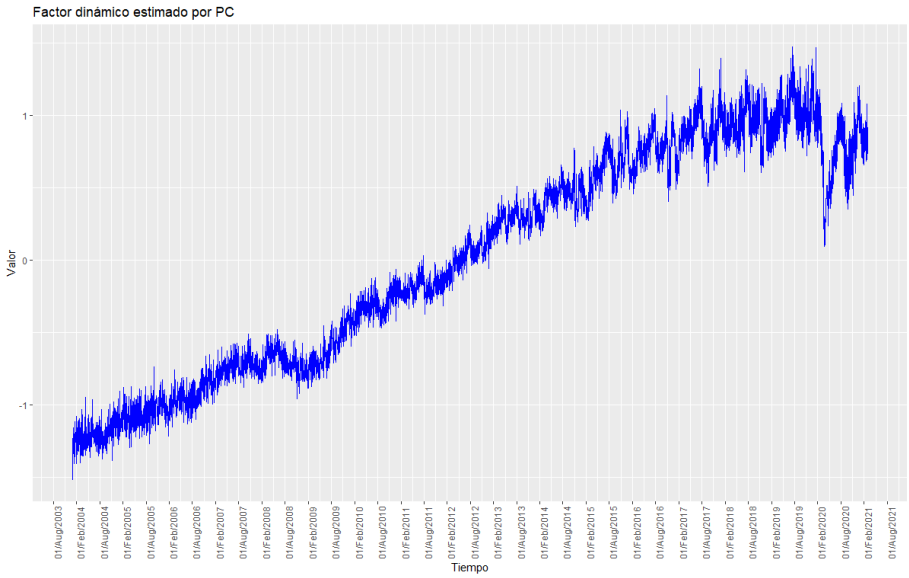

Figura 3.18: Factor dinámico estimado por el método PC.

Se observa que el factor tiene una tendencia creciente, pero refleja adecuadamente la caída de la economía mexicana durante la crisis del 2008 y durante la pandemia del COVID-19, para ir creciendo de manera oscilatoria en el los últimos meses.

3.4.3. Estimación por el método 2SKS

A partir del factor dinámico preliminar estimado por PC se estima un nuevo factor dinámico por el método 2SKS, como se puede notar en la Figura 3.19.

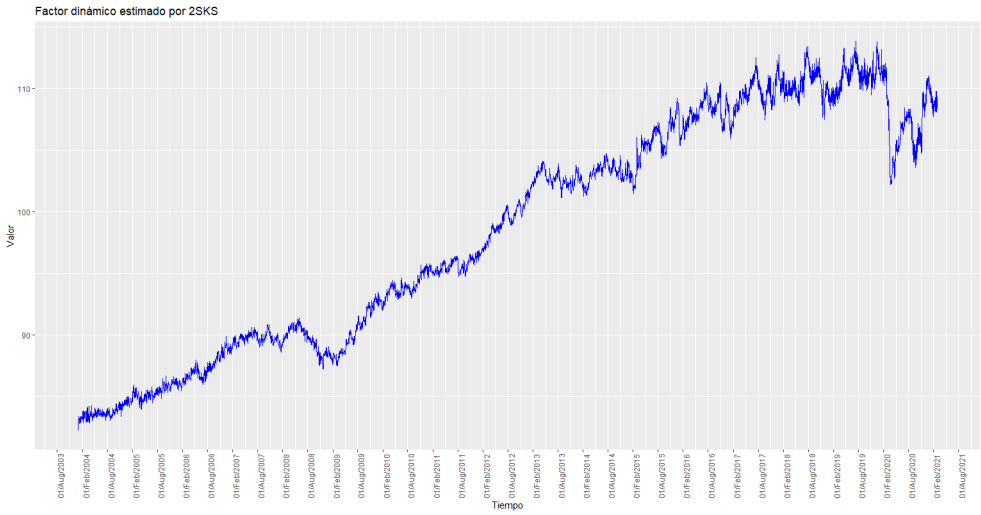

Figura 3.19: Factor dinámico estimado por el método 2SKS.

El factor estimado por 2SKS tiene una tendencia muy similar al factor estimado por PC, pero con una varianza mucho menor, consecuencia del suavizamiento.

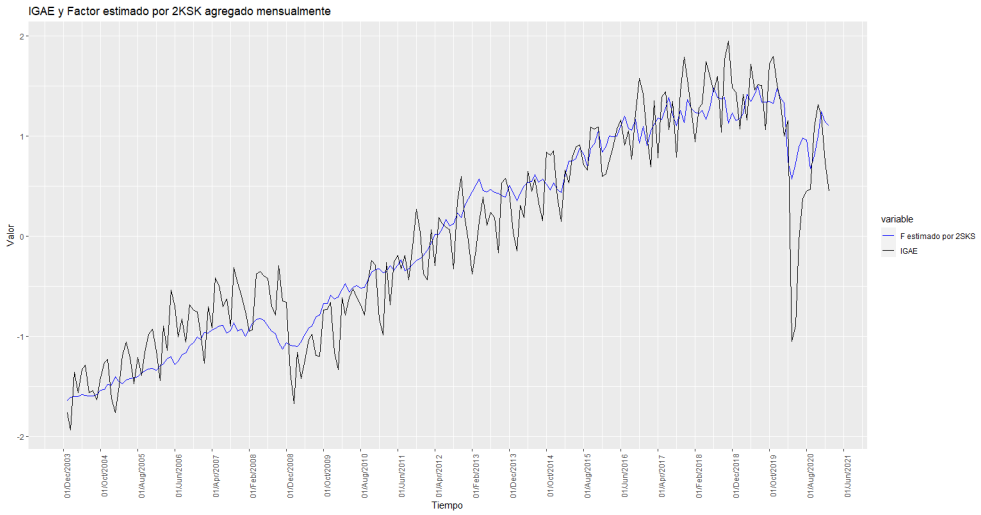

Figura 3.20: IGAE y el factor dinámico estimado por 2SKS agregado mensualmente.

En la Figura 3.20 se encuentra graficado en negro el IGAE y en azul el factor dinámico estimado por 2SKS agregado mensualmente. Nótese que el factor dinámico tiene un comportamiento muy similar al del IGAE, lo cual es un indicio de que refleja adecuadamente el comportamiento de la actividad económica en México.

3.4.4. Verificación de supuestos

Ya que solamente se estimó un factor dinámico ($\hat{r} = 1$) la prueba PANIC se reduce a aplicar una prueba ADF tradicional al factor dinámico. Se aplicaron primer puebas ADF con tendencia y sin tendencia al factor.

Hipótesis: H_0: El factor dinámico $\hat{F}_t$ tiene una raíz unitaria vs. H_1: El factor dinámico $\hat{F}_t$ es estacionario.

Nivel de significancia: $\alpha = 0.05$.

p-valores obtenidos: 0.2857 y 0.9784.

Resultado: A un nivel de significancia $\alpha = 0.05$ no existe suficiente evidencia estadística para rechazar la hipótesis nula a favor de la hipótesis alternativa, es decir, no existe suficiente evidencia para afirmar que el factor dinámico $\hat{F}_t$ es estacionario.

Se aplican entonces pruebas ADF con tendencia y sin tendencia al factor diferenciado $\Delta\hat{F}_t$.

Hipótesis: H_0: El factor dinámico $\Delta\hat{F}_t$ tiene una raíz unitaria vs. H_1: El factor dinámico $\Delta\hat{F}_t$ es estacionario.

Nivel de significancia: $\alpha = 0.05$.

p-valores obtenidos: 0.01y 0.01.

Resultado: A un nivel de significancia $\alpha = 0.05$ existe suficiente evidencia estadística para rechazar la hipótesis nula a favor de la hipótesis alternativa, es decir, existe suficiente evidencia para afirmar que el factor dinámico $\Delta\hat{F}_t$ es estacionario.

Del análisis anterior se concluye que el factor dinámico es integrado de orden uno, es decir, basta una diferencia para obtener un factor estacionario.

Luego, se aplica la prueba PANIC para errores idiosincráticos a los residuales del MDF.

Hipótesis: H_0: Los residuales del MFD tienen al menos una raíz unitaria vs. H_1: Los residuales del MFD son estacionarios.

Nivel de significancia: $\alpha = 0.05$.

p-valor obtenido: 0.00.

Resultado: A un nivel de significancia $\alpha = 0.05$ existe suficiente evidencia estadística para rechazar la hipótesis nula a favor de la hipótesis alternativa, es decir, existe suficiente evidencia para afirmar que los residuales del MFD son estacionarios.

Entonces, se concluye que se cumple el supuesto de que los residuales del MFD son estacionarios.

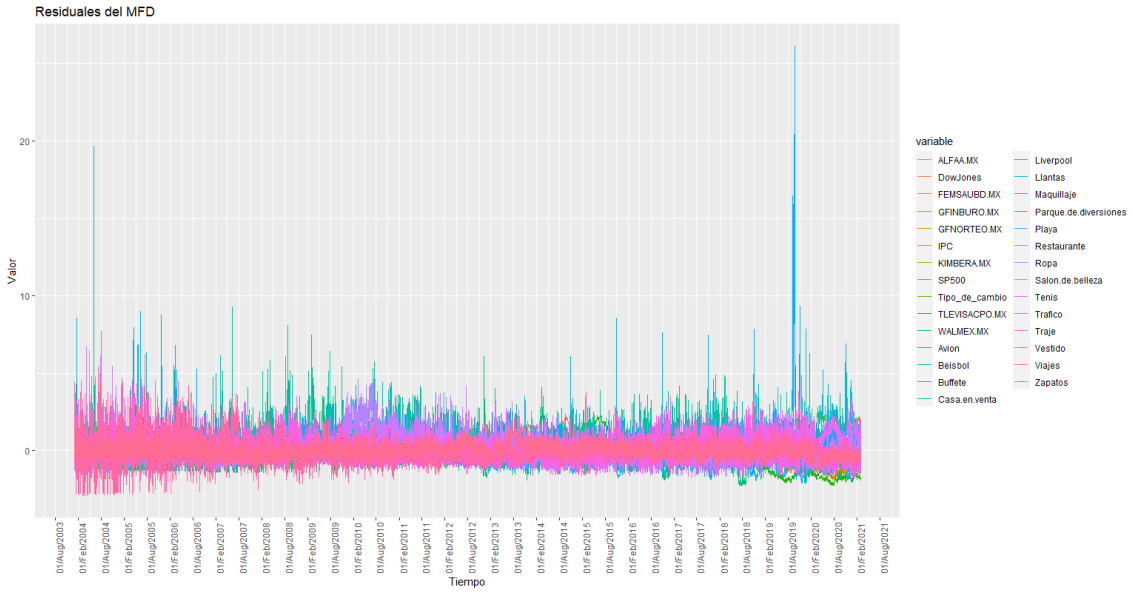

Figura 3.21: Residuales del MFD.

En la Figura 3.21 se observa que los residuales tienen una media alrededor del 0, y con una varianza constante, salvo ciertos picos que sobresalen, lo cual apoya la conclusión de que son estacionarios.

3.4.5. Regla de Combinación

Se aplica la RC partiendo del resultado de Guerrero y Nieto (1999):

$$\widehat{IDAE} = F + A(IGAE - CF),$$

siendo en éste caso F el factor dinámico estimado por el método 2SKS.

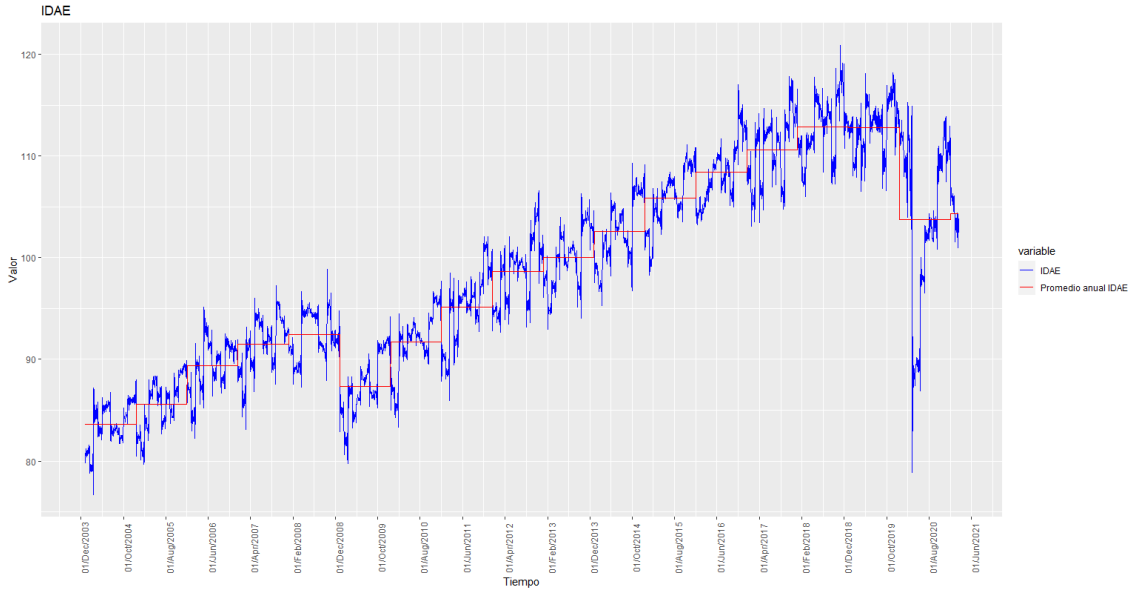

Figura 3.22: IDAE.

El indicador económico diario $\widehat{IDAE}$ se encuentra en color azul en la Figura 3.22, mientras que en rojo están sus promedios anuales. Como se puede observar, el indicador es prácticamente un análogo al IGAE pero de periodicidad diaria, lo cual significa que el $IDAE$ es un indicador que captura la información de la dinámica económica de México. Se tiene una media creciente alrededor del tiempo, pero con caídas durante la crisis económica del 2008 y la pandemia de COVID-19 de 2020, mientras que la varianza es constante.

Capítulo 4

Conclusiones

Se ajustó un MFD que permitió estimar al IDAE, un indicador diario de la actividad económica de nuestro país, lo cual es de suma importancia para la toma de decisiones, enfatizando la aplicación para el caso de México, pues el entorno económico de un país cambia rápidamente, y el hecho de tener un indicador económico de alta frecuencia permite identificar a mayor velocidad los cambios que puedan ser relevantes.

El modelo ajustado cumple con los supuestos necesarios para confirmar que es un modelo válido y bien fundamentado, por lo que existe evidencia estadística que da soporte a las conclusiones que se obtengan del IDAE.

Una técnica importante y provechosa para la selección de las variables fue el clústering de series de tiempo, pues permite identificar comportamientos o características de las series que de otra manera no hubiesen sido tan fáciles de observar.

Fue necesario adecuar los temas seleccionados para el conjunto de variables no tradicionales, lo cual es un indicador de la importancia que tiene el proponer un indicador fundamentado no solamente por la estadística, sino también por el sentido estructural y los aspectos empíricos que se observan en la realidad.

Este indicador puede ser utilizado en Estadística Oficial para realizar análisis de la coyuntura macroeconómica mexicana.

En trabajos futuros se podrían considerar diferentes variables, sería interesante, por ejemplo, agregar índices o reportes de movilidad. Otro trabajo futuro radica en utilizar el IDAE en modelos de nowcasting, de tal manera que se puedan hacer estimaciones de la actividad económica en tiempo real.

Referencias

Ahn, S. C., y Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, *81*(3), 1203–1227.

Aprigliano, V., Foroni, C., Marcellino, M., Mazzi, G., y Venditti, F. (2017). A daily indicator of economic growth for the euro area. *International Journal of Computational Economics and Econometrics*, *7*(1-2), 43–63.

Bai, J., y Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.

Bai, J., y Ng, S. (2004). A panic attack on unit roots and cointegration. *Econometrica*, *72*(4), 1127–1177.

Bai, J., y Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics*, *25*(1), 52–60.

Bai, J., y Ng, S. (2008). *Large dimensional factor analysis*. Now Publishers Inc.

Barigozzi, M., Lippi, M., y Luciani, M. (2016). Non-stationary dynamic factor models for large datasets. *Available at SSRN 2741739*.

Chamberlain, G., y Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, *51*(5), 1281–1304.

Corona, F., González-Farías, G., y López-Pérez, J. (2021). A nowcasting approach to generate timely estimates of mexican economic activity: An application to the period of covid-19. *arXiv preprint arXiv:2101.10383*.

Corona, F., Muriel, N., y González-Farías, G. (2021). Dynamic factor structure of team performances in liga mx. *Journal of Applied Statistics*, 1–13.

Corona, F., Poncela, P., y Ruiz, E. (2020). Estimating non-stationary common factors: implications for risk sharing. *Computational Economics*, *55*(1), 37–60.

Doz, C., Giannone, D., y Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics*, *164*(1), 188–205.

Eraslan, S., y Götz, T. (2021). An unconventional weekly economic activity index for germany. *Economics Letters*, *204*, 109881.

Giannone, D., Reichlin, L., y Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*(4), 665–676.

Guerrero, V. M., y Corona, F. (2018). Retropolating some relevant series of mexico's system of national accounts at constant prices: The case of mexico city's gdp. *Statistica Neerlandica*, *72*(4), 495–519.

Guerrero, V. M., y Nieto, F. H. (1999). Temporal and contemporaneous disaggregation of multiple economic time series. *Test*, *8*(2), 459–489.

Kaufman, L., y Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Lewis, D., Mertens, K., y Stock, J. H. (2020). *Us economic activity during the early weeks of the sars-cov-2 outbreak* (Inf. Téc.). National Bureau of Economic Research.

Liao, T. W. (2005). Clustering of time series data-a survey. *Pattern recognition*, *38*(11), 1857–1874.

Lourenço, N., y Rua, A. (2021). The daily economic indicator: tracking economic activity daily during the lockdown. *Economic Modelling*, *100*, 105500.

Maharaj, E. A., D'Urso, P., y Caiado, J. (2019). *Time series clustering and classification*. Chapman and Hall/CRC.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, *92*(4), 1004–1016.

Sakoe, H. (1971). Dynamic-programming approach to continuous speech recognition. En *1971 proc. the international congress of acoustics, budapest.*

Stock, J. H., y Watson, M. W. (2011). The oxford handbook of economic forecasting. En M. P. Clements y D. F. Hendry (Eds.), (cap. Dynamic Factor Models).

Oxford University Press.

Stock, J. H., y Watson, M. W. (2012). *Disentangling the channels of the 2007-2009 recession* (Inf. Téc.). National Bureau of Economic Research.

Apéndice A

Códigos de R

Selección de variables por medio de PLS

```
#cargando librerias
library(reshape2) #para porder usar la funcion melt
library(ggplot2) #para plotear
library(gridExtra) #para plotear en una misma ventana
library(scales) #para poder usar la funcion date_breaks

source("functions_agregaciones.R")
source("functions_PLS_bootstrap.R")
source("functions_PCA_TS.R")
source("functions_adf.R")
source("functions_dynamic_factors.R")

#cargando los datos
datos <- read.csv("variables.csv", header=TRUE, row.names = NULL)
attach(datos)
#escalando los datos excepto la primera columna por ser las fechas
datos[,-1] = scale(datos[,-1])
```

```
#formato adecuado a las fechas para poder graficarlo adecuadamente
datos$Date = as.Date(datos$Date)

#melt para poder graficar todas las series juntas
meltdf <- melt(datos,id="Date")

#grafico de todas las series juntas
ggplot(meltdf,aes(x=Date,y=value,colour=variable,group=variable))
  + geom_line() + #para graficar las lineas
  scale_x_date(date_labels = "%d/%b/%Y",
  breaks=date_breaks("6 month")) + #para graficar las fechas
  theme(axis.text.x = element_text(angle = 90)) +
  #para rotar las fechas 90 grados
  labs(title = "Variables tradicionales",
  x = "Tiempo", y = "Valor") #etiquetas

#IGAE
IGAE_base <- read.csv("IGAE_2004.csv", header=TRUE, row.names = NULL)
attach(IGAE_base)

#################################################################
#Agregados mensuales de las series
agregados_mensuales = monthly_aggregates(datos)

#formato adecuado a las fechas para poder graficarlo adecuadamente
agregados_mensuales$Date = as.Date(agregados_mensuales$Date)

#melt para poder graficar todas las series juntas
meltdf <- melt(agregados_mensuales,id="Date")
```

```
#grafico de todas las series juntas
ggplot(meltdf,aes(x=Date,y=value,colour=variable,group=variable)) +
  geom_line() + #para graficar las lineas
  scale_x_date(date_labels = "%d/%b/%Y",
  breaks=date_breaks("6 month")) + #para graficar las fechas
  theme(axis.text.x = element_text(angle = 90)) +
  #para rotar las fechas 90 grados
  labs(title = "Variables tradicionales agregadas mensualmente",
  x = "Tiempo", y = "Valor") #etiquetas

##########################################################
#Ajustes PLS

PLS_99 = IC_bootstrap_PLS(X=agregados_mensuales,
          y=IGAE, S=1000, alpha = 0.01)
PLS_99$Variables_sig
PLS_99$Variables_NO_sig
PLS_99$Loadings_finales
PLS_99$Datos_Sig

GT_sig =  PLS_99$Datos_Sig

#formato adecuado a las fechas para poder graficarlo adecuadamente
GT_sig$fecha = as.Date(GT_sig$fecha)

#melt para poder graficar todas las series juntas
meltdf <- melt(GT_sig,id='fecha')

#grafico de todas las series juntas
```

```
ggplot(meltdf,aes(x=fecha,y=value,colour=variable,group=variable))
  + geom_line() + #para graficar las lineas
  scale_x_date(date_labels = "%d/%b/%Y",
  breaks=date_breaks("6 month")) + #para graficar las fechas
  theme(axis.text.x = element_text(angle = 90)) +
  #para rotar las fechas 90 grados
  labs(title = "Variables tradicionales seleccionadas",
  x = "Tiempo", y = "Valor") #etiquetas

##################################################################
#PCA Mensual

PCA_mensual = PCA_ts(X=agregados_mensuales[,PLS_99$Variables_sig],
                y=IGAE, fecha=agregados_mensuales$Date,
                col_comp = "blue",
                main = "IGAE vs. Primer componente")
PCA_mensual$IC_corr
PCA_mensual[[3]]

#Prueba ADF a los errores de la regresion entre el IGAE y comp1
reg = lm(IGAE ~ PCA_mensual$Componente1)
#se rechaza H0, sin raiz unitaria, regresion no espuria
adf(reg$residuals, "trend")$p.value
#se rechaza H0, sin raiz unitaria, regresion no espuria
adf(reg$residuals, "none")$p.value

PCA_diario = PCA_diario(X=datos[,PLS_99$Variables_sig], fecha=datos$Date,
              color_plot ="blue",
              main = "Primer componente")
PCA_diario[[2]]
```

Ajuste del MFD

```
#cargando librerias
library(reshape2) #para porder usar la funcion melt
library(ggplot2) #para plotear
library(gridExtra) #para plotear en una misma ventana
library(scales) #para poder usar la funcion date_breaks
library(pls)
library(dlm)
library(forecast)
library(FitARMA)
library(zoo)
library(tidyr)
library(MASS)

#cargando funciones
source("functions_adf.R")
source("functions_dynamic_factors.R")
source("Bai-Ng_Ahn_Horenstein.R")
source("TWO_STEPS.R")

#cargando los datos
datos <- read.csv("Base_final.csv", header=TRUE, row.names = NULL)
attach(datos)
#escalando los datos excepto la primera columna por ser las fechas
datos[,-1] = scale(datos[,-1])

#formato adecuado a las fechas para poder graficar
datos$Date = as.Date(datos$Date)
```

```
#melt para poder graficar todas las series juntas
meltdf <- melt(datos,id="Date")

#grafico de todas las series juntas
ggplot(meltdf,aes(x=Date,y=value,colour=variable,group=variable))
    + geom_line() + #para graficar las lineas
  scale_x_date(date_labels = "%d/%b/%Y",
  breaks=date_breaks("10 month")) +
  #para graficar las fechas
  theme(axis.text.x = element_text(angle = 90)) +
  #para rotar las fechas 90 grados
  labs(title = "Google Trends, Variables tradicionales y stocks",
  x = "Tiempo", y = "Valor") #etiquetas

#IGAE
IGAE_base <- read.csv("IGAE_2004.csv", header=TRUE, row.names = NULL)
attach(IGAE_base)

##########################################################
#Criterios para la seleccion de r

#matriz con los datos estandarizados y sin fechas
X <- as.matrix(datos[,-1])
#X <- as.matrix(datos[2:26])
N = ncol(X)
T = nrow(X)
rmax = floor(1.55*min(T^(2/5),N^(2/5)))
rmax

#descomposicion en valores propios
```

```
XtX <- t(X) %*% X
ed <- eigen(XtX)

# criterio tradicional
cumsum(ed$values/sum(ed$values))

#criterio de Bai y Ng
bai_ng = c()
for(i in 1:rmax){
  mod_aux <- pcfest(X, i, demean = 0, constant = 1)
  bai_ng = cbind(bai_ng,mod_aux$ICPk)
}
bai_ng #se sugiere 1 factor

#criterio de onatski
rhat <- onatski2010(X, demean = 0)["ed"]
rhat #se sugieren 2 factores

#criterio de AHn y Horenstein
ahn_hor = ratio.test(X, kmax = rmax, demean = 0)
ahn_hor #se sugieren 1 factores

#numero de componentes principales elegido
p=1

##########################################################################
#Estimacion por PC

#estimacion de la matriz de cargas  por componentes principales
mod1 <- pcfest(X, p, demean = 0, constant = 1)
```

```
#matriz de cargas estimada
F_estimado = as.data.frame(-mod1$Fhat)

#uniendo la matriz Fhat estimada con las fechas
comp_princ = cbind(as.Date(datos$Date),F_estimado)
#cambiando el nombre de la columna x por fecha
colnames(comp_princ)[1] = "fecha"

#plot del componente principal 1
comp1 = ggplot(comp_princ,aes(x=fecha,y=f1)) +
  geom_line(col="blue") + #para graficar las lineas
  scale_x_date(date_labels = "%d/%b/%Y",
  reaks=date_breaks("6 month")) + #para graficar las fechas
  theme(axis.text.x = element_text(angle = 90)) +
  #para rotar las fechas 90 grados
  labs(title = "Factor dinamico estimado por PC",
  x = "Tiempo", y = "Valor") #etiquetas
comp1

#######################################################################
#Estimacion por 2SKS

est_2sks = two_step(X)
fhat_s = as.matrix(-est_2sks$fs_hat[,2])
rhat = 1
X_s <- X
ed <- eigen(t(X_s)%*%X_s)
N <- ncol(X_s)
```

```
##########################################################################
#errores idiosincraticos 2sks

#ehat_2sks <- X_s - fhat_s %*% t(Phat)
#hipotesis alternativa residuales son estacionarios
ehat_2sks = est_2sks$ehat
pooled.test(ehat_2sks)

#uniendo la matriz et con las fechas
residuales= data.frame(datos$Date,ehat_2sks)
#cambiando el nombre de la columna x por fecha
colnames(residuales)[1] = "fecha"
#melt para poder graficar todos los residuales juntos
melt_residuales <- melt(residuales,id="fecha")

#grafico de los residuales
ggplot(melt_residuales,aes(x=fecha,y=value,colour=variable,
  group=variable)) + geom_line() + #para graficar las lineas
  scale_x_date(date_labels = "%d/%b/%Y",
  breaks=date_breaks("6 month")) + #para graficar las fechas
  theme(axis.text.x = element_text(angle = 90)) +
  #para rotar las fechas 90 grados
  labs(title = "Residuales del MFD",
  x = "Tiempo", y = "Valor") #etiquetas

##########################################################################
#Grafica del factor estimado por 2SKS

#factor agregado mensual
df_fhat_s = data.frame(datos$Date,fhat_s)
```

```
factor_mensual = monthly_aggregates(df_fhat_s)
factor_mensual = factor_mensual[,-1]

#para normalizar
Tt = length(IGAE)
fhat_s_m <-  monthly_aggregates(df_fhat_s)
regre <- lm(IGAE ~ fhat_s_m$fhat_s)
fhat_s_m$fhat_s <- fitted(regre)

# indicador normalizado
fhat_s <- coef(regre)[1] + coef(regre)[2]*fhat_s

#grafico del factor estimado por 2SKS
factor_ks = data.frame(datos$Date,fhat_s)
colnames(factor_ks) = c("Date","fhat_s")
ggplot(factor_ks,aes(x=Date,y=fhat_s)) +
  geom_line(col="blue") + #para graficar las lineas
  scale_x_date(date_labels = "%d/%b/%Y",
  breaks=date_breaks("6 month")) + #para graficar las fechas
  theme(axis.text.x = element_text(angle = 90)) +
  #para rotar las fechas 90 grados
  labs(title = "Factor dinamico estimado por 2SKS",
  x = "Tiempo", y = "Valor") #etiquetas

#######################################################################
#Grafica del IGAE y el factor estimado por 2SKS

IGAE_y_2SKS= data.frame(IGAE_base$Date,factor_mensual,IGAE)
colnames(IGAE_y_2SKS) = c("Date","F estimado por 2SKS","IGAE")
IGAE_y_2SKS[,-1] = scale(IGAE_y_2SKS[,-1])
```

```
IGAE_y_2SKS$Date = as.Date(IGAE_y_2SKS$Date)

#melt para poder graficar todas las series juntas
melt_IGAE_2SKS <- melt(IGAE_y_2SKS, id="Date")

#grafico del igae y el factor 2sks agregado mensual
ggplot(melt_IGAE_2SKS, aes(x=Date, y=value, colour=variable,
  group=variable)) + geom_line(aes(color=variable)) +
  #para graficar las lineas
  scale_color_manual(values=c("blue","black")) +
  scale_x_date(date_labels = "%d/%b/%Y",
  breaks=date_breaks("10 month")) + #para graficar las fechas
  theme(axis.text.x = element_text(angle = 90)) +
  #para rotar las fechas 90 grados
  labs(title = "IGAE y Factor estimado
  por 2KSK agregado mensualmente",
  x = "Tiempo", y = "Valor") #etiquetas

#################################################################
#Pruebas adf al factor

adf(fhat_s, "trend")$p.value
adf(fhat_s, "none")$p.value
adf(diff(fhat_s), "trend")$p.value
adf(diff(fhat_s), "none")$p.value

#################################################################
#Regla de combinacion
```

```
igae = IGAE
igae = as.matrix(igae)
rownames(igae) = IGAE_base$Date
fhat_s = data.frame(fhat_s)
rownames(fhat_s) = datos$Date

C <- matrix(0, length(igae), nrow(fhat_s))
rownames(C) <- rownames(igae)

for(i in 1 : nrow(C)){
  idx <- substring(rownames(fhat_s), 3, 4) ==
    substring(rownames(igae)[i], 3, 4) &
    substring(rownames(fhat_s), 6, 7) ==
    substring(rownames(igae)[i], 6, 7)
  C[i,idx] <- 1/sum(idx)
}

# arima
arimam <- auto.arima(fhat_s)

# ma coefficients
arc <- coef(arimam)[paste("ar",1:arimam$arma[1], sep = "")]
mac <- coef(arimam)[paste("ma",1:arimam$arma[2], sep = "")]

s=20
# ma coefficients
psi <- ImpulseCoefficientsARMA(arc, mac, s)
psi[is.na(psi)] <- 0

# matrix Psi
```

```
Psi <- diag(ncol(C))

# generating
for(j in 1 : s){
  for(i in 1 : (nrow(Psi)-j))
    Psi[i+j,i] <- psi[j]
}

# A matrix
A <- Psi %*%t(Psi) %*%t(C) %*%ginv(C%*%Psi %*%t(Psi) %*%t(C))

# dissagregation
Zd <- fhat_s + A%*%(as.matrix(igae - C%*%as.matrix(fhat_s)))
ave <- matrix(0, nrow(fhat_s), 1)
for(i in 1 : length(2004:2021)){
  idx <- unique(substring(rownames(igae),3, 4))[i] ==
                substring(rownames(fhat_s), 3, 4)
  ave[idx,] <- mean(igae[unique(substring(rownames(igae), 3, 4))[i] ==
                substring(rownames(igae), 3, 4),])
}

#Df de Zd agregado mensual
Zd_monthly = data.frame(datos$Date, Zd)
Zd_monthly = monthly_aggregates(Zd_monthly)
colnames(Zd_monthly) = c("Date","Promedio mensual Y")

#se completan las fechas faltantes entre la fecha minima
#y la fecha maxima
df_Zd = complete(Zd_monthly,Date = seq.Date(min(as.Date(Zd_monthly$Date)),
as.Date("2021-02-28"), by="day"))
```

```
#se rellenan los NA con el inmediato anterior
df_Zd = fill(df_Zd,colnames(df_Zd)[-1])

#se agrega el factor diario y los promedios anuales
df_Zd = data.frame(datos$Date, Zd, ave)
colnames(df_Zd) = c("Date", "IDAE", "Promedio anual IDAE")

#melt para poder graficar el indicador economico
melt_Zd<- melt(df_Zd,id="Date")

#grafico del indicador economico
ggplot(melt_Zd,aes(x=Date,y=value,colour=variable,group=variable))
  + geom_line(aes(color=variable)) +
  #para graficar las lineas
  scale_color_manual(values=c("blue","red")) +
  scale_x_date(date_labels = "%d/%b/%Y",
  breaks=date_breaks("10 month")) +
  #para graficar las fechas
  theme(axis.text.x = element_text(angle = 90)) +
  #para rotar las fechas 90 grados
  labs(title = "IDAE",
  x = "Tiempo", y = "Valor") #etiquetas
```

Función de agregaciones mensuales

```
#Funcion que calcula las agregaciones mensuales
#de un conjunto X de series de tiempo
#argumentos:
#   X conjunto de series de tiempo con la PRIMERA COLUMNA DE FECHAS
```

```
#cargando librerias
library(xts) #para poder hacer la agregacion

monthly_aggregates = function(X){

  T = dim(X)[1] #Numero de observaciones T
  N = dim(X)[2] #Numero de series N, considerando a las fechas

  #formato adecuado a las fechas para poder graficarlo adecuadamente
  X[,1] = as.Date(X[,1])

  #fechas entre el primer y el ultimo dia en formato mensual
  fechas_mensuales = seq.Date(min(X[,1]),max(X[,1]),by = "months")

  #numero de meses
  T_mensual = length(fechas_mensuales)

  #matriz en la que se guardaran las series agregadas
  agregados <- matrix(0,nrow=T_mensual, ncol=N)

  #ciclo que realiza la agregacion
  for(i in 2:N){
  #formato xts para poder hacer la agregacion
    ts_aux = xts(x=X[,i], order.by = X[,1])
    #se separa por meses
    ts_aux_month <- split(ts_aux, f = "months")
    #se calcula la media por mes
    agregados[,i] <- sapply(X = ts_aux_month, FUN = mean)
  }
```

```
  #nombres a las columnas
  colnames(agregados) = colnames(X)
  #convirtiendo a dataframe para poder agregar las fechas
  agregados = as.data.frame(agregados)
  #agregando fechas
  agregados[,1] = fechas_mensuales

  #se regresan los agregados mensuales
  return(agregados)
}
```

Función de ajuste PLS con bootstrap

```
#Funcion que calcula IC bootstrap al primer
#componente de una regresion PLS
#argumentos:
#   X = conjunto de series de tiempo, con la primera
#       columna igual a las fechas
#   y = variable respuesta
#   S = numero de muestras bootstrap
#   alpha = nivel de significancia

#se cargan librerias
library(pls)
library(tibble)

IC_bootstrap_PLS = function(X, y, S=10000, alpha=0.05){

  #semilla
```

```
set.seed(10)

fecha = X[,1]

#datos X en formato matriz sin la primera columna
X <- data.matrix(X[,-1])

T = dim(X)[1] #Numero de observaciones T
N = dim(X)[2] #Numero de series N

# matriz donde guardamos los datos
Phat_sample <- matrix(0, N, S)

# remuestramos con muestras dependientes
for(i in 1 : S){ print(i)
  #genera las muestras
  samples <- sample(T/4)[1]:sample((T/2):T)[1]
  #ajusta PLS a las muestras
  pls_regre_sample <- plsr(y[samples] ~ X[samples,])
  #guarda los loadings del primer componente
  Phat_sample[,i] <- loadings(pls_regre_sample)[,1]
}

#calcular IC
Phat_conf_2 <- round(cbind(
  #Limite inferior
  apply(Phat_sample, 1, function(x) quantile(x, alpha)),
  #Valor medio
  apply(Phat_sample, 1, function(x) quantile(x, 0.5)),
  #limite superior
```

```
  apply(Phat_sample, 1, function(x) quantile(x, 1-alpha))), 4)

#vector vacio para guardar los IC que no contienen al 0
loadings_significativos = c()
#ciclo que encuentra los loadings significativos,
#es decir, cuyo intervalo no contiene al 0
for(i in 1:N){
  if(unname((Phat_conf_2[i,1] < 0 && Phat_conf_2[i,3] < 0)
  || (Phat_conf_2[i,1] > 0 && Phat_conf_2[i,3] > 0)) ){
    #guarda el i-esimo loading significativo
    loadings_significativos = c(loadings_significativos,i)
  }
}

#datos seleccionados
df_significativos = X[,loadings_significativos]
#nombres de las variables significativas
var_sig = colnames(df_significativos)
#agregando la columna fechas
df_significativos = data.frame(fecha, df_significativos)

loadings_finales = Phat_conf_2[loadings_significativos,]
row.names(loadings_finales) = var_sig
colnames(loadings_finales) = c("q_alpha","q_0.5","q_(1-alpha)")

#datos NO seleccionadas
df_no_sig = X[,-loadings_significativos]
#nombres de las variables NO significativas
var_no_sig = colnames(df_no_sig)
#agregando la columna fechas
```

```
  df_no_sig = data.frame(fecha, df_no_sig)

  #se guardan los resultados calculados en una lista
  resultado = list(Phat_conf_2,loadings_significativos,
  loadings_finales, df_significativos, var_sig, df_no_sig, var_no_sig)
  names(resultado) = c("IC","Index_sig", "Loadings_finales",
  "Datos_Sig","Variables_sig","Datos_NO_sig","Variables_NO_sig")
  #se regresan los resultados
  return(resultado)
}
```

PCA para series de tiempo

```
#Funcion que ajusta PCA a un conjunto de series de tiempo
#y calcula la correlacion con una v. respuesta y
#argumentos:
#   X = conjunto de series de tiempo
#   y = variable respuesta
#   fecha = fechas para graficar
#   main = titulo

PCA_ts = function(X, y, fecha, main="",col_comp="blue"){

  #ajuste del PCA
  pca <- prcomp(X, center = TRUE,scale. = TRUE)
  #se guarda el primer componente principal
  comp1 = pca$x[,1]
  #correlacion entre el primer componente principal
  #y la variable respuesta y
```

```
IC_corr = cor.test(comp1,y)

#se guardan el primer componente y la variable respuesta
datos_pca = cbind(comp1,y)
#se escalan los datos
datos_pca = scale(datos_pca)
#se agregan las fechas a la matriz
datos_pca = cbind(fecha,datos_pca)
#se convierte a formato data frame para poder graficar
datos_pca= as.data.frame(datos_pca)
#se agregan nombres a las columnas
colnames(datos_pca) = c("fecha","Componente 1", "IGAE")
#se agrega el formato correcto
datos_pca$fecha = as.Date(datos_pca$fecha)
#melt para poder graficar todas las series juntas
melt_pca <- melt(datos_pca ,id="fecha")
#grafico de todas las series juntas
plot_PCA = ggplot(melt_pca ,aes(x=fecha,y=value,
  colour=variable,group=variable)) +
  geom_line(aes(color=variable)) + #para graficar las lineas
  scale_color_manual(values=c(col_comp,"black")) +
  scale_x_date(date_labels = "%d/%b/%Y",
  breaks=date_breaks("6 month")) + #para graficar las fechas
  #para rotar las fechas 90 grados
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = main, x = "Tiempo", y = "Valor") #etiquetas

#se regresan el primer componente, el IC de la correlacion y el plot
resultado = list(comp1,IC_corr, plot_PCA)
names(resultado) = c("Componente1","IC_corr","Plot")
```

```
  return(resultado)
}

#Funcion que ajusta PCA a un conjunto de series de tiempo
#y plotea el primer componente
#argumentos:
#   X = conjunto de series de tiempo
#   fecha = fechas para graficar
#   main = Titulo

PCA_diario = function(X, fecha, main="", color_plot){

  #ajuste del PCA
  pca <- prcomp(X, center = TRUE, scale. = TRUE)
  #se guarda el primer componente principal
  comp1 = pca$x[,1]
  #se agregan las fechas a la matriz
  datos_pca = cbind(fecha,comp1)
  #se convierte a formato data frame para poder graficar
  datos_pca= as.data.frame(datos_pca)
  #se agregan nombres a las columnas
  colnames(datos_pca) = c("fecha","Componente 1")
  #se agrega el formato correcto
  datos_pca$fecha = as.Date(datos_pca$fecha)

  plot_PCA = ggplot(datos_pca,aes(x=fecha,y=comp1)) +
    geom_line(color=color_plot) + #para graficar las lineas
    scale_x_date(date_labels = "%d/%b/%Y",
    breaks=date_breaks("6 month")) + #para graficar las fechas
    #para rotar las fechas 90 grados
```

```
    theme(axis.text.x = element_text(angle = 90)) +
    labs(title = main, x = "Tiempo", y = "Valor") #etiquetas

  #se regresan el primer componente, el IC de la correlacion y el plot
  resultado = list(comp1,plot_PCA)
  names(resultado) = c("Componente1","Plot")
  return(resultado)
}
```

Apéndice B

Códigos de Python

Clústering de series de tiempo

```
#se cargan los datos como un dataframe de pandas
import pandas as pd
import numpy as np
df = pd.read_csv("VTyS_99.csv",index_col = ['Date'])
df = df.T
df

#nombre de las columnas
nombres = list(df.T.columns)

#se convierte el df a una lista de listas por
#el formato que requiere tslearn
from tslearn.utils import to_time_series_dataset
list_ts = df.values.tolist()
formatted_ts = to_time_series_dataset(list_ts)
print(formatted_ts.shape)
```

```
#se cargan los modulos necesarios
from tslearn.clustering import TimeSeriesKMeans
from tslearn.preprocessing import TimeSeriesScalerMeanVariance
import matplotlib.pyplot as plt
import numpy as np
from tslearn.clustering import GlobalAlignmentKernelKMeans
from tslearn.metrics import sigma_gak
import random
import matplotlib.dates as mdates

#Estandarizacion de las series
series = TimeSeriesScalerMeanVariance().fit_transform(formatted_ts)

#diferentes tipos de lineas para los plots
typeline = [ '-', '--', '-.', ':' ]

#implementacion en HTML para imprimir varias tablas juntas
from IPython.core.display import HTML
def multi_table(table_list):
    ''' Acceps a list of IpyTable objects and returns
    a table which contains each IpyTable in a cell
    '''
    return HTML(
        '<table><tr style="background-color:white;">' +
        ''.join(['<td>' + table._repr_html_() +
        '</td>' for table in table_list]) +
        '</tr></table>'
    )

#Soft DTW k-means
```

```
k=3  #numero de clusters
seed=10 #semilla

#se imprime en la salida el metodo aplicado
print("Soft DTW k-means")

#parametros
soft_km = TimeSeriesKMeans(n_clusters=k,
                           metric="softdtw",
                           metric_params={"gamma": .01},
                           n_jobs=-1,
                           random_state=seed)

#asuste de DBA k-means
y_pred_soft_km = soft_km.fit_predict(series)

#se guardan las etiquetas de los clusters
soft_km_etiquetas = soft_km.labels_
#se guarda la informacion de las claves y las etiquetas en un dataframe
soft_km_dic = {'Clave': nombres, 'Cluster': soft_km_etiquetas+1}
soft_km_tabla = pd.DataFrame(data = soft_km_dic)

#se cambia la longitud de la figura
fig = plt.figure(figsize=(20, 20))
#ciclo que grafica cada cluster
for k_i in range(k):
    #se grafica cada cluster como parte de un subplot
    ax = plt.subplot(k,1, k_i + 1)
    #df auxiliar con las etiquetas de cada serie
    df_aux_soft_km = soft_km_tabla[soft_km_tabla.Cluster == k_i+1]
```

```
    #grafico de las series del i-esimo cluster
    for x_i in series[y_pred_soft_km == k_i]:
        plt.plot(x_i.ravel(), alpha=1,
        linestyle=random.choice(typeline))
    #grafico de los centroides de los clusters
    plt.plot(soft_km.cluster_centers_[k_i].ravel(), "r-")
    #se agregan los nombres de cada cluster
    plt.text(0.55, 0.85,'Cluster %d' %(k_i + 1),
    transform=plt.gca().transAxes)
    #agregando legends a cada cluster
    plt.legend(list(df_aux_soft_km.iloc[:,0]), loc='lower right' )

#Resultado final Soft DTW k-means
print("Soft DTW k-means clusters")
multi_table([soft_km_tabla[soft_km_tabla.Cluster == 1],
            soft_km_tabla[soft_km_tabla.Cluster == 2],
            soft_km_tabla[soft_km_tabla.Cluster == 3]])
```