Funcionamiento en muestras finitas de técnicas de imputación y retropolación:

caso de las series de encuestas económicas nacionales del INEGI

Finite Sample Performance of Imputation and Retropolation Techniques:

the INEGI's National Economic Surveys' case

Francisco de Jesús Corona Villavicencio,* Jesús López-Pérez y Nelson Omar Muriel Torrero*****

Nota: se agradecen los comentarios y sugerencias realizadas por el personal de la Dirección General Adjunta de Encuestas Económicas del INEGI, en particular a Araceli Martínez Gama, Santiago Ávila Ávila, Juan José Ríos Franco, Francisco Reyes Piña, Ramón Bravo Cepeda, Ramón Sánchez Trujano, Roberto Tovar Soria, Diana Gabriela Cedeño Robles, Rodrigo G. Carranza Trinidad y Zelic Marco Antonio Rosa Vara; también, nuestro agradecimiento a Gerardo Leyva Parra, director general adjunto de Investigación por sus valiosas aportaciones realizadas.

* Instituto Nacional de Estadística y Geografía (INEGI), franciscoj.corona@inegi.org.mx.
** INEGI, jesus.lopezp@inegi.org.mx
*** Universidad Iberoamericana CDMX, nelson.muriel@ibero.mx

Snow and Rosss Geese, Bosque Del Apache, New Mexico, USA, Winter/Education Images/Getty Images

El objetivo de este trabajo es analizar el funcionamiento en muestras finitas de diferentes técnicas de imputación y retropolación en el contexto de series de tiempo. Para estos fines, se realiza un experimento Monte Carlo simulando diversas series de tiempo a través de modelos de factores dinámicos estacionales, considerando factores estacionarios, no estacionarios y la combinación de ambos; lo anterior, bajo distintas especificaciones en el componente idiosincrático. Este proceso generador de datos es validado ajustando dichos modelos a cinco bases de datos de las encuestas económicas nacionales del INEGI. Los principales resultados indican que, para imputar datos, es conveniente usar información adicional; de otra forma, los métodos basados en el filtro de Kalman son una buena alternativa. En retropolación, los mejores resultados se obtienen incluyendo restricciones sobre el pasado de la serie de tiempo a retropolar; caso contrario, empalmar series de tiempo ofrece una solución útil en la práctica.

Palabras clave: cointegración; encuestas económicas nacionales; modelos de factores dinámicos; experimento Monte Carlo; raíz del error cuadrático medio.

Recibido: 4 de marzo de 2019.
Aceptado: 9 de mayo de 2019.

The purpose of this research is to analyze the finite sample performance of frequently used imputation and retropolation time series techniques. In this way, we carry out a Monte Carlo experiment simulating several time series through seasonal Dynamic Factor Models (DFMs) considering stationary factors, non-stationary common factors and the combination of both, along with different specifications in the idiosyncratic component. This data-generating process is validated through DFMs estimation on five database of National Economic Surveys (EEN) from INEGI. The main results indicate that for data imputation, using information of correlated time series helps to minimize the estimation error; on the other hand, the methods based on the Kalman filter are a useful alternative. For retropolation, the better results are obtained using restrictions over the past of the time series; otherwise, the splice method is a useful alternative in practice.

Key words: Cointegration; National Economic Surveys; Dynamic Factor Models; Monte Carlo Experiment; Root mean squared error.

1. Introducción

Las instituciones encargadas de generar series de tiempo económicas de carácter oficial se enfrentan a diversos retos metodológicos para construir las que reflejen un concepto económico, ofrecer al público las que cubran periodos mayores, incorporar nuevos aspectos metodológicos ante los cambios de año base, etcétera. Todos estos ejemplos requieren el uso de técnicas estadísticas o econométricas que deben tener sustento teórico y empírico para garantizar la calidad de la información oficial publicada.

Centrándonos en el problema de generar indicadores económicos con temporalidad larga, observamos que, con frecuencia, estos dependen de series de tiempo que, por diversas causas (problemas técnicos ocurridos al momento de realizar la recolección de la información, falta de recursos económicos para observarla, inexistencia de la misma, etc.), tienen datos faltantes. Por otro lado, para construir indicadores económicos que tengan las características de pertinencia y utilidad al público en general, es necesario, en ocasiones, recurrir a información de series de tiempo complementarias para poder construir otras más largas. En estos casos, es necesario implementar técnicas ya sea de imputación de datos, de retropolación de series de tiempo o ambas para construir indicadores adecuados.

La generación de los indicadores económicos disponibles en el Banco de Información Económica (BIE) del Instituto Nacional de Estadística y Geografía (INEGI) depende de los datos disponibles en series de tiempo de estudios de dominios menores. En ellas, por razones múltiples, suele haber datos faltantes e incompatibilidades, producto —estas últimas— de los cambios metodológicos a

los que su registro se ha sujetado. En particular, las encuestas económicas nacionales (EEN) —entre ellas, la Mensual sobre Empresas Comerciales (EMEC), la Mensual de la Industria Manufacturera (EMIM), la Mensual de Servicios (EMS), la Nacional de Empresas Constructoras (ENEC) y la Mensual de Opinión Empresarial (EMOE)— presentan estas características, lo cual hace que su integración en los índices del BIE sea un reto y que la búsqueda de indicadores económicos con temporalidades más largas se beneficie de las técnicas de retropolación e imputación.

Existen algunos trabajos previos que proponen y/o evalúan el funcionamiento de diversas técnicas de imputación en diferentes áreas del conocimiento. En cuanto a propuestas metodológicas tenemos, por ejemplo, a Harvey & Pierse (1984), quienes recomiendan un método basado en la representación espacio-estado del Modelo Autorregresivo Integrado de Media Móvil (ARIMA, por su nombre en inglés) y la aplicación del filtro de Kalman. Peña & Tiao (1989), por otro lado, plantean que se puede hacer uso de series de tiempo complementarias para imputar la información faltante utilizando esperanzas condicionales o promedios como estimadores. En el ámbito de series de tiempo múltiples, Guerrero & Gaspar (2010) proponen una metodología que considera la edición e imputación de la serie de forma conjunta basándose en procesos de la familia de vectores autorregresivos (VAR). En cuanto a la evaluación de técnicas de imputación, destacan los trabajos de Pfeffermann & Nathan (2001) concentrados en encuestas; Schafer & Graham (2002), quienes presentan una evolución histórica de las técnicas de imputación de datos faltantes en cortes transversales; Moritz *et al*. (2015), donde el problema se aborda desde el ámbito computacional; Schmitt *et al*. (2015), concentrados en aplicaciones en las ciencias naturales y Pratama *et al*. (2016), quienes basan su análisis en la causa de la falta de datos. En el caso particular de las técnicas de retropolación —como se explica en De la Fuente-Moreno (2014)—, la técnica usada con más frecuencia por organismos internacionales se basa en el encadenamiento o empalme de series de tiempo. De manera alternativa, Guerrero & Corona (2018) sugieren el concepto de retropolación restringida para *llevar hacia atrás* el Producto Interno Bruto (PIB) trimestral por entidad federativa de México, base 2008.

De esta forma, aunque existen trabajos previos que evalúan el funcionamiento de diversas técnicas de imputación y retropolación, no hay uno que sea integral y que valore a fondo el funcionamiento en muestras finitas de diversas técnicas en el contexto de series de tiempo y bajo diferentes estructuras estocásticas, como lo pueden ser la estacionariedad y la variabilidad. En este trabajo nos proponemos cubrir esa laguna en la literatura con el propósito particular de que nuestra propuesta sea de utilidad en el contexto de las EEN del INEGI, pero rescatando la aplicabilidad de las técnicas en escenarios más generales. En definitiva, el objetivo fundamental de esta investigación es evaluar el funcionamiento de diversas técnicas de imputación y retropolación usadas de manera frecuente en el contexto de series de tiempo. Partimos del hecho de que existe información que puede servir para imputarlas y/o retropolarlas; capturamos esta relación proponiendo como proceso generador de datos uno de factores dinámicos, en el cual las series de tiempo comparten factores comunes y se relacionan a través de ellos pero tienen, a su vez, una dinámica individual, también estocástica. Para el caso de imputación, usamos la especificación espacio-estado aplicando el suavizamiento de Kalman como método para imputar datos. También, se ponen a prueba diferentes tipos de *splines*, medias móviles y métodos de reemplazo basados en la información disponible dentro de la muestra. Asimismo, se considera la técnica basada en los *vecinos más cercanos*, la cual es usada con frecuencia en el caso de corte transversal, pero aplicable también a series de tiempo. Para el caso de retropolación, se utiliza el enfoque del empalme de series de tiempo, un algoritmo fácil de implementar y muy usado por distintos organismos internacionales generadores de información. Se especifican, además, dos técnicas novedosas de retropolación, ambas basadas en la estimación de modelos VAR, la primera realizando *backasting* o pronósticos *hacia atrás* y la segunda utilizando

dichos pronósticos, pero imponiendo restricciones lineales en la línea de Guerrero & Corona (2018).

El resto del artículo se organiza como sigue: en la segunda sección se describen los métodos de imputación y retropolación usados en este trabajo; en la siguiente se detalla el diseño del experimento Monte Carlo implementado para evaluar el rendimiento de los diferentes métodos de imputación y retropolación en muestras finitas; en la cuarta se evalúan las condiciones del experimento Monte Carlo en el contexto de las EEN; en la posterior se presentan los principales resultados de este experimento y, por último, se llega a las conclusiones y se hacen algunas recomendaciones.

2. Métodos de imputación y retropolación

2.1. Técnicas de imputación

2.1.1. Filtro de Kalman

La especificación general está dada por la representación gaussiana espacio-estado, donde la serie observada, denotada por y_t, viene dada por:

$$\begin{aligned} y_t &= Z_t \alpha_t + u_t, \\ \alpha_{t+1} &= T_t \alpha_t + R_t e_t, \end{aligned} \qquad (1)$$

para $t = 1,\ldots,T$. En la ecuación (1), $u_t \sim N(0,H_t)$, $e_t \sim N(0,Q_t)$ y $\alpha_1 \sim N(a_1,P_1)$. Para consultar más detalles sobre el modelo espacio-estado dado por la expresión (1) ver, entre otros, a Helske (2017). Usando el suavizamiento de Kalman, podemos estimar de forma recursiva, con toda la información disponible, la ecuación de estado como:

$$\hat{\alpha}_t = E(\alpha_t \mid y_1,\ldots, y_T). \qquad (2)$$

El primer método de imputación, llamado TKS, supone que $Z_t = T_t = R_t = 1$, $\alpha_t = t$ y $H_t = Q_t = 0.01$; es decir, es un modelo donde las observaciones se mueven alrededor de una tendencia determinística y donde existe poca variabilidad en los términos de error en la representación espacio-estado. El segundo, denominado LLKS, supone que α_t es una caminata aleatoria y que H_t y Q_t son estimadas por máxima verosimilitud (ver Moritz & Bartz-Beielstein, 2017). Nótese que este modelo, a diferencia de TKS, supone una tendencia estocástica.

2.1.2. *Splines*

El segundo grupo de métodos de imputación está basado en el uso de *splines* que, en esencia, representan una serie de interpoladores lineales y no-lineales para unir puntos entre datos faltantes. Matemáticamente diríamos que una función $s(y)$ es una *spline* en el intervalo $[a,b]$, si existe una partición del intervalo,

$$P = \{a = y_0 < y_1 < \ldots < y_T = b\} \qquad (3)$$

de tal forma que $s(y)$ es un polinomio, tal vez distinto, en cada intervalo $[y_i, y_{i+1}]$, $i = 0,1,\ldots,T$ - 1. En este caso, los puntos y_t se denominan los nodos del *spline*.

En este trabajo se utiliza la *spline* lineal, donde se unen los puntos a través de una proporcionalidad lineal entre los datos faltantes. También, se usan los procedimientos sugeridos por Forsythe *et al*. (1977), donde se ajusta una posición cúbica exacta a través de los cuatro puntos en cada extremo de los datos y el algoritmo de Stineman (1980), donde se computan pendientes más bajas cerca de pasos o picos abruptos en la secuencia de puntos. Al primero se le denota como LIN; al segundo, SPL; y al tercero, STI.

2.1.3. Medias móviles

Otro grupo de métodos de imputación está basado en la estimación de promedios o medias móviles basadas en el promedio de un número determinado, n, de observaciones centrales; en él se encuentran la simple (SMA), la ponderada de forma lineal

(LWMA) y la ponderada de manera exponencial (EWMA). En la SMA se pondera con el mismo peso a todas las observaciones que forman el promedio. En la LWMA, los pesos decrecen aritméticamente, es decir, las observaciones junto a un valor central i tienen un peso$1/2$; las observaciones siguientes (i - 2, i + 2), $1/3$; las que siguen (i - 3, i + 3), $1/4$ y así, en lo sucesivo. Por último, para el caso de la EWMA, los pesos van decreciendo de manera exponencial de la forma $(1/2)^i$.

2.1.4. Imputación basada en la información disponible en la muestra

Este grupo de procedimientos está sustentado en argumentos netamente heurísticos y se basan solo en utilizar el dato anterior o el siguiente disponible para rellenar los faltantes. En términos estadísticos, el método de arrastre de la última observación realizada (LOCF) está dado por:

$$\hat{y}_t = y_{t-J}, \quad J = \min(i | y_{t-i} \neq NA). \tag{4}$$

Por otro lado, el método de arrastre de la primera observación realizada (NOCB):

$$\hat{y}_t = y_{t+J}, \quad J = \max(i | y_{t+i} \neq NA). \tag{5}$$

En ambos casos, $\hat{y}_t$ es el estimador de y_t usando la información disponible en el momento i.

2.1.5. Método basado en los k vecinos más cercados (KNN)

Utiliza la información de una serie de tiempo, x_{jt}, para imputar los datos faltantes en la serie de tiempo de interés, y_t. De forma estadística, y para simplificar, si definimos k =1, el estimador KNN está definido de la siguiente manera:

$$\tilde{y}_t = \{x_t \mid \delta(x_{it}, y_t) < \delta(x_{jt}, y_t) \forall j\} \tag{6}$$

donde $\hat{y}_t$ representa el estimador KNN, $\delta(x_{it}, y_t) = | y_t - x_{jt} |$ es la distancia entre las series de tiempo y_j y x_{jt} al tiempo t para $j = 1, \ldots, K$, siendo K el número de series de tiempo disponibles como candidatos a vecinos más cercanos. De manera adicional, se puede reformular δ_j usando la información de $K \geq k > 1$ vecinos más cercanos. También, nótese que diversas formas funcionales de x_{jt} pueden ser usadas para imputar y_t. Para mayor detalle sobre este método de imputación, ver Kowarik & Templ (2016).

2.2. Técnicas de retropolación

2.2.1. Empalme de series de tiempo

Consiste en heredar el comportamiento pasado de una serie de tiempo a otra; por ejemplo, cuando una está desactualizada, pero tiene una temporalidad hacia atrás mayor que una más reciente, puede *llevarse hacia atrás* la más reciente con la temporalidad pasada de la desactualizada. Esto, claro está, resulta en la herencia de comportamientos específicos de la serie de tiempo desactualizada sin perder el nivel de la de interés y es una práctica común en los organismos encargados de generar información estadística oficial.

De manera más general, podemos explotar la correlación instantánea entre dos series de tiempo siempre que estas compartan un periodo común y que una de ellas tenga mayor información temporal en el pasado. En este contexto, se puede definir el estimador por empalme $\tilde{y}_h$ (SPLICE) de la siguiente manera:

$$\tilde{y}_h = f(x_h), \tag{7}$$

donde x_h se obtiene de maximizar la correlación de y_{t^*} con alguna variable del vector $X_{t^*} = (x_{1t^*}, \ldots, x_{Kt^*})'$, para $t^* = H + 1, \ldots, T$, donde $h = 1, \ldots, H$. Vale la pena destacar que es necesario que el vector (y_{t^*}, x_{jt^*}) esté compuesto por series de tiempo cointegradas para que la estimación de la correlación sea estadísticamente válida.

2.2.2. Retropolación: VAR

Denotemos las observaciones como el vector $Y_{t^*} = (Y_{t^*}, X_{t^*})'$ de dimensión N x 1, y supongamos que siguen un proceso VAR(p), es decir, existen matrices de coeficientes de dimensiones N x N y un vector de errores tales que:

$$Y_{t^*} = A_1 Y_{t^*-1} + \ldots + A_p Y_{t^*-p} + v_t \quad (8)$$

Es conocido que, ya que cada variable dependiente tiene las mismas variables independientes, la estimación de las A_j puede realizarse a través de mínimos cuadrados ordinarios (MCO). Asimismo, puede mostrarse que el mejor estimador lineal insesgado (MELI) de Y_{t^*+h} es $Y_{t^*+h|T}$ o, sin pérdida de generalidad, $Y_{h|T}$ es el MELI de Y_h. De esta forma, podemos utilizar los coeficientes de MCO y realizar la cadena de pronóstico para obtener:

$$Y_{h|T} = \hat{A}_1 Y_{h+1|T} + \ldots + \hat{A}_p Y_{h+p|T}. \quad (9)$$

Por consiguiente, pueden obtenerse las retropolaciones para Y_h usando el estimador $Y_{h|T}$ que se denota en este trabajo como VAR. Es importante considerar las características estocásticas de Y_{t^*}, pues si el modelo expresado en la ecuación (8) no es estacionario, estaremos ignorando relaciones de cointegración. La prueba más utilizada en el contexto de modelos multivariados viene dada por Johansen (1991); sin embargo, cuando las series son cointegradas y el objetivo es pronosticar, Lütkepohl (2005) argumenta que una representación VAR(p) en niveles es apropiada.

2.2.3. Retropolación restringida

Utilizando la regla general de combinación de Guerrero & Peña (2003) se pueden utilizar las estimaciones dadas por la expresión (9) e incorporar restricciones lineales para obtener predicciones restringidas de Y_h. Dichas restricciones cumplen la siguiente relación:

$$z = C y_{D(h-1)+d}, \quad (10)$$

donde $y_{D(h-1)+d} = (y_{D(h-1)+1}, \ldots, y_{D(h-1)+D})'$ y $C = \frac{1}{D}\mathbf{1}_D$, siendo D la frecuencia de la serie (por ejemplo, D =12 para datos mensuales) y $\mathbf{1}_D$ un vector de unos de dimensión D x 1. En particular, sean $Y_{h|T}$ las estimaciones preliminares de y_h, de forma que $Y_h = Y_{h|T} + E_h$, donde E_h es un proceso estacionario vectorial que admite la misma representación VAR(p), el MELI de y_h basado en $Y_{h|T}$ y en z está dado por:

$$\hat{y}_h = y_{h|t} + \hat{a}(z - C y_{D(h-1)+d}), \quad (11)$$

donde $\hat{a}$ se obtiene al expresar el VAR(p) en su forma de medias móviles, estimándose los coeficientes de manera recursiva. Para más detalles sobre la estimación de los coeficientes ver Guerrero & Corona (2018). Dicho estimador lo denotamos como RVAR.

Nótese que esta representación es multivariada, por lo que se obtienen todos los pronósticos restringidos para Y_h, de tal forma que las expresiones (10) y (11) están dadas para representaciones de series de tiempo múltiples, es decir, modelos VAR(p); sin embargo, decidimos enfocarnos en la variable que se desea retropolar con el objetivo de aligerar la notación.

3. Diseño del experimento Monte Carlo

El proceso generador de datos para evaluar el funcionamiento empírico de los métodos de imputación y retropolación está basado en una estructura de factores dinámicos. Una de las implicaciones más importantes de este diseño experimental es que las series de tiempo están relacionadas por uno o más factores comunes. Por ello, al imputar o retropolar, se podría estar realizando con series relacionadas entre sí. En este estudio, consideramos diferentes estructuras de dependencia y variabilidad en el componente idiosincrático de cada serie de tiempo. De esta manera, las observaciones, Y_t, siguen el proceso de factores dinámicos:

$$
\begin{aligned}
Y_t &= \mu + \lambda F_t + \varepsilon_t, \\
(\mathrm{I} - \phi L)(I - \Phi L^D) F_t &= (I + \Theta L^D)\eta_t, \\
\varepsilon_t &= \Gamma \varepsilon_{t-1} + a_t,
\end{aligned} \tag{12}
$$

donde μ es el vector de medias de las observaciones de dimensión N x 1, F_t y η_t son los factores comunes y sus respectivas innovaciones de dimensión r x 1, λ es una matriz de dimensión N x r que denota la contribución de los factores sobre las observaciones, ε_t y a_t son vectores de dimensión N x 1 que representan al componente idiosincrático y sus errores, respectivamente. Por otra parte, ϕ, Φ, Θ y Γ son matrices de coeficientes que, por simplicidad, se suponen diagonales. Las dimensiones de las primeras tres son r x r y de la última, N x N. Finalmente, L es el operador de diferencias tal que $LY_t = Y_{t-1}$. Conocemos al sistema de ecuaciones de la expresión (12) como un modelo de factores dinámicos (DFM, por sus siglas en inglés).

Asumiendo estacionariedad en la parte estacional, algunos resultados importantes recaen sobre las matrices de coeficientes, en específico sobre ϕ y Γ; por ejemplo, si algún elemento de la diagonal de ϕ es 1, entonces los factores no son estacionarios, por lo que Y_t no es estacionaria (si $\lambda \neq 0$). Más aún, si el componente idiosincrático es estacionario, cada par de series de tiempo del vector Y_t es cointegrado, es decir, los factores comunes son las tendencias comunes de las observaciones. Por otra parte, si el componente idiosincrático no es estacionario, habrá relaciones espurias en Y_t, es decir, cada serie de tiempo es una caminata aleatoria independiente.

Nuestro experimento Monte Carlo considera los siguientes modelos generales:

$$
\begin{aligned}
&\mathrm{M1}: \phi = 1, \\
&\mathrm{M2}: \phi = 0.5, \\
&\mathrm{M3}: \phi = \begin{bmatrix} 1 & 0 \\ 1 & 0.5 \end{bmatrix}.
\end{aligned} \tag{13}
$$

El primero (M1) corresponde al caso en el que las observaciones son generadas por una caminata aleatoria; el segundo (M2), a donde el factor es estacionario; y el tercero (M3) es la combinación de los dos anteriores. Asimismo, se consideran tres casos en la matriz de coeficientes VAR del componente idiosincrático, a saber, $\Gamma_1 = diag(-0.8)$, $\Gamma_2 = 0$ y $\Gamma_3 = diag(0.99)$, es decir, solo se consideran modelos cointegrados o estacionarios, aunque al asumir elementos de la diagonal en la matriz de coeficientes del componente idiosincrático cercanos a la unidad, nos acercamos al caso espurio. En todos se toma $\mu = 100$, $\lambda \sim U(0,1)$, $\Sigma_n = diag(1)$, $\Phi = diag(0.7)$ y $\Theta = diag(0.3)$, para $D = 12$.

Consideramos tres distintas matrices de covarianzas para las innovaciones del componente idiosincrático cuando es homocedástico, a saber $\Sigma_{a1} = diag(0.1)$, $\Sigma_{a2} = diag(1)$ y $\Sigma_{a3} = diag(10)$. Por otra parte, bajo heterocedasticidad especificamos las matrices de covarianzas como $\Sigma_{a1} = diag(U\sim[0.05, 0.15])$, $\Sigma_{a2} = diag(U\sim[0.5, 1.5])$ y $\Sigma_{a3} = diag(U\sim[5, 15])$. Por último, también consideramos el caso de errores correlacionados transversal y contemporáneamente siguiendo la sugerencia de Corona *et al.* (2017), donde se propone utilizar una estructura válida de correlación débil.

Los tamaños de muestra seleccionados son $N_1 = 15$, $T_1 = 100$ y $N_2 = 30$, $T_2 = 200$, considerando paneles de dimensiones similares a los de las EEN, tanto si fuesen por subsectores económicos como por entidad federativa. De esta forma, considerando tres dinámicas en el componente idiosincrático, tres variabilidades, tres tipos de error y dos tamaños de muestra, tenemos 3 x 3 x 3 x 2 = 54 distintas especificaciones por cada uno de los tres modelos en el experimento. Consideramos $R = 500$ réplicas y para cada diferente especificación eliminamos de forma aleatoria 24 o 48 observaciones, alrededor de 25% de la muestra. Para el estudio de la imputación, estas observaciones se eliminan de posiciones seleccionadas de manera aleatoria y para el de la retropolación del inicio de la muestra. En cada réplica, cada especificación y para cada serie de tiempo del DFM se aplican las técnicas de imputación y retropolación. La estimación del VAR(p) es bivariada, seleccionando la variable a retropolar y aquella que tenga más correlación con la variable objetivo. Cuando se incorporan restricciones lineales se asume que estas son conocidas y, asimismo,

de carácter temporal, es decir, que el promedio de cada tres periodos es conocido.

Para evaluar el funcionamiento de cada técnica, se considera como estadístico a la raíz cuadrada del error cuadrático medio (RMSE, por sus siglas en inglés). En cada réplica, y para cada panel de series de tiempo, se estima el valor promedio de los RMSE y su desviación estándar. Al finalizar las 500 réplicas se toman los promedios globales tanto para la media como para la desviación estándar para cada especificación.

Las medias móviles tienen un horizonte de $n = 4$ y usamos $k = 5$ para el método de KNN. El experimento Monte Carlo es programado y estimado con ayuda del programa R, usando las librerías de KFAS para la implementación del método TKS e imputeTS para la estimación de LLKS, LIN, SPL, STI, LOCF y NOCB. Para la aplicación del KSS se utiliza la librería VIM y, por último, para la estimación de los modelos VAR se usa la librería vars.

4. Estimación del Modelo de Factores Dinámicos: EEN

Con el objetivo de establecer que las condiciones en las que se genera el experimento Monte Carlo son válidas para los datos de las EEN, se estiman los componentes de diferentes DFM y se ajustan posibles modelos ARIMA estacionales (SARIMA, por su nombre en inglés) a los factores y modelos autorregresivos de orden 1, a los componentes idiosincráticos con el fin de estimar los valores de los parámetros autorregresivos y las varianzas del error del componente idiosincrático. La estimación de los factores y sus cargas asociadas se realizan a través de componentes principales (PC, por sus siglas en inglés) como lo propone Bai (2004), es decir, se asume *a priori* que el componente idiosincrático es estacionario. También, el número de factores, $\hat{r}$, es determinado por el criterio de Ahn & Horenstein (2013). Para la implementación de este ejercicio, se centran las observaciones en 100, de tal forma que tengan la misma media que las series simuladas en el experimento.

Vale la pena destacar que, si los factores comunes son estacionales, es conveniente usar otros métodos de estimación tal como Nieto *et al*. (2016); no obstante, nuestro objetivo no es desentrañar las características particulares de los DFM, sino solo investigar si la estimación tradicional de un DFM a través de un método no paramétrico como lo es PC captura características similares a las condiciones en las que es generado el experimento. Nótese que el supuesto fundamental para la estimación consistente de los factores y sus cargas a través de PC es que conforme N tiende a infinito, el efecto de los factores sobre las observaciones es lo que permanece, convergiendo a cero el efecto del componente idiosincrático. Corona *et al*. (2017) prueban que, en muestras finitas y con factores estacionarios y no estacionarios, esto ocurre con N alrededor de 15. Así, la estimación del modelo a través de PC es considerada suficiente para nuestros fines.

En consecuencia, para la estimación de los parámetros del DFM, se utilizaron las siguientes bases de datos, todas con año base = 2008:

- EMEC. Personal ocupado total de comercio al por mayor por entidad federativa, 2008:01-2018:07 ($N = 32$, $T = 127$).
- EMIM. Índice de personal ocupado por rama de la industria manufacturera, 2007:01-2018:07 ($N = 21$, $T = 139$).
- EMS. Ingresos totales por la prestación de servicios según sector y dominio, transportes, correos y almacenamiento, 2008:01-2018:07 ($N = 13$, $T = 127$).
- ENEC. Personal ocupado total por entidad federativa, 2006:01-2018:07 ($N = 132$, $T = 151$).
- EMOE. Indicador de confianza empresarial del sector, 2008:01-2018:09 ($N = 7$, T = 129).

El cuadro presenta los resultados de las estimaciones de los parámetros asociados al DFM, mostrándose el número estimado de factores y la especificación de los modelos SARIMA ajustados a los factores. Es importante señalar que dicha especificación a los factores es realizada a través de

Cuadro

Resumen de la estimación de los parámetros del DFM

Encuesta	$\hat{r}$	Modelo SARIMA $(p,d,q)(P,D,Q)$: F_t	Cuantiles 2.5, 50 y 97.5% de $diag(\Gamma)$	Cuantiles 2.5, 50 y 97.5% de $diag(\Sigma_a)$
EMEC	1	SARIMA(0,1,0)(1,0,0)[12]	0.54 0.91 0.98	0.08 0.64 5.78
EMIM	1	SARIMA(2,2,1)(2,0,0)[12]	0.93 0.99 1.00	0.00 0.01 0.51
EMS	1	SARIMA(0,1,1)(0,1,2)[12]	0.53 0.81 0.89	8.59 15.24 112.49
ENEC	1	SARIMA(1,1,0)(1,0,0)[12]	0.39 0.77 0.92	13.95 71.29 241.22
EMOE	1	ARIMA(1,1,1)	0.48 0.77 0.84	0.84 2.05 2.91

Nota: p = orden autorregresivo de la serie observada, d = orden de integración, q = orden autorregresivo de medias móviles, P = orden autorregresivo del componente estacional, D = orden de integración del componente estacional, Q = orden autorregresivo de medias móviles y 12 = estacionalidad mensual.

la selección automática que otorga la librería *forecast* del programa R, presentándose también los cuantiles 2.5, 50 y 97.5% tanto de $diag(\Gamma)$ como de $diag(\Sigma_a)$.

Se puede apreciar que, en todos los casos, de acuerdo con el criterio de Ahn & Horenstein (2013), un factor es suficiente para resumir la parte común de las series de tiempo para cada encuesta, aunque tal vez existan más factores si utilizamos otros criterios, por ejemplo, el de Onatski (2010). En todos los casos, con excepción de la EMOE, hay evidencia para afirmar que el factor es estacional y todos los factores tienen, al menos, una raíz unitaria en la parte no estacional. Solo para el caso de las EMS se detectó una raíz unitaria en la parte estacional. Los cuantiles asociados a la diagonal de Γ son positivos y sus valores varían entre 0.4 y 1, es decir, todas las autocorrelaciones de los componentes idiosincráticos son positivas y, en algunos casos, da lugar a DFM espurios. Para finalizar, los valores de $diag(\Sigma_a)$ llegan hasta 241.22, los cuales pueden considerarse muy grandes aunque para la EMEC, EMIM y EMOE los valores máximos de varianza se alcanzan en 5.78, siendo el máximo límite inferior de 13.95 para las cinco encuestas. De esta forma, el M1 parece adaptarse de buena forma a las bases de datos de las EEN tomadas como ejemplo; es decir, son paneles de series de tiempo cointegrados, con errores heterocedásticos y de muy distinta variabilidad, donde en algunos casos podemos encontrar relaciones espurias entre las variables.

5. Resultados del experimento Monte Carlo

5.1. Imputación

En esta subsección se presentan los RMSE para los métodos de imputación y retropolación descritos de forma previa. En ambos casos se hace énfasis en modelos con errores heterocedásticos dado que no se encontraron diferencias significativas entre modelos con errores homocedásticos, heterocedásticos y correlacionados de manera transversal, pero sí en términos de la estructura de autocorrelación en los componentes idiosincráticos y en los tamaños de la varianza de su respectivo error y de las muestras. En todas las gráficas se presentan los intervalos de confianza aproximados del RMSE a través de las 500 réplicas. En ellas, las columnas muestran los tres tamaños de la varianza del error del componente idiosincrático y los renglones, las dos medidas de muestra consideradas. Se presentan los resultados para los casos más representativos; no obstante, el resto está disponible bajo petición.

La gráfica 1 presenta los resultados para el M1 considerando $\Gamma = diag(-0.8)$, es decir, cuando los errores idiosincráticos tienen correlación serial negativa. En este caso, las observaciones son generadas por un factor común no estacionario y, dada la estructura del error, las series de tiempo están

cointegradas para cada par de estas y, también, los factores comunes representan la tendencia común de las observaciones.

Se puede apreciar que, cuando la varianza del error del componente idiosincrático es pequeña, todos los criterios presentan errores menores a la unidad y que SPL es el método cuyo RMSE se incrementa conforme la varianza del componente idiosincrático lo hace. Nótese que cuando la varianza del error del componente idiosincrático alcanza su máximo valor, $\sigma_a^2 = 10$, todos los criterios incrementan su variabilidad, siendo TKS el método con menor RMSE, tanto en el promedio de sus medias como de sus desviaciones estándar a través de las réplicas. El efecto del tamaño de la muestra es casi imperceptible, excepto cuando $\sigma_a^2 = 10$, caso en el que los intervalos de confianza son más pequeños cuando la muestra es más grande, i.e., cuando $N = 30$ y $T = 200$.

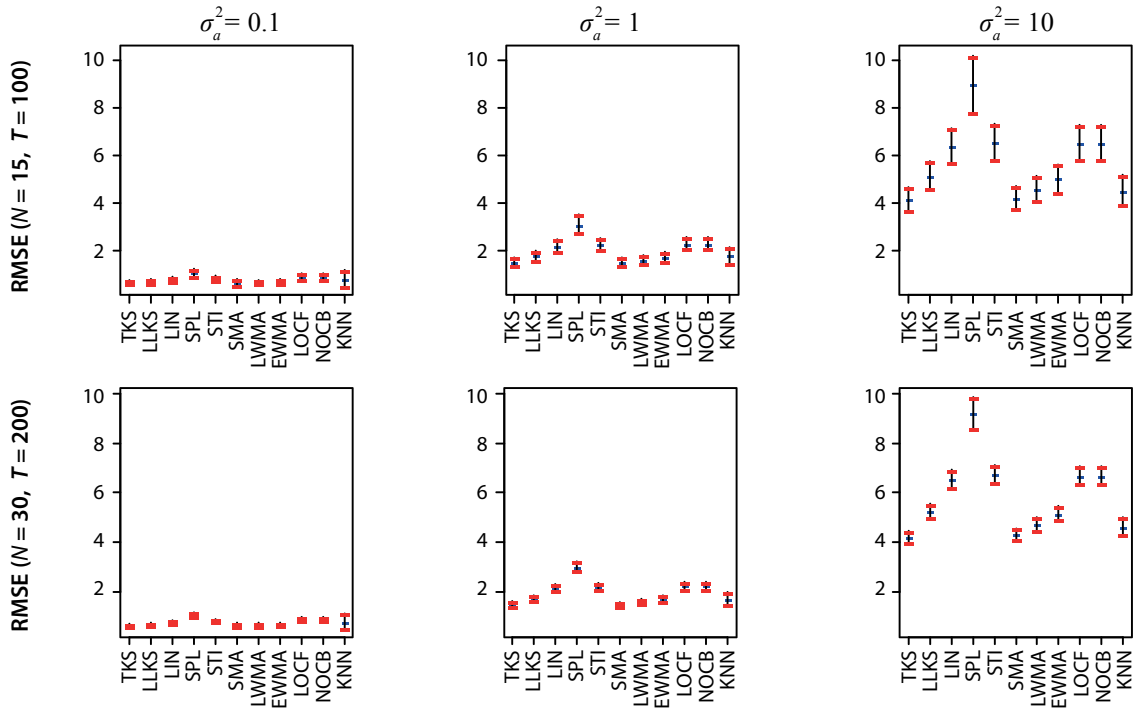
La gráfica 2 muestra los mismos resultados que la 1 cuando $\Gamma = diag(0)$, es decir, cuando los componentes idiosincráticos son ruido blanco. En esta situación, al igual que el caso anterior, las series de tiempo son cointegradas.

En la gráfica 2 se puede denotar que KNN presenta una mayor variabilidad respecto a los otros métodos. Llama la atención que su variabilidad es más grande que la del resto de los procedimientos cuando $\sigma_a^2 = 0.01$. Asimismo, es interesante observar que conforme σ_a^2 aumenta, SPL es el procedimiento que tiende a resentirlo más, incrementándose la media del RMSE a través de las réplicas. De forma similar a lo obtenido en la gráfica 1, el tamaño de muestra tiene el mismo efecto en todos los procedimientos, siendo los intervalos de confianza menores en el tamaño de muestra más grande cuando $\sigma_a^2 = 10$.

En la gráfica 3 se presentan los resultados de los intervalos de confianza de los RMSE estimados a través de las réplicas para cada uno de los métodos de imputación para el M1 cuando $\Gamma = diag(0.99)$.

Gráfica 1

Intervalos de confianza (95%) de los RMSE para los distintos métodos de imputación para M1 con errores heterocedásticos y $\Gamma = diag(-0.8)$



Nótese que esta parametrización representa el caso cuando las variables están siendo simuladas casi como unas caminatas aleatorias independientes, aunque en términos teóricos, las series de tiempo siguen siendo cointegradas.

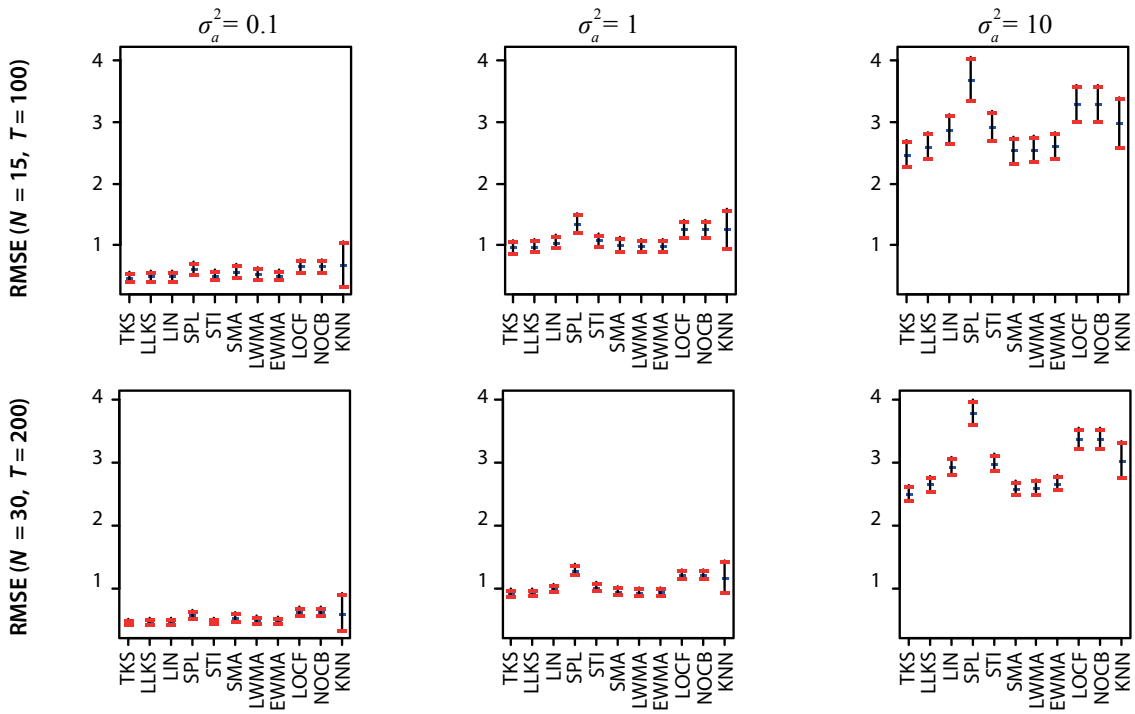
Puede observarse que cuando $\sigma_a^2 = 0.1$ o $\sigma_a^2 = 1$, todos los procedimientos de imputación, salvo el KNN y los métodos basados en la información disponible en la muestra, tienden a funcionar de manera similar mostrando un límite superior del intervalo confianza que no rebasa a 1. Por otra parte, cuando $\sigma_a^2 = 10$, los intervalos superiores se incrementan, pero lo realizan apenas sobrepasando el 1 en algunos casos, con excepción del KKN, que llega ahora a rondar las tres unidades. Este resultado es esperado, dado que KNN funciona bien conforme *los vecinos* aportan información relevante para imputar datos faltantes y esto se da en el caso de paneles cointegrados, es decir, cuando la diagonal de Γ sea menor a la unidad en valor absoluto. Asimismo, nótese que la varianza del componente idiosincrático es también una función de σ_a^2 y conforme esta sea más grande, más variabilidad habrá entre las series que forman al DFM, lo que puede afectar el funcionamiento de este método.

Para M2 y M3 solo presentamos los casos cuando $\Gamma = diag(0)$ y $\Gamma = diag(0.99)$, respectivamente, ya que, como se denotó en la subsección anterior, este tipo de dinámicas autorregresivas son las situaciones que más se presentan en la práctica. La gráfica 4 muestra el funcionamiento de los procedimientos de imputación para el M2 cuando $\Gamma = diag(0)$. En este caso, el factor común que genera a las observaciones es estacionario, por lo que las series de tiempo se mueven alrededor de una variable que tanto su media como su varianza no dependen del tiempo.

Se puede apreciar que el método del KNN presenta tanto valores medios como intervalos de confianza un poco más pequeños en los RMSE que los otros procedimientos, sobre todo cuando

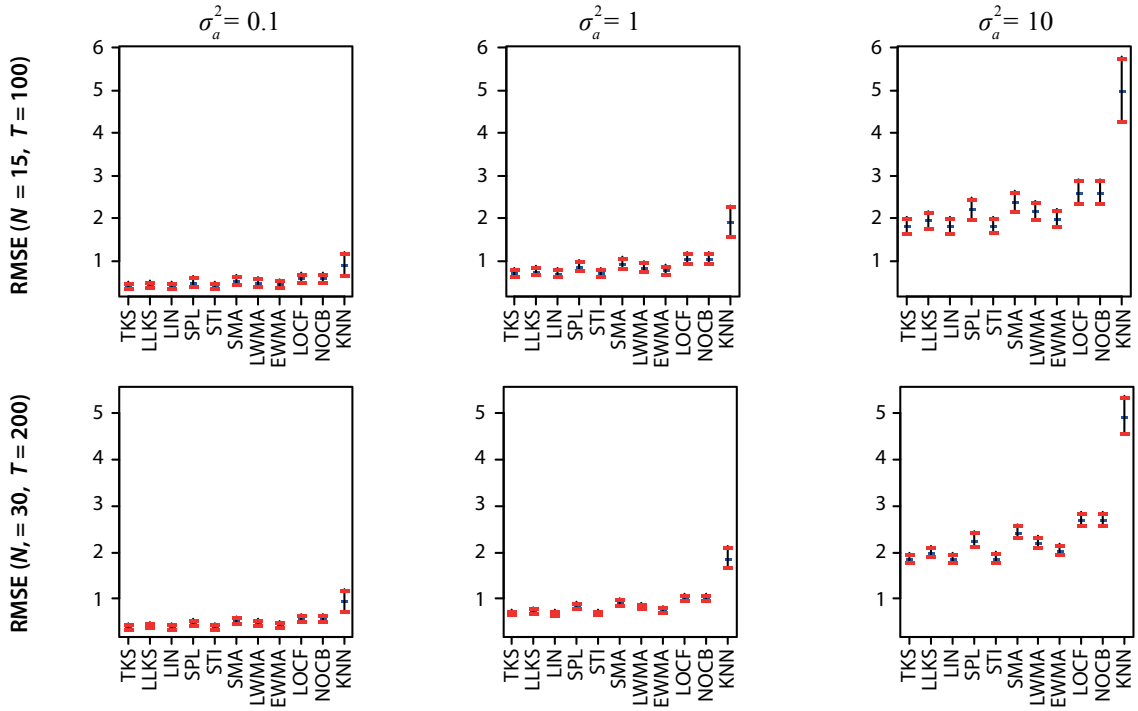
Gráfica 2

Intervalos de confianza (95%) de los RMSE para los distintos métodos de imputación para M2 con errores heterocedásticos y $\Gamma = diag(0)$



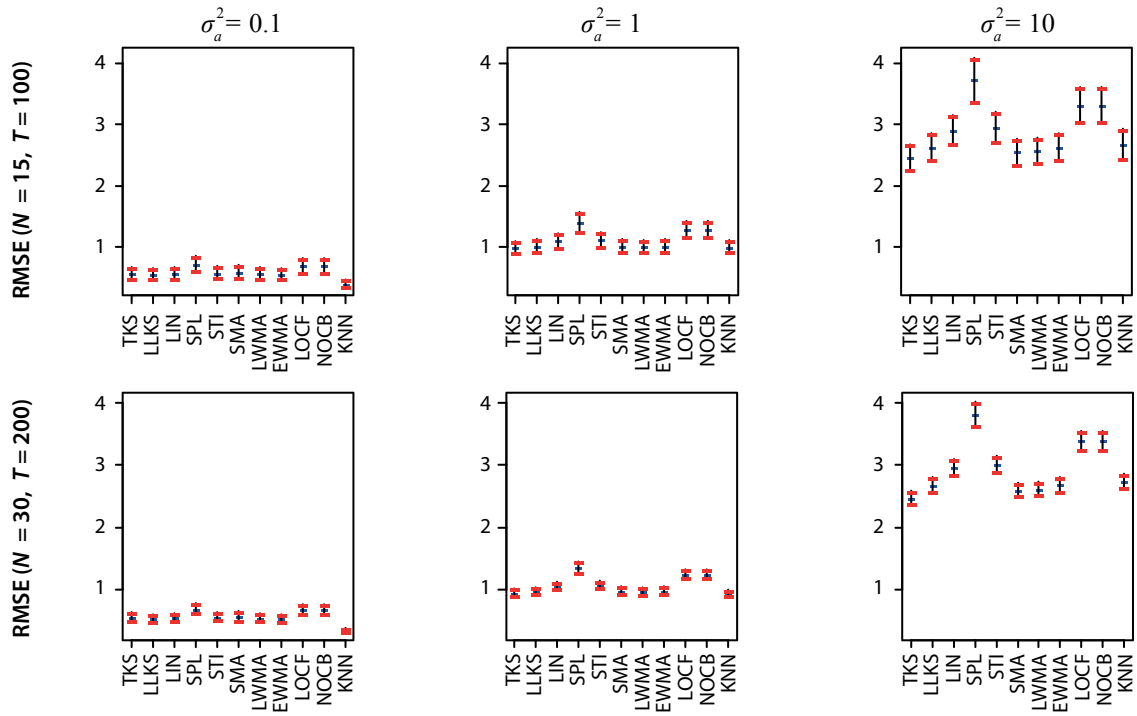
Gráfica 3

Intervalos de confianza (95%) de los RMSE para los distintos métodos de imputación para M1 con errores heterocedásticos y $\Gamma = diag(0.99)$



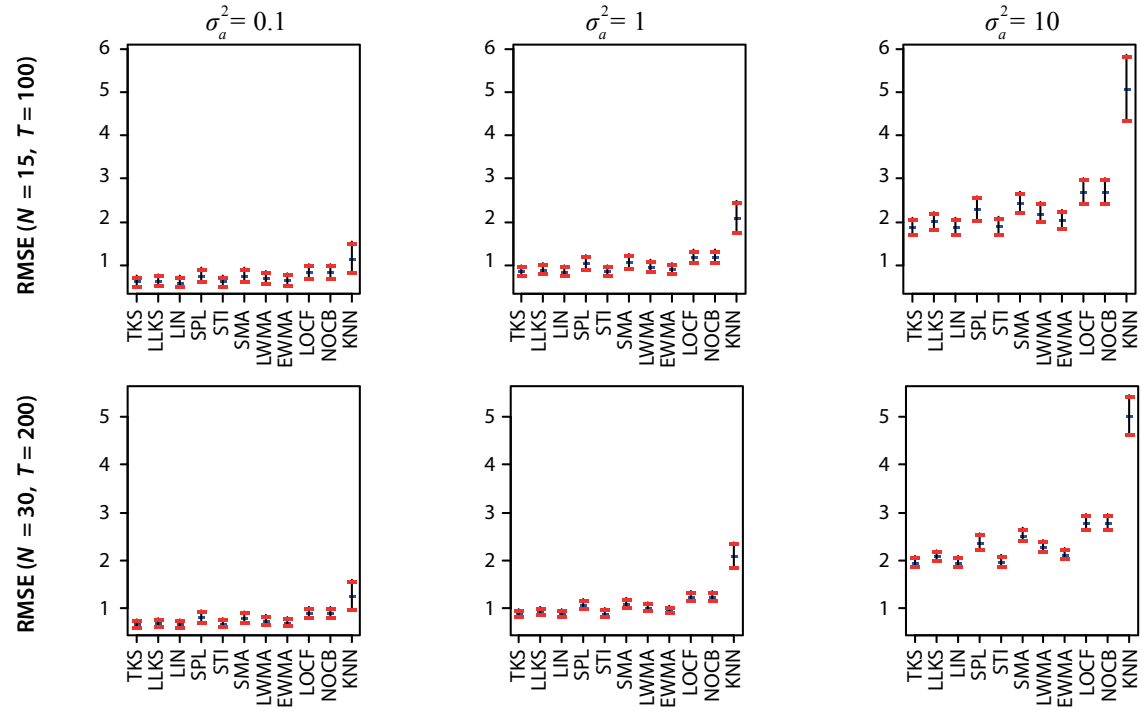
Gráfica 4

Intervalos de confianza (95%) de los RMSE para los distintos métodos de imputación para M2 con errores heterocedásticos y $\Gamma = diag(0)$



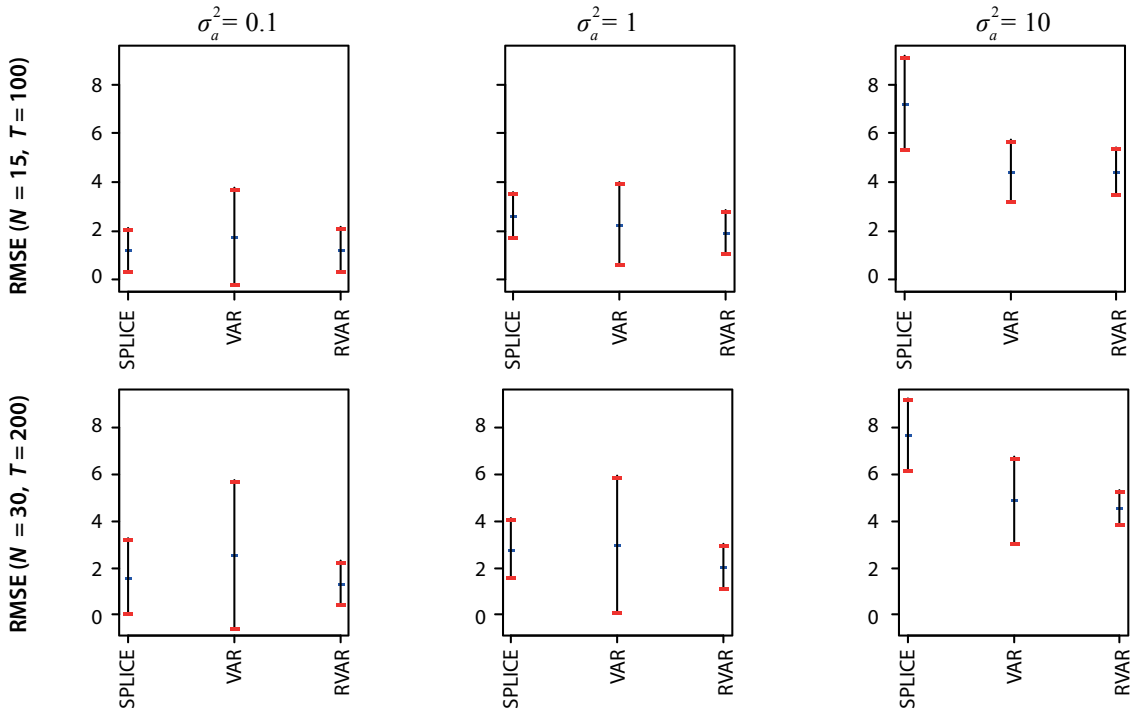
Gráfica 5

Intervalos de confianza (95%) de los RMSE para los distintos métodos de imputación para M3 con errores heterocedásticos y $\Gamma = diag(0.99)$



Gráfica 6

Intervalos de confianza (95%) de los RMSE para los distintos métodos de retropolación para M1 con errores heterocedásticos y $\Gamma = diag(-0.8)$



$\sigma_a^2 < 10$. Sin embargo, cuando $\sigma_a^2 = 10$, los métodos presentan una variabiliad similar y, más aún, los intervalos de los RMSE tienden a intersectarse con excepción de SPL, LOCF y NOCB, por lo que, en la práctica, el resto de los métodos de imputación que no son estos tres tienden a funcionar casi igual cuando la varianza del error del componente idiosincrático no es grande.

Para finalizar, la gráfica 5 muestra los mismos resultados que la 4, pero ahora para el M3 considerando el caso cuando el componente idiosincrático tiene una matriz de coeficientes autorregresivos $\Gamma = diag(0.99)$. Este caso es de interés porque, aunque en teoría, las series de tiempo son cointegradas, notemos que, dependiendo de las cargas de λ, las series pueden estar al límite de ser caminatas aleatorias independientes, por ejemplo cuando $y_{it} = \lambda_1 F_{1t} + \varepsilon_{it} (\lambda_2 = 0)$ o bien, el caso de cuando comparten un factor estacionario, pero domina el comportamiento individual no estacionario, es decir, cuando $y_{it} = \lambda_2 F_{2t} + \varepsilon_{it}$ $(\lambda_1 = 0)$, bajo el argumento de que ε_t tiende a ser una serie de tiempo no estacionaria.

Se puede apreciar que KNN funciona de manera deficiente. Esto tiene que ver con que estamos utilizando información de *vecinos* con los cuales la relación puede considerarse, casi, espuria. Este efecto se vuelve aún más importante conforme σ_a^2 crece. La presencia del factor estacionario es irrelevante y esto está relacionado con el hecho de que la variación en las observaciones es dominada por la variación del factor común no estacionario.

Como conclusión, y dadas las características consideradas en este experimento, los métodos basados en el filtro de Kalman tienden a funcionar, en promedio, un poco mejor que el resto de los procedimientos, sobre todo el TKS. Otro resultado importante es que, ya que la variabilidad alrededor de la media es similar entre los basados en medias móviles y el filtro de Kalman, podemos concluir que, ya que los intervalos se intersectan, dichas familias de métodos tienden a funcionar de manera similar. También, el utilizar un *spline* lineal funciona, incluso, mejor que el de la misma familia, es decir, el interpolador denotado como STI. Asimismo, los métodos más débiles para imputar datos resultaron ser el SPL y los basados en la información disponible en la muestra. Otra conclusión muy importante es que, aunque el KNN tiene un pobre funcionamiento cuando no hay *buenos vecinos,* funciona bien cuando los vecinos utilizados aportan información relevante y esto ocurre con frecuencia en la práctica.

5.2. Retropolación

En este contexto, como comentamos en la sección 3, nos interesa saber qué método *retropola* de forma más acertada las observaciones faltantes. La gráfica 6 presenta los resultados de los métodos de retropolación para el M1 cuando $\Gamma = diag(-0.8)$.

Resulta interesante observar que, sin considerar el tamaño de muestra y de la varianza asociada al error del componente idiosincrático, el RVAR tiende a funcionar hasta cierto punto mejor, dado que obtenemos medias y varianzas más pequeñas en los RMSE calculadas a partir de las réplicas. En otras palabras, tenemos estimaciones con una mayor precisión y una menor incertidumbre. No obstante, cuando la varianza del error del componente idiosincrático es pequeña, SPLICE es muy competitivo. Esto es relevante si consideramos lo flexible que resulta implementar este método respecto al VAR y al RVAR, los cuales requieren una mayor cantidad de información y supuestos para su respectiva implementación. Solo en el caso cuando $\sigma_a^2 = 10$, el método RVAR —que supone la existencia de restricciones, en este caso temporales— es claramente el mejor.

Los resultados obtenidos para el M1 al ir variando el parámetro de autocorrelación en los componentes idiosincráticos no muestran cambios relevantes respecto a lo obtenido en la gráfica 6, motivo por el cual solo presentamos este caso para el M1.

La gráfica 7 muestra los resultados encontrados para el M2 cuando $\Gamma = diag(0)$, es decir, cuando

las observaciones tienen un factor común estacionario y los errores individuales son ruido blanco.

Podemos apreciar resultados parecidos a los mostrados en la gráfica 6, aunque los intervalos para RVAR se hacen más grandes, lo cual nos muestra que las predicciones tienen una incertidumbre más alta en el caso estacionario que en el no estacionario. Nótese que al generar series de tiempo cointegradas, aprovechamos su tendencia común al pronosticarlas, motivo por el cual las predicciones son más certeras en el caso no estacionario. En este modelo también se observa que cuando $\sigma_a^2 = 10$, el método RVAR es radicalmente mejor tanto respecto al SPLICE como al VAR.

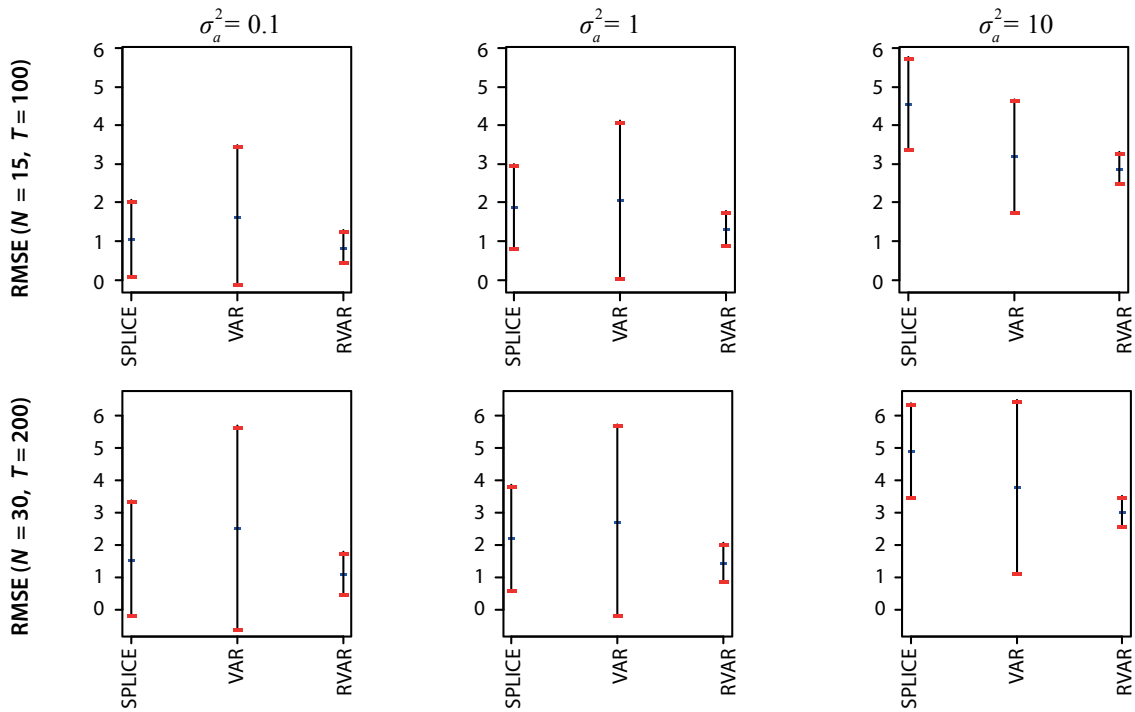
Por último, la gráfica 8 presenta lo obtenido para el M3 cuando $\Gamma = diag(0.99)$.

Solo se pueden apreciar algunos cambios respecto a la gráfica 6, es decir, y al igual que en el caso de imputación, el factor no estacionario domina sobre el estacionario; no obstante, podemos apreciar ahora que, cuando $\sigma_a^2 = 1$ y el tamaño de muestra es más grande, el RVAR presenta intervalos de confianza más pequeños que el SPLICE y el VAR.

Un resultado importante es que, aun cuando el RVAR da resultados más certeros y con menor incertidumbre, el SPLICE puede considerarse estadísticamente competitivo, sobre todo en los casos cuando la variabilidad individual de las observaciones es pequeña, incluso en algunos casos funciona mejor que el VAR y esto resulta de utilidad porque su implementación requiere solo del supuesto que la variable de empalme esté correlacionada con la variable a retropolar. Nótese que el VAR y el RVAR requieren de un número suficiente de observaciones para poder ser implementados (se estiman $N + Np^2$ parámetros), además de que, de manera empírica, es necesario validar que las series son cointegradas (pronóstico en niveles) y que los residuales están libres de autocorrelación.

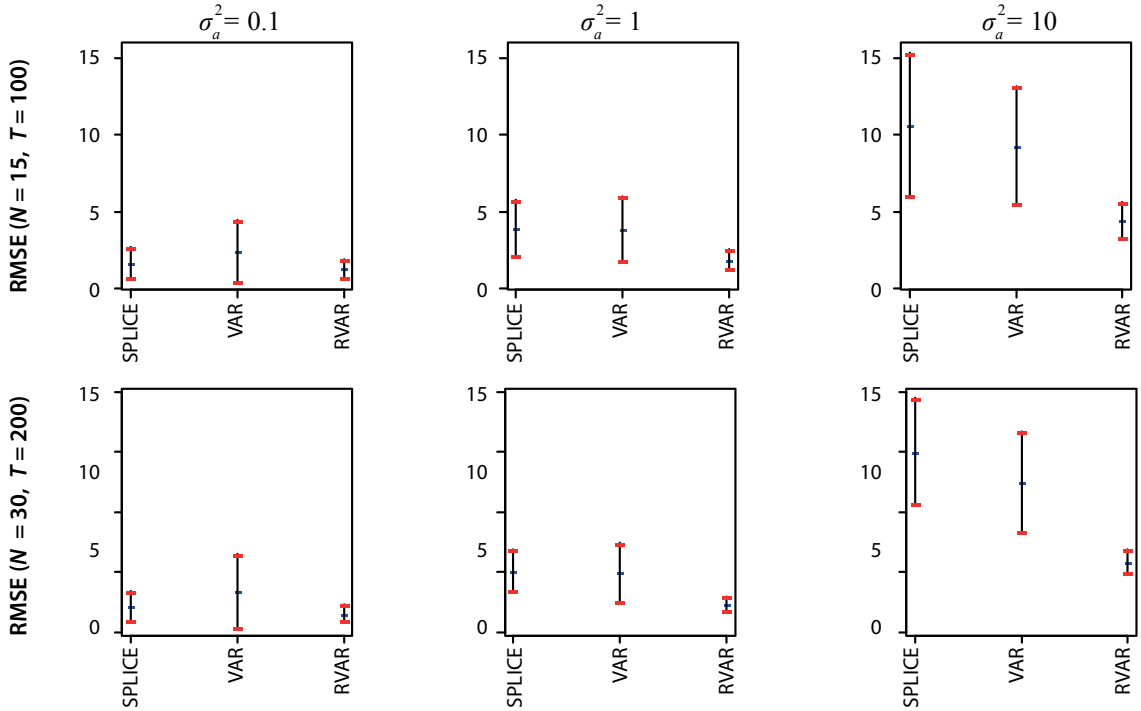
Gráfica 7

Intervalos de confianza (95%) de los RMSE para los distintos métodos de retropolación para M1 con errores heterocedásticos y $\Gamma = diag(0)$



Gráfica 8

Intervalos de confianza (95%) de los RMSE para los distintos métodos de retropolación para M3 con errores heterocedásticos y $\Gamma = diag(0.99)$



6. Conclusiones y recomendaciones

Este trabajo se orientó en evaluar el funcionamiento en muestras finitas de diferentes métodos de imputación y retropolación en el contexto de series de tiempo, cuyos resultados puedan ser útiles para los generadores de información oficial. Nos enfocamos en simular paneles de series de tiempo como DFM, cuyas características fueron validadas para cinco bases de datos de las EEN del INEGI. En este sentido, se consideraron DFM estacionales con factores no estacionarios, estacionarios y la combinación de ambos, variando sobre todo el tamaño de muestra, los parámetros de autocorrelación en el componente idiosincrático y diferentes estructuras de dependencias en estos, como la homocedasticidad, heterocedasticidad y la autocorrelación cruzada. En otras palabras, se simularon series de tiempo que se observan con frecuencia en la práctica. Se pudo ver que, en el contexto de imputación, los métodos que utilizan en el filtro de Kalman son los más competitivos, aunque su rendimiento es similar a los que se basan en medias móviles. También, el KNN es el más competitivo cuando la información de los *vecinos más cercanos* es de calidad, situación relacionada con el caso de DFM cointegrados. Los métodos basados en la información disponible en la muestra y *splines* cúbicas resultaron ser los menos precisos. En lo que respecta a la retropolación, el RVAR es de forma clara el mejor de los tres analizados; no obstante, se requiere conocer, en cada caso particular, las restricciones lineales adecuadas y necesarias para su implementación. En caso de no conocer estas restricciones, se observa que el SPLICE otorga con frecuencia mejores resultados que los obtenidos a través de un VAR.

De esta forma, se recomienda a los generadores de información que, para imputar, es útil usar in-

formación de variables la cual esté correlacionada en algún sentido, estadístico o económico, con la serie de tiempo de interés. En caso de no contar con este tipo de información, se recomienda preferir los métodos basados en el filtro de Kalman o en medias móviles. Para retropolar, es claro que tener más información del pasado de la serie de tiempo permite minimizar el error atribuible a la estimación, lo anterior restringiendo los pronósticos obtenidos en una fase anterior a través de un modelo VAR; no obstante, si no existen dichas restricciones, empalmar series de tiempo correlacionadas entre sí resulta una alternativa flexible y útil.

Fuentes

Ahn, Seung C. y Alex R. Horenstein. "Eigenvalue ratio test for the number of factors", en: *Econometrica*. Vol. 81, núm. 3. 2013, pp. 1203-1227.

Bai, Jushan. "Estimating cross-section common stochastic trends in nonstationary panel data", en: *Journal of Econometrics*. Vol. 122, núm. 1. 2004, pp. 137-183.

Corona, Francisco, Pilar Poncela y Esther Ruiz. "Estimating non-stationary common factors: Implications for risk sharing", en: *DES-Working Papers. Statistics and Econometrics*. WS 24585. Departamento de Estadística, Universidad Carlos III de Madrid, 2017.

De la Fuente Moreno, Ángel. "A mixed splicing procedure for economic time series", en: *Estadística Española*. Vol. 56, núm. 183. 2014, pp. 107-121.

Forsythe, George. E., Cleve B. Moler y Michael A. Malcolm. *Computer Methods for Mathematical Computations*. Prentice-Hall, 1977.

Guerrero, Víctor M. y Blanca I. Gaspar. "Edition and Imputation of Multiple Time Series Data Generated by Repetitive Surveys", en: *Journal of Data Science*. Vol. 8, núm. 4. 2010, pp. 555-577.

Guerrero, Víctor M. y Daniel Peña. "Combining multiple time series predictors: a useful inferential procedure", en: *Journal of Statistical Planning and Inference*. Vol. 116, núm 1. 2003, pp. 249-276.

Guerrero, Víctor M. y Francisco Corona. "Retropolating some relevant series of Mexico's System of National Accounts at constant prices: The case of Mexico City's GDP", en: *Statistica Neerlandica*. Vol. 72, núm. 4. 2018, pp. 495-519.

Harvey, Andrew C. y Richard G. Pierse. "Estimating missing observations in economic time series", en: *Journal of the American Statistical Association*. Vol. 79, núm. 385. 1984, pp. 125-131.

Helske, Jouni. "KFAS: Exponential Family State Space Models in R", en: *Journal of Statistical Software*. Vol. 78, núm 10. 2016, pp. 1-39.

Hyndman, Robert J. y Yeasmine Khandakar. "Automatic time series forecasting: the forecast package for R", en: *Journal of Statistical Software*. Vol. 26, núm. 3. 2007, pp. 1-22.

Johansen, Søren. "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models", en: *Econometrica*. Vol. 59, núm. 6. 1991, pp. 1551-1580.

Kowarik, Alexander and Matthias Templ. "Imputation with R package VIM", en: *Journal of Statistical Software*. Vol. 74, núm. 7. 2016, pp. 1-16.

Lütkepohl, Helmut. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Luzi, Orietta., *et al*. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. ISTAT, CBS, SFSO, Eurostat, 2007.

Moritz, Steffen *et al. Comparison of different methods for univariate time series imputation in R*. arXiv preprint arXiv:1510.03924, 2015.

Moritz, Steffen y Thomas Bartz-Beielstein. "ImputeTS: time series missing value imputation in R", en: *The R Journal*. Vol. 9, núm. 1. 2017, pp. 207-218.

Nieto, Fabio. H., Daniel Peña y Dagoberto Saboyá. "Common seasonality in multivariate time series", en: *Statistica Sinica*. Vol 26, núm. 4. 2016, pp. 1389-1410.

Onatski, Alexei. "Determining the number of factors from empirical distribution of eigenvalues", en: *The Review of Economics and Statistics*. Vol. 92, núm 4. 2010, pp. 1004-1016.

Peña, Daniel & George C. Tiao. "A note on likelihood estimation of missing values in time series", en: *The American Statistician*. Vol. 45, núm. 3. 1991, pp. 212-213.

Pfaff, Bernhard. "VAR, SVAR and SVEC Models: Implementation Within R Package vars", en: *Journal of Statistical Software*. Vol. 27, núm. 4. 2008, pp. 1-32.

Pfeffermann, Danny y Gad Nathan. "Imputation for wave nonresponse: Existing methods and a time series approach", en: Groves, Robert *et al*. (eds.). *Survey Nonresponse*. New York, Wiley, 2001, pp. 417-429.

Pratama, Irfan, *et al*. "A review of missing values handling methods on time-series data", en: *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*. Bandung, Indonesia, octubre de 2016, pp. 1-6.

R Core Team. "R: A language and environment for statistical computing", en: *R Foundation for Statistical Computing*. Vienna, Austria, 2017 (DE) https://www.R-project.org/

Schafer, Joseph L. y John W. Graham. "Missing data: our view of the state of the art", en: *Psychological Methods*. Vol. 7, núm. 2. 2002, pp. 147-177.

Schmitt, Peter, Jonas Mandel y Mickael Guedj. "A comparison of six methods for missing data imputation", en: *Journal of Biometrics & Biostatistics*. Vol. 6, núm. 1. 2015, pp. 1-6.

Stineman, Russel W. "A Consistently Well Behaved Method of Interpolation", en: *Creative Computing*. Vol. 6, núm. 7. 1980, pp. 54-57.