CAPÍTULO IV

Enfoque de *big data* para generar y analizar datos de actividad económica en México

Víctor M. Guerrero*
Francisco Corona
Juan Antonio Mendoza

El objetivo de este artículo es presentar dos trabajos realizados con un enfoque basado en el uso eficiente, desde un punto de vista estadístico, de los datos más relevantes disponibles para solucionar problemas que enfrentan en la actualidad las agencias de estadística oficial, especialmente en México. Los casos que se presentan son: 1) estimación del producto interior bruto (PIB) desde el espacio exterior, que considera la combinación de datos oficiales provenientes de las Cuentas Nacionales, con datos de luminosidad nocturna producidos por mediciones de satélites; 2) retropolación de series de las Cuentas Nacionales, es decir, extrapolación hacia atrás, con apoyo en fuentes diversas y heterogéneas. El lazo unificador de estos trabajos se encuentra en el hecho de que en los dos casos se enfrenta una o más de las 5 Vs que caracterizan a los problemas relacionados con big data o sea, la presencia de un gran *volumen* de datos, con alta *velocidad* de aparición de nuevos datos, amplia *variedad* de fuentes de información, que aportan diferente *valor* y con distintos grados de *veracidad.* En los casos estudiados se buscó obtener información útil, a partir de los datos disponibles y se hizo uso de metodología estadística validada por los datos mismos, para asignar optimalidad a los resultados obtenidos.

Palabras clave: combinación de información, estadística oficial, medición económica, modelos de series de tiempo, retropolación.

* V. M. Guerrero agradece a la Asociación Mexicana de Cultura, A. C. el apoyo brindado, mediante la Cátedra de Análisis de Series de Tiempo y Pronósticos en Econometría, para la realización de este proyecto. Asimismo, los autores agradecen a los editores del presente volumen sus comentarios y sugerencias, que permitieron mejorar la presentación de este artículo, así como la invitación a contribuir con el mismo a esta obra.

1. INTRODUCCIÓN

Dentro del ámbito de la estadística oficial, es común que se requiera aplicar alguna herramienta metodológica de carácter estadístico para ampliar la cobertura o tratar de mejorar la medición económica de alguna(s) variable(s) relevante(s) para la toma de decisiones, tanto a nivel gubernamental, como a nivel de las empresas y de los individuos en general. Lograr la ampliación de cobertura o mejorar la medición de alguna variable requiere que el analista a cargo de dicha labor sea capaz de incorporar nuevos datos o cambiar algún sistema ya establecido en la Agencia Oficial de Estadística (AOE) respectiva y, sobre todo, de convencer con argumentos sólidos de teoría estadística y de resultados con validez empírica, a los encargados de los sistemas estadísticos dentro de la AOE. Para esto, no es suficiente con tener alguna idea "ingeniosa e innovadora" acerca de dónde y cómo obtener datos alternativos, sino en proponer al mismo tiempo la manera de traducir dichos datos en información de calidad, que contribuya a ampliar la oferta que ya brinda la AOE.

Una de estas situaciones se presenta al tratar de mejorar la calidad de la medición económica que se hace con el cálculo oficial del producto interno bruto (PIB). Es bien sabido que el PIB presenta defectos asociados, por ejemplo, con la medición de la economía informal o ilegal y es por ello que recurrir a una medición alternativa del PIB resulta atractivo. Esto se logra precisamente con la incorporación de cifras que se obtienen de la luminosidad, medida a través de imágenes satelitales (*e. g.* Ghosh *et al.,* 2009). Un trabajo pionero de este tipo de enfoque es el que realizaron Henderson, Storeygard y Weil (2012) quienes combinaron datos satelitales con datos oficiales de crecimiento económico para varios países. Ellos usaron datos de tipo panel para diversos países y los ponderaron de acuerdo con las calificaciones de calidad de los datos de cada país que determina el Banco Mundial. Para el caso de México, Guerrero y Mendoza (2019) optaron por el uso exclusivo de datos correspondientes a cada país de forma individual, sin necesidad de recurrir a evaluaciones de la calidad de los datos estadísticos oficiales. Esto va más de acuerdo con lo que se hace para calcular las cifras oficiales del PIB, que no deben incorporar datos externos al país en cuestión, sino únicamente información relacionada con la actividad realizada en dicho país y, de preferencia, provista por fuentes internas.

Otro tema de ampliación de cobertura se refiere al ámbito temporal, pues en ocasiones la longitud de las series de datos oficiales es relativamente corta y no permite realizar un análisis adecuado de la economía. Esta situación se presentaba en México en el año 2016, ya que las series trimestrales de la contabilidad nacional que estaban disponibles al público de manera oficial, a nivel de las entidades federativas del país, cubrían de manera homogénea solamente los años 2003 a 2015. Esto no significaba que no hubiera datos disponibles para años previos, sino que los datos para dichos años no satisfacían criterios de homogeneidad adecuados. Por ello existía la necesidad de homogeneizar los datos, en lo que toca a diversos criterios que permiten realizar comparaciones válidas.

El problema que enfrentaron Guerrero y Corona (2018a, 2018b) fue del tipo recién descrito y por ello aplicaron una variedad de procedimientos estadísticos que condujeron a generar una base de datos uniforme en lo que se refiere a los siguientes cuatro criterios de

clasificación: sectorial, temporal, geográfico y de año base, según se describe más adelante. Esa base de datos permitió ampliar, con datos de 1993 en adelante, la oferta de información oficial del Instituto Nacional de Estadística y Geografía (INEGI), que es la AOE de México. Para esto se hizo uso de las diferentes bases heterogéneas y se emplearon modelos de series de tiempo. Esto comúnmente no es aceptado por las AOE (véase al respecto Braaksma y Zeelenberg, 2015) y el INEGI no es la excepción al respecto; por ello hubo necesidad de someter los resultados a diversas verificaciones empíricas realizadas por los técnicos encargados de mantener actualizado el Sistema de Cuentas Nacionales de México en el INEGI, hasta que se convencieron de la fiabilidad de los datos generados y fue entonces que ya se les pudo considerar como datos "oficiales". Otro elemento que fortaleció la decisión de considerar los resultados de la retropolación efectuada como oficiales fue la publicación de Corona y López (2020), que les brindó a las cifras estimadas validez desde una perspectiva de análisis econométrico.

La relación que existe entre los dos problemas considerados se refiere a la presencia de una o más de las 5 Vs que caracterizan el análisis de big data, es decir, la presencia de un gran *volumen* de datos, con alta *velocidad* de aparición de nuevos datos, amplia *variedad* de fuentes de información, que aportan datos con diferente *valor* y con distintos grados de *veracidad* (Gupta *et al.,* 2018). La organización de este documento es como sigue: en la sección segunda se presenta el caso de la combinación de datos oficiales de actividad económica con datos provenientes de imágenes de satélite, ahí se describe el tipo de datos producidos por los satélites y la forma en que se pueden combinar con las cifras oficiales de un país en específico. También se muestra en esa sección una aplicación para estimar el crecimiento del PIB verdadero de México. La sección tercera se ocupa del tema de la retropolación, es decir, de la extrapolación hacia atrás, de las series de tiempo trimestrales del PIB de México hasta 1993, clasificado por las tres grandes actividades económicas, para los 32 estados que conforman al país y medido todo a precios constantes del año 2013. Para lograr esto se tienen en cuenta todas las fuentes de información de carácter oficial disponibles. Finalmente, en la cuarta sección se emiten algunas conclusiones y recomendaciones.

2. COMBINACIÓN DE DATOS OFICIALES CON DATOS PROVENIENTES DE IMÁGENES DE SATÉLITES

La contabilidad nacional es un tema relevante para todos los países, ya que la información que surge de ella brinda una base sólida para la toma de decisiones relacionadas con la política económica. Prácticamente todos los países generan indicadores de actividad económica según los lineamientos establecidos por diversos organismos multilaterales; ello no impide que existan divergencias entre las cifras reportadas de manera oficial y las que surgen al aplicar métodos alternativos. En particular, una variable que se usa como referencia fundamental para referirse a la actividad económica de un país es el PIB, que es pieza fundamental para el análisis macroeconómico, por lo cual es importante calcularlo correctamente. Una vez contabilizada la producción del país, se puede usar el indicador para establecer comparaciones entre distintas economías y distinguir así las diferencias entre países

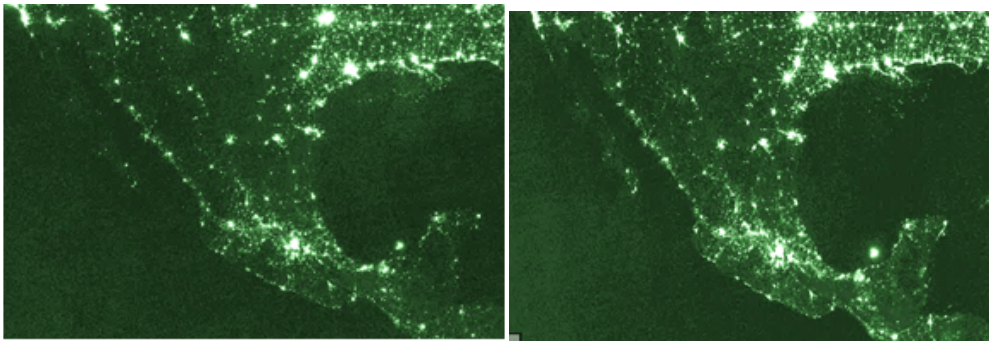
o entre distintas regiones dentro de un país. Tales comparaciones son una herramienta clave para el análisis del crecimiento económico del país. El uso de imágenes de satélites para medir el crecimiento del PIB, a través de la intensidad de las luces, es una herramienta que mejora dicha medición, sobre todo para lograr credibilidad en las comparaciones que se hacen, ya que con este enfoque se analiza en esencia la capacidad instalada del país.

El departamento meteorológico satelital de la fuerza aérea de Estados Unidos cuantifica la intensidad de las luces mediante un sistema operativo de escaneo que monitorea imágenes vía satélite, aproximadamente catorce veces al día y cuenta con una base de datos desde el año 1992. La Administración Nacional Oceanográfica y Atmosférica junto con el Centro Nacional de Geofísica de Estados Unidos, manipulan la base de datos satelitales para hacerla amigable al público en general. El proceso de filtrado de la información que realizan estas agencias remueve puntos que podrían sesgar la información, como las auroras polares y los incendios forestales, para capturar exclusivamente la intensidad de la luz artificial. Una vez removidos los sesgos potenciales, se promedian los datos sobre todas las órbitas de cada satélite por año, con lo que se obtiene información agrupada por satélite y por año. La manera como se reporta la información para cada celda dentro de la matriz que clasifica por latitud y longitud, es a través de un número digital (ND) que va de 0 a 63, donde 0 denota ausencia de luz y 63 es la máxima intensidad posible de luz. Se debe mencionar que la comparación del número *per se* puede cambiar a lo largo de los años debido a la obsolescencia de los satélites. Un claro ejemplo de esto ocurrió en 2002 con el satélite llamado F15, cuando la obsolescencia del satélite se evidenció como una caída en la actividad de las luces para todos los países en ese año. Por ello, en la base de datos de Henderson, Storeygard y Weil (2012) se excluyeron los datos de algunos países para el año 2002.

Las imágenes sobre la intensidad de la luz, captadas durante la noche, son un campo muy interesante que no se ha estudiado con profundidad en México. El presente trabajo busca abrir la puerta para el empleo más amplio de esta medición, para realizar análisis de la macroeconomía desde una perspectiva diferente. Aunque existen otras aplicaciones de

Figura 1.

Luminosidad nocturna en México en 1992 (izquierda) y 2008 (derecha)



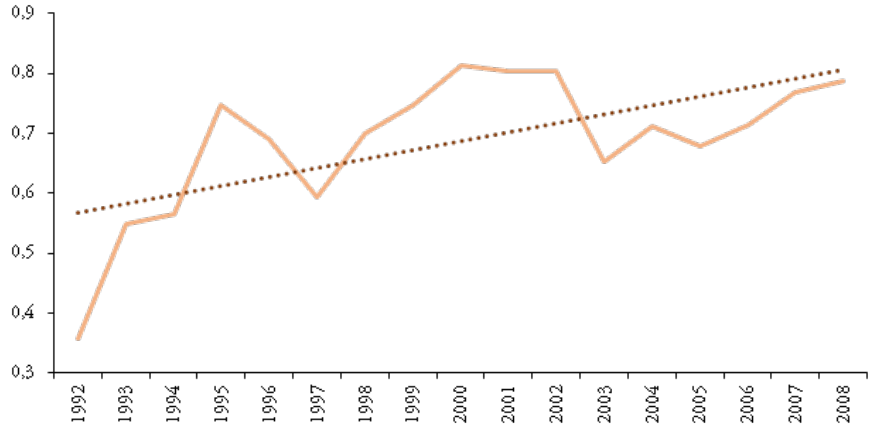
Fuente: Oficina Nacional de Administración Oceánica y Atmosférica de Estados Unidos. https://sos.noaa.gov/datasets/nighttime-lights-comparison-1992-2000-and-2008/

esta fuente de información, como lo señalan Nordhaus y Chen (2015), al usar datos sobre la intensidad de la luz, los analistas macroeconómicos podrían estimar el PIB, incluso mejor que el INEGI, según se aprecia con la metodología que aquí se presenta. Para tener una idea de la importancia de las luces nocturnas, en la figura 1 se muestran las imágenes para México en 1992 y 2008, donde es claro que hay mayor luminosidad en algunas regiones del país, en el año más reciente.

El valor de los números digitales para México que se muestra en la figura 2 para la serie de ND en logaritmos (naturales), marca una tendencia ascendente. Se aprecia también la caída de la economía durante la crisis de 1995, aunque con un rezago, que podría deberse al hecho de que cuando cae la actividad económica, las empresas recortan de manera inicial los costos variables y luego los fijos (dentro de los cuales se sitúa la energía eléctrica).

Figura 2.

Trayectoria de ln(ND) para México en el periodo 1992-2008



Fuente: Elaboración propia.

La tendencia creciente que presenta la serie de ND no ha sido lineal y los cambios que se han presentado año con año parecen coincidir con los del PIB; sin embargo, es importante subrayar que la caída que se origina en 2002 no coincide con lo ocurrido con el PIB para ese año o el siguiente.

2.1. Estimación del crecimiento del PIB verdadero para un país individual

La estimación que se propone, parte de la existencia de dos series de datos, del PIB y de la intensidad de las luces, para los años t = 1, …, N en un país específico. Se supone entonces un modelo de señal más ruido para ligar los crecimientos del PIB verdadero (Y_t) y oficial (Z_t), medidos como diferencias logarítmicas, *i.e,* $Dy_t = y_t - y_{t-1}$, con $y_t = \ln(Y_t)$ y $Dz_t = z_t - z_{t-1}$, con $z_t = \ln(Z_t)$. Así se obtiene:

$$Dz_t = Dy_t + \eta_t \qquad \text{para } t = 2,...,N, \qquad [1]$$

donde η_t es el ruido que oscurece la señal Dy_t, con $\eta_2, ..., \eta_N$ una sucesión de errores aleatorios tales que $Cov(\eta_t, \eta_{t'}) = 0$ si $t \neq t'$, con $E(\eta_t) = 0$ y $Var(\eta_t) = \sigma^2_\eta$. En Guerrero y Mendoza (2019) se supone –y se brinda justificación para ello– que $Cov(Dy_t, \eta_t) = 0$ y $Var(Dy_t) = \sigma^2_{Dy}$, donde los errores son estacionarios y no-correlacionados, lo que implica que la discrepancia entre el crecimiento oficial y el verdadero no se acarrea de un periodo al siguiente. Por otro lado, se supone una relación de elasticidad constante entre las luces nocturnas observadas (X_t) y el ingreso del país, de donde surge la expresión $X_t = KY_t^\beta$, con K una constante positiva y β la elasticidad de las luces respecto al PIB. Por lo cual,

$$Dx_t = \beta Dy_t + \varepsilon_t \qquad \text{para} \qquad t = 2,...,N, \qquad [2]$$

donde $\varepsilon_2,..., \varepsilon_N$ son errores aleatorios no-correlacionados, con $E(\varepsilon_t) = 0$, $Var(\varepsilon_t) = \sigma^2_\varepsilon$, $Cov(Dy_t, \varepsilon_t) = 0$ y $Cov(\eta_t, \varepsilon_t) = 0$ para t = 2,....N.

Al definir los vectores, de dimensión N-1, con los crecimientos de los PIB y de las luces, $D\mathbf{y} = (Dy_2, ..., Dy_N)'$, $D\mathbf{z} = (Dz_2, ..., Dz_N)'$ y $D\mathbf{x} = (Dx_2, ..., Dx_N)'$, se puede escribir a [1] y [2] como un sistema de ecuaciones lineales. Entonces, si se supone que los parámetros β, σ^2_ε y $\alpha = \sigma^2_\eta / \sigma^2_\varepsilon$ son conocidos, se puede usar el método de mínimos cuadrados generalizados (MCG) para obtener el Mejor Estimador Lineal e Insesgado de D**y**, dado por:

$$\widehat{D\mathbf{y}} = \lambda D\mathbf{z} + (1 - \lambda)\widetilde{D\mathbf{z}} \qquad [3]$$

con $\widetilde{D\mathbf{z}} = \beta^{-1} D\mathbf{x}$ y $\lambda = \frac{\alpha^{-1}}{\alpha^{-1}+\beta^2} \in (0, 1)$. Un procedimiento factible para obtener el estimador $\widehat{D\mathbf{y}}$ se propone más adelante. Además, la matriz de varianza-covarianza del error de predicción se obtiene como

$$Var(\widehat{D\mathbf{y}} - D\mathbf{y}) = \sigma^2_\varepsilon(\alpha^{-1} + \beta^2)^{-1} I_{N-1}. \qquad [4]$$

con

$$\widehat{\sigma}^2_\varepsilon = \frac{\hat{\lambda}\,\hat{\beta}^2}{N-3}\left(D\mathbf{z} - \widetilde{D\mathbf{z}}\right)'\left(D\mathbf{z} - \widetilde{D\mathbf{z}}\right). \qquad [5]$$

Es interesante notar que σ^2_ε tiende a 0 conforme $\hat{\lambda} \to 0$ (o, de forma equivalente, $\hat{\alpha}$ tiende a ∞. Este hecho se usa más adelante para estimar α.

Con el fin de estimar β se usan las ecuaciones [1] y [2], de manera que:

$$Dz_t = \beta^{-1}(Dx_t - \varepsilon_t) + \eta_t = \beta_1 Dx_t + \gamma_t, \qquad [6]$$

donde:

$$\beta_1 = \beta^{-1} \qquad y \qquad \gamma_t = \eta_t - \beta^{-1}\varepsilon_t, \qquad [7]$$

con $E(\gamma_t) = 0$ y $Var(\gamma_t) = \sigma^2_\varepsilon(\alpha + \beta^{-2})$. Luego, como:

$$\text{Cov}(\text{Dx}_t, \gamma_t) = \text{Cov}(\text{Dx}_t, \eta_t - \beta^{-1}\,\varepsilon_t) = -\,\beta^{-1}\,\sigma_\varepsilon^2 \qquad [8]$$

El estimador de mínimos cuadrados ordinarios (MCO), $\hat{\beta}_{1,\text{MCO}} = \widehat{\text{Cov}}(\text{Dx}_t, \text{Dz}_t)/\widehat{\text{Var}}(\text{Dx}_t)$, involucra a $\widehat{\text{Cov}}(\text{Dx}_t, \text{Dz}_t) = \beta_1\widehat{\text{Var}}(\text{Dx}_t) - \beta_1\hat{\sigma}_\varepsilon^2$ Por lo que tiene sesgo, o sea,

$$\text{E}\left(\hat{\beta}_{1,\text{OLS}}\right) = \beta_1\text{E}\left(\frac{\widehat{\text{Var}}(\text{Dx}_t) - \hat{\sigma}_\varepsilon^2}{\widehat{\text{Var}}(\text{Dx}_t)}\right) \neq \beta_1. \qquad [9]$$

Para corregir este problema, Guerrero y Mendoza (2019) muestran que se podría usar el cociente de crecimientos promedio del PIB oficial y las luces nocturnas. Sin embargo, dicho estimador únicamente utiliza el crecimiento de largo plazo de ambas variables involucradas, por lo cual no se considera confiable para estimar el crecimiento anual del PIB. Una alternativa es utilizar la mediana del crecimiento de las luces que, adicionalmente, brinda protección contra la influencia de mediciones satelitales anómalas, que se sabe pueden ocurrir como se mencionó al inicio de esta sección. Al hacer esto surge el estimador insesgado

$$\hat{\beta}_1 = \left\{ \begin{array}{ll} \frac{\text{Dz}\{\text{Dx}_{(m+1)}\}}{\text{Dx}_{(m+1)}} & \text{si } N-1 = 2m+1 \\ \frac{\text{Dz}\{\text{Dx}_{(m)}\}+\text{Dz}\{\text{Dx}_{(m+1)}\}}{\text{Dx}_{(m)}+\text{Dx}_{(m+1)}} & \text{si } N-1 = 2m \end{array} \right\} \qquad [10]$$

donde $\text{Dz}\{\text{Dx}_{(t)}\}$ denota el crecimiento del PIB oficial correspondiente al crecimiento de las luces en el momento t.

Con el estimador $\hat{\beta} = \hat{\beta}_1^{-1}$ se obtiene la estimación preliminar insesgada

$$\widetilde{\text{Dz}}_t = \hat{\beta}_1\text{Dx}_t \quad \text{para } t = 2, \ldots, N, \qquad [11]$$

que se combina con las cifras de crecimiento oficial del PIB mediante la expresión [3]. Para ello, falta estimar el parámetro α, lo cual se hace a partir de la expresión $\lambda = \alpha^{-1}/(\alpha^{-1} + \beta^2)$, de manera que,

$$\hat{\alpha} = \left(1-\lambda\right)/\left(\hat{\beta}^2\lambda\right) \qquad [12]$$

con $\lambda \in (0,1)$ elegida de manera apropiada. Por ello se propone analizar la sensibilidad de los resultados ante diferentes valores de λ. Esto se logra al considerar intervalos de ±2 errores estándar para el verdadero crecimiento del PIB y elegir el valor de λ como el menor valor que hace válida la afirmación probabilística del Teorema de Tchebysheff. Con esta propuesta se obtiene el menor valor de varianza estimada $\hat{\sigma}_\varepsilon^2$ según se hizo notar después de la ecuación [5]. De esta forma, para t = 2, ..., N, debe cumplirse que:

$$\text{Pr}\left[\left|\widehat{\text{Dy}}_t - \text{Dy}_t\right| \geq 2\hat{\sigma}_\varepsilon\left(\hat{\alpha}^{-1} + \hat{\beta}^2\right)^{-1/2}\right] \leq 1/4, \qquad [13]$$

a fin de considerar que una serie de tiempo de datos oficiales del crecimiento del PIB $\{\text{Dz}_t\}$ sea suficientemente cercana a $\{\text{Dy}_t\}$ si, a lo más, ¼ de las observaciones de $\{\text{Dz}_t\}$ están fuera de los intervalos:

$$\widehat{Dy}_t \pm 2\sqrt{\hat{\sigma}_\varepsilon^2\left(\hat{\alpha}^{-1}+\hat{\beta}^2\right)^{-1}} \quad \text{para} \quad t = 2, \ldots N. \qquad [14]$$

Por último, la varianza del error de predicción está dada por:

$$\mathrm{Var}\left(\widehat{Dy}_t - Dy_t\right) = \hat{\sigma}_\varepsilon^2\left(\hat{\alpha}^{-1}+\hat{\beta}^2\right)^{-1} = \frac{\hat{\lambda}\,\hat{\beta}^2}{(N-3)\left(\hat{\alpha}^{-1}+\hat{\beta}^2\right)}\left(\mathbf{Dz}-\widetilde{\mathbf{Dz}}\right)'\left(\mathbf{Dz}-\widetilde{\mathbf{Dz}}\right) \qquad [15]$$

así que, con $\hat{\alpha}^{-1} = \hat{\lambda}\,\hat{\beta}^2/\left(1-\hat{\lambda}\right)$, se obtiene:

$$\mathrm{Var}\left(\widehat{Dy}_t - Dy_t\right) = \frac{\hat{\lambda}\left(1-\hat{\lambda}\right)}{N-3}\left(\mathbf{Dz}-\widetilde{\mathbf{Dz}}\right)'\left(\mathbf{Dz}-\widetilde{\mathbf{Dz}}\right), \qquad [16]$$

y $\mathrm{Var}\left(\widehat{Dy}_t - Dy_t\right)$ crece a $\frac{0.25}{N-3}\left(\mathbf{Dz}-\widetilde{\mathbf{Dz}}\right)'\left(\mathbf{Dz}-\widetilde{\mathbf{Dz}}\right)$ conforme $\hat{\lambda} \longrightarrow 1/2$ y decrece a 0 si $\hat{\lambda} \longrightarrow 0$ o $\hat{\lambda} \longrightarrow 1$. Es por ello que la máxima incertidumbre ocurre cuando $\hat{\lambda} \longrightarrow 1/2$ y corresponde al caso de igual ponderación para los dos crecimientos económicos –el medido con el satélite y el de las cifras oficiales–. Adicionalmente, debe notarse que el estimador del crecimiento verdadero del PIB surgió sin haber supuesto alguna distribución de probabilidad, por lo cual no es factible establecer inferencias del tipo de intervalos de predicción para el crecimiento anual del PIB.

2.2. Aplicación al crecimiento del PIB de México

La aplicación empírica para México hace uso de los datos de los años 1992 a 2008. En principio, se obtuvo la mediana del crecimiento de las luces durante dicho periodo, Med(Dx), calculada como el promedio de los crecimientos observados en 2006 y 2008, de donde se calculó $\hat{\beta}_1 = 1.2377$ de acuerdo con [10], lo que implica una elasticidad de las luces respecto al ingreso de $\hat{\beta} = 0.808$. Para validar el supuesto de no-correlación serial, se estimó el coeficiente autorregresivo de orden 1 para los residuos (0.24 con error estándar de 0.29 y valor-p 0.42), y se concluyó que el coeficiente no es significativamente diferente de cero, con ello se brindó apoyo empírico al supuesto. En la tabla 1 se presentan los valores de $\hat{\alpha}$ y de la varianza estimada del error σ_ε^2, para distintos valores de λ y haciendo uso de la elasticidad estimada. Los resultados muestran el grado de sensibilidad del estimador $\hat{\alpha}$ ante distintos valores de λ.

Tabla 1.

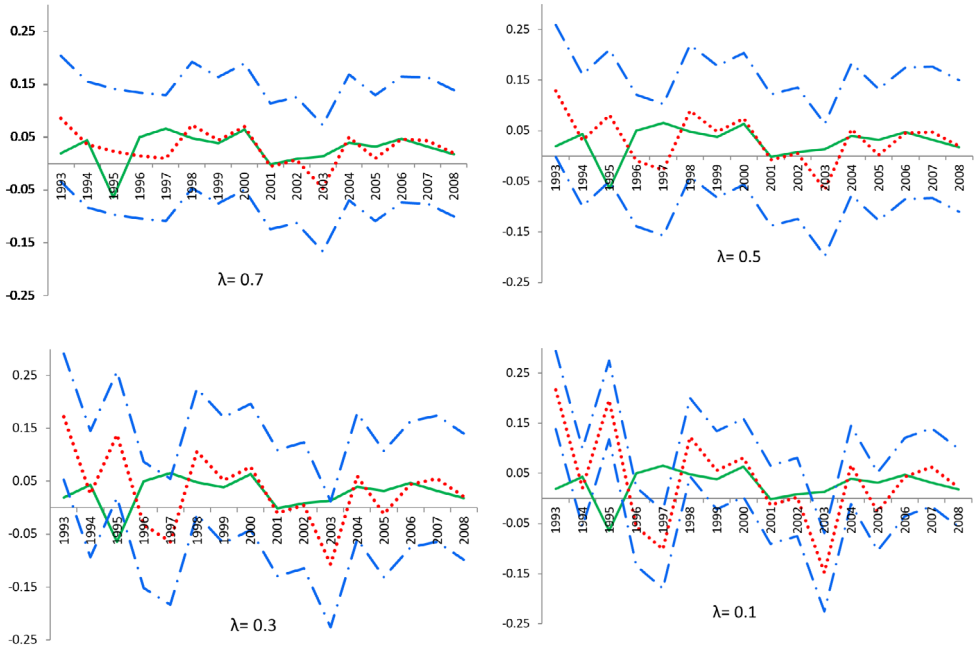
Resultados de la estimación para México con $\hat{\beta}$ = 0.808 y diferentes valores de λ

λ	*0.9*	*0.8*	*0.7*	*0.6*	*0.5*	*0.4*	*0.3*	*0.2*	*0.1*
$\hat{\alpha}$	0.170	0.383	0.656	1.021	1.532	2.298	3.574	6.127	13.786
$\hat{\sigma}_\varepsilon^2$	0.010	0.009	0.008	0.007	0.006	0.004	0.003	0.002	0.001

En la figura 3 se muestran los intervalos de ±2 errores estándar del tipo [14] que se obtuvieron con cuatro valores distintos de λ. Se aprecia que con λ = 0.7 no existen observaciones fuera de la banda, con λ = 0.5 una observación sale de la banda, con λ = 0.3 hay cuatro observaciones fuera y con λ = 0.1 hay cinco.

Figura 3.

Intervalos de ±2 errores estándar para λ = 0.7, 0.5, 0.3 y 0.1 Crecimiento oficial (línea verde) y estimado (línea roja), con sus cotas (guiones)



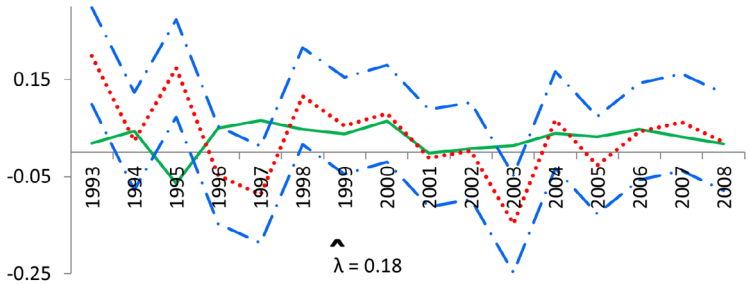
Fuente: Elaboración propia.

La figura 4 permite ver cuatro observaciones fuera de la banda, que es el número buscado, o sea, un valor entero no mayor que (N-1)/4 = 4. En este caso, $\hat{\lambda}$= 0.18 es el valor más pequeño que produce tal resultado, pues con $\hat{\lambda}$= 0.17 hay cinco valores fuera de la banda, de manera que se elige $\hat{\lambda}$= 0.18 como la estimación apropiada. La estimación del promedio del crecimiento verdadero del PIB de la figura 4, para los años 1993-2008 es de 3.27 %, mientras que el promedio de crecimiento oficial fue de 2.82 %, lo cual conduce a concluir que el PIB oficial produce una subestimación del 0.45 % anual.

El resultado obtenido debe validarse respecto a los supuestos del modelo que lo produjo. En lo que toca al supuesto implícito de estacionariedad de los residuos $\left\{Dz_t - \widetilde{Dz}_t\right\}$ mostrados en la figura 5, se aplicó la prueba de raíz unitaria de Phillips-Perron, que es estrictamente válida para muestras grandes y en este caso, con sólo 16 datos, se debe considerar como un

Figura 4.

Intervalos de ±2 errores estándar para el verdadero PIB de México, con $\hat{\lambda}$ = 0.18. Crecimiento oficial (línea verde) y estimado (línea roja), con sus cotas (guiones)

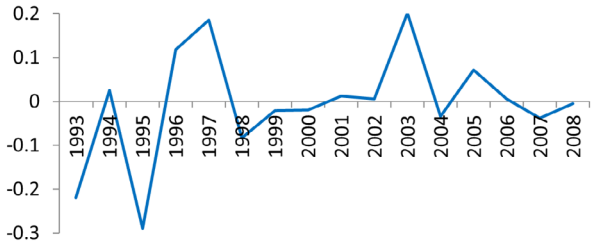


Fuente: Elaboración propia.

mero indicador de posible no-estacionariedad. Así se obtuvieron los estadísticos calculados: con 0 retrasos, -5.31, con 1 retraso, -5.34, y con 2 retrasos, -5.67, todos ellos conducentes al rechazo de la hipótesis nula de raíz unitaria al nivel de significancia del 1 %.

Figura 5.

Serie de residuos para el modelo del verdadero PIB de México



Fuente: Elaboración propia.

Por último, como también se supuso que el error en la expresión [6] tiene media 0, se usó la prueba del rango-con-signo de Wilcoxon de dos lados, para datos apareados $\left(Dz_t, \widetilde{Dz_t}\right)$ y el resultado fue que la suma de los rangos positivos es 68, mientras que el valor crítico al 10 % es 37 para n = N-1 = 16 parejas de observaciones. En consecuencia, no hay evidencia en contra de la hipótesis nula de que ambas series tienen la misma localización, lo que conduce a concluir que la diferencia no es significativa, incluso al nivel del 10 %. Por lo tanto, Guerrero y Mendoza (2019) concluyeron que los supuestos subyacentes en el procedimiento estadístico eran válidos en los datos.

3. RETROPOLACIÓN DE LAS SERIES DE CUENTAS NACIONALES COMBINANDO FUENTES DIVERSAS

Las series de tiempo que se generan de manera oficial y que ponen a disposición del público las AOE, como sucede con el INEGI, no siempre tienen la longitud, ni la frecuencia deseada de observación, así como tampoco la cobertura geográfica o sectorial requerida. Además, los cambios de año base producen desarticulaciones entre los datos de las series de tiempo, que son ocasionadas por cambios en la estructura de las actividades económicas para distintos años base. Esto implica que los indicadores económicos más importantes que se generan de manera oficial deben someterse a procesos de revisión, desde el punto de vista de la clasificación económica de actividades. Por ello se deben usar herramientas estadísticas para estimar valores pasados de las series y que sirvan para desagregar datos originales.

El problema que se trata en esta sección consiste en compatibilizar y homogeneizar las distintas bases de datos disponibles, tanto en formato electrónico como en documentos impresos en papel. La compatibilización de cifras se debe cumplir en los diversos ámbitos en los que se presenta la información: (i) por cobertura geográfica, de forma que el nivel estatal –de los 32 estados del país– sea compatible con el nacional; (ii) por cobertura sectorial, para que los sectores –y en algunos casos incluso las ramas de actividad económica– sean compatibles con las tres grandes actividades (GA) económicas, o sea, GA 1 las actividades primarias, GA 2 actividades secundarias y GA 3 actividades terciarias; (iii) por cobertura temporal, para que las cifras a nivel trimestral sean compatibles con las cifras anuales; y (iv) homogénea en lo toca al año base, que debe ser el mismo para todo el periodo de análisis (2013 en este caso). Adicionalmente, dicho periodo debe ser de la mayor longitud posible, en el caso del INEGI este periodo va de 1980 a 2016, de acuerdo con las series que se usan como insumos, sin embargo, en este documento solamente se muestra la aplicación que cubre el periodo de 1993 a 2016 –el lector interesado puede encontrar la aplicación extendida hasta 1980 en el documento de investigación de Guerrero y Corona (2017)–.

Al inicio de este estudio la situación era la siguiente, existían tres bases de datos a nivel estatal: 1) con cifras anuales del PIB (en esta sección, por PIB se entenderá PIB real), que cubrían el periodo 1993-2006, con distinta clasificación de actividades económicas que la del actual Sistema de Clasificación Industrial de América del Norte (SCIAN) y con año base 1993; 2) con datos del Indicador Trimestral de la Actividad Económica Estatal (ITAEE), de 2003 a 2015, clasificada a nivel de GA, con año base 2008; y 3) con datos del PIB anual, clasificado por sector de actividad, para el periodo 2003-2015, también con año base 2008. Con estos datos se puede generar una base con datos a nivel estatal, trimestral, clasificados por GA, para los años 1993-2015, con año base 2008. Dicha base de datos se complementa con dos bases de datos disponibles a nivel nacional –sin desglose estatal–: 4) trimestral y clasificada por subsectores para 1993-2015, con año base 2008; 5) base de datos clasificada según la clasificación de actividades previa al SCIAN, también trimestral, para 1980-2015 y expresada con año base 2008. Finalmente, se contaba con dos bases de datos más, ambas con año base 2013 y con datos hasta 2016: 6) la del ITAEE, clasificado por GA, para 2003-2016; y 7) la del PIB nacional clasificado por subsectores, para 1993-2016.

Las bases de datos resultantes de las distintas fases del proyecto fueron compatibilizadas entre sí mediante la aplicación de diferentes técnicas, que incluyen: (i) conversión de datos para cambiar de año base, incluyendo la nueva clasificación de actividades económicas; (ii) desagregación temporal y contemporánea, para generar datos con mayor desglose que los de las bases de datos originales, tanto en la dimensión temporal como en la contemporánea –la cual cubre las dimensiones geográfica y sectorial–; (iii) Retropolación restringida, para extender el rango de los datos observados hacia atrás en el tiempo, respecto a los que están disponibles en las bases de datos oficiales; y (iv) Reconciliación de cifras para que satisfagan estrictamente las relaciones contables que existen entre los datos estatales y los nacionales. Todas estas técnicas son del tipo de macrodatos, puesto que surgen de métodos que no requieren del uso de los microdatos que se registraron en su momento –y a los que el público no tiene acceso–, ni pretenden reconstruir los datos originales que se observaron en el pasado. Las diferentes técnicas dan lugar a distintas etapas para lograr la retropolación restringida global; en particular las técnicas (ii), (iii) y (iv) son óptimas en términos estadísticos, pues corresponden a aplicaciones de la regla de combinación de información que se presenta en Guerrero y Peña (2003).

3.1. Descripción de los procedimientos estadísticos

En este apartado se describen someramente los procedimientos estadísticos que permiten cumplir con los objetivos planteados. Para mayores detalles, se invita al lector a revisar los documentos de Guerrero y Corona (2017, 2018a, 2018b) donde se presentan los detalles de las técnicas usadas aquí.

Conversión

Los métodos para efectuar la conversión no pretenden reconstruir la base de datos original –la del año base antiguo–, ni generar la verdadera base de datos que tenga el nuevo año base. Esto es, se busca únicamente obtener una aproximación a lo que se pudo haber observado, pero se carece de una medida de la incertidumbre asociada, con la cual se pueda juzgar su validez. Es por ello que se recurre al juicio visual y a la justificación que brinda el *principio de preservación del movimiento* usado en el tema de desagregación de series y en el del ajuste a un valor de referencia, como se describe en Dagum y Cholette (2006). El método que aquí se sugiere usar se conoce como método proporcional y tiene dos vertientes, la primera anual y la segunda trimestral. En ambos casos, la sugerencia es aplicar la técnica al nivel con mayor desglose de actividades económicas que sea posible, con lo cual se podría pensar que se aproxima el valor que se produciría con el enfoque de microdatos. De hecho, en el manual de Eurostat (Roulin y Eidmann, 2007, p. 14) se indica que los coeficientes de conversión deben calcularse al nivel más detallado posible, que en el presente caso corresponde al nivel de sectores, o sea, a dos dígitos de la clasificación del SCIAN.

La conversión o empalme es equivalente a asignar a la serie de tiempo con año base más reciente, y para los valores donde no haya información, la variación porcentual anual de

la serie de tiempo con año base anterior, de tal forma que la serie de tiempo convertida (año base reciente) mantiene sus niveles y exhibe el movimiento de la serie de tiempo con año base anterior. Para más detalles, véase Parrot y McKenzie (2003). Finalmente, es de subrayar que este método de conversión es válido solamente para cifras expresadas a precios constantes. Cuando se requiera realizar una tarea semejante, pero con cifras a precios corrientes, se debe usar otro tipo de encadenamiento, como lo señalan Correa, Escandón, Luengo y Venegas (2003) o Hellberg (2010).

Desagregación univariada

La desagregación, al igual que la retropolación restringida que aquí se aplican, puede ser multivariada o univariada. Así, se postula un modelo en donde la señal, o serie por estimar, es la suma de una serie preliminar más un ruido que se supone se comporta como un proceso estacionario con media cero y se modela como un proceso autorregresivo y de promedios móviles (ARMA). Con estos supuestos, se utiliza la regla de combinación de Guerrero y Peña (2003) para series de tiempo multivariadas, que es consistente con el método multivariado de retropolación restringida propuesto por Guerrero y Nieto (1999) que se describe más adelante.

Se supone ahora que la serie por ser estimada admite el mismo modelo AR, aunque con distintas varianzas. Desde luego, la no estacionariedad se supone capturada por los elementos determinísticos del modelo, con ello Nieto (1998) dedujo un método que produce resultados óptimos en términos estadísticos, si se cumplen los supuestos del modelo. Este procedimiento produce resultados óptimos cuando el modelo AR para la serie preliminar es adecuado. Para ello se deben verificar los supuestos que fundamentan a dicho modelo. En particular, la verificación de estacionariedad se obtiene si las raíces de la ecuación característica del polinomio AR involucrado, están fuera del círculo unitario –en el plano complejo–. La no autocorrelación de los errores se verifica mediante el estadístico de Ljung-Box y si se rechaza el supuesto se debe modificar el modelo hasta lograr el no-rechazo. Por último, para lograr que la media de los residuos sea cero, se incluye una constante en el modelo, con lo cual también se evitan sesgos potenciales. De igual forma se tiene que validar el modelo para las discrepancias. Cuando la verificación brinda resultados favorables, se puede concluir la validez de la serie preliminar y de su respectivo modelo, con lo cual se deduce también la validez de los resultados desagregados.

Retropolación restringida

El planteamiento de Guerrero y Nieto (1999) para la desagregación temporal y contemporánea de series de tiempo múltiples se usa para la retropolación restringida, que difiere de la desagregación tan solo en la manera como se genera la estimación preliminar. Para obtener el estimador, se aplica un método bietápico del tipo de mínimos cuadrados generalizados (MCG) factibles. Finalmente, la verificación de los supuestos de los modelos VAR involucrados en la retropolación restringida se realiza como en la desagregación univariada. Así que se debe validar el modelo VAR que se usa para producir las retropolaciones irrestrictas, al igual que el modelo VAR que representa el comportamiento de las discrepancias entre retropola-

ciones restringidas e irrestrictas. En principio debe verificarse la estacionariedad, mediante el cálculo de las raíces de las matrices de los polinomios asociados con las ecuaciones determinantes de los modelos. Acto seguido se debe probar que no hay autocorrelación, con la prueba de Ljung-Box para el caso multivariado.

Reconciliación de cifras estatales y nacionales

La aplicación de los métodos anteriores produce la estimación lineal más eficiente que se puede lograr con las bases de datos estatales oficiales –disponibles en el INEGI–. Dicha estimación cumple con las restricciones contables requeridas para su credibilidad, no obstante, existe otra base de datos que no ha sido utilizada porque no contiene datos a nivel estatal, sino nacional. Ahora se usa esta base con el fin de no omitir el uso de fuentes de información oficial. Además, la base de datos estimada hasta este punto contiene datos a nivel estatal que en este caso serán considerados como preliminares. Estos datos se deben ajustar para que cada trimestre sume al total nacional de cada una de las GA, provenientes de la nueva base, los cuales constituyen el conjunto de restricciones contemporáneas por satisfacer de manera estricta, para que los resultados sean creíbles desde el punto de vista de la contabilidad nacional.

El procedimiento es aplicable a cada una de las GA y produce resultados compatibles contablemente entre los niveles estatal y nacional. Para lo anterior, se consideran las participaciones de los estados respecto su total obtenidas a través de sumas para cada uno de los trimestres usando las series de tiempo retropoladas. Dichas participaciones se restringen multiplicando su respectiva participación con el PIB total nacional obtenido a través de fuentes oficiales. La formalización de este procedimiento puede consultarse en Guerrero y Corona (2018b).

3.2. Aplicación numérica de la metodología estadística

A manera de ilustración de los resultados que se obtienen al aplicar los métodos previamente descritos, en este apartado se muestran los resultados que se obtuvieron para un solo estado, la Ciudad de México –así llamado a partir de enero de 2016 y antes denominado distrito federal– que fue seleccionado por tener la mayor participación en el PIB nacional. Los resultados para los otros 31 estados pueden verse en el documento de investigación de Guerrero y Corona (2017).

Conversión de año base 1993 a año base 2008

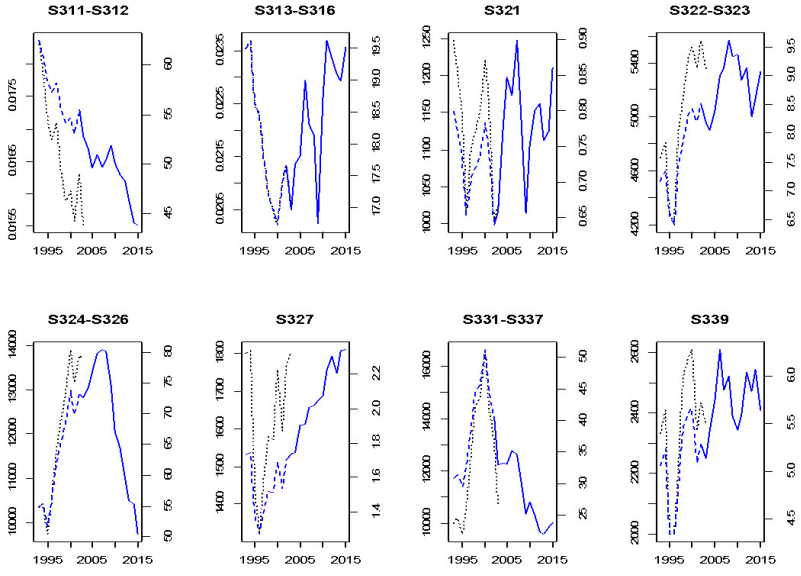
Los datos que surgieron como resultado de la aplicación de la fase de conversión están expresados en millardos de pesos a precios constantes de 2008 y estas son las unidades que se usan en las figuras y tablas de la parte numérica de este trabajo. Cabe recordar que la conversión surge de un procedimiento que no está basado en modelos estadísticos y por ello no requiere de validación empírica mediante pruebas estadísticas formales, sino de la

inspección visual de los resultados que produce. Por tal motivo, los resultados se presentan a través de gráficas que contienen las series obtenidas por conversión de los subsectores de industrias manufactureras, mientras que la segunda muestra el resultado de la conversión aplicada a nivel de sectores de cada GA. En estas gráficas aparece la serie observada con año base 2008 así como la serie observada con año base 2003 y la serie convertida al año base 2008. Finalmente en la tercera figura aparecen las series convertidas de las GA; en estas gráficas se muestran las series generadas de manera indirecta por agregación de los sectores a GA y las que se obtienen por conversión directa de las GA. La conversión más apropiada es la indirecta, cuyos resultados fueron considerados razonables por funcionarios del INEGI, a quienes se les presentaron resultados parciales de este trabajo, conforme se fueron produciendo.

La figura 6 presenta las gráficas de los subsectores manufactureros obtenidos por conversión de los datos de la Ciudad de México. La única comparación sensible que puede hacerse entre la serie con base 1993 y la serie con base 2008, es de su respectivo crecimiento, el cual se observa a través de la dirección de las series, sin tener en cuenta magnitudes, y que se aprecia razonable. Para conocer las claves de los subsectores, se recomienda consultar el artículo de Yuskavage (1990). Igual que con los resultados obtenidos al convertir las series de subsectores manufactureros de año base 1993 a 2008, en la figura 7 se observa que la conversión de cifras para los sectores de actividad en la Ciudad de México refleja la dirección, aunque no la magnitud, de los movimientos del año base 1993.

Figura 6.

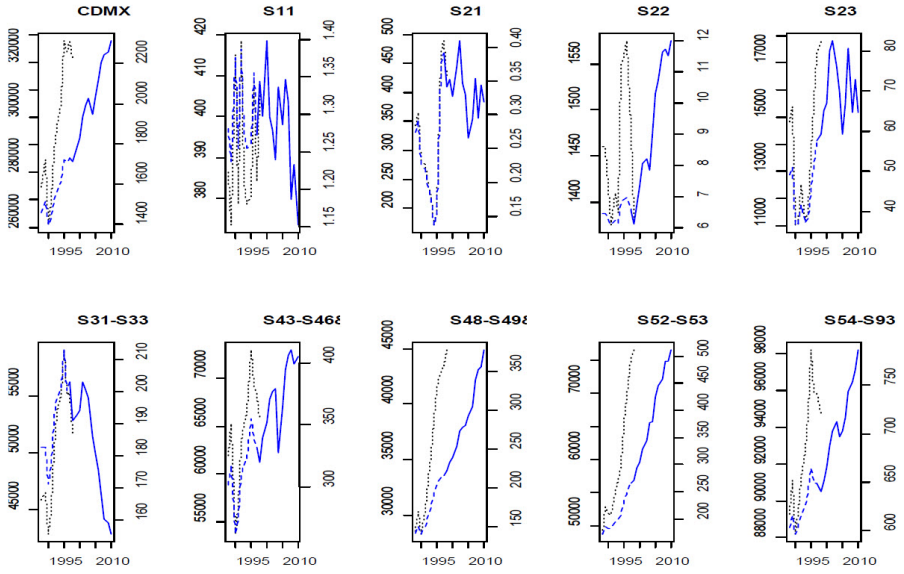
Conversión por subsectores de manufacturas. Serie observada (base 1993) en línea negra punteada, periodo de conversión en línea azul punteada y serie base 2008, en línea azul



Fuente: Elaboración propia.

Figura 7.

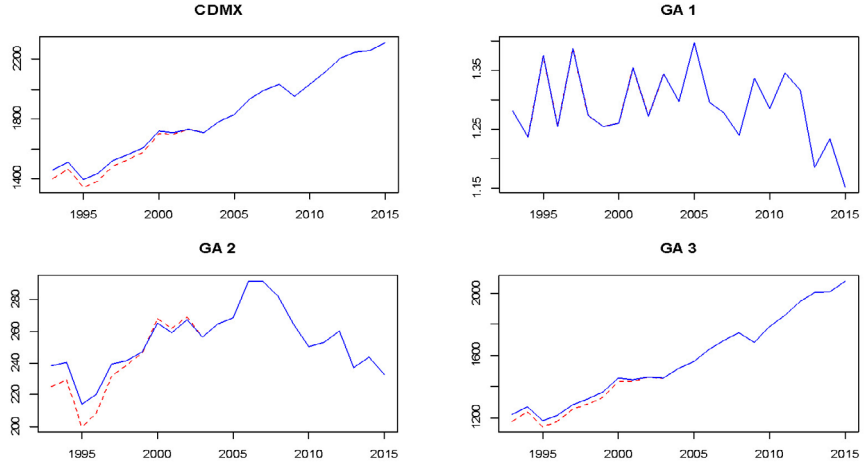
Conversión por sectores. Serie observada (base 1993) en línea negra punteada, periodo de conversión en línea azul punteada y serie observada base 2008, en línea azul continua



Fuente: Elaboración propia.

Figura 8.

Conversión de grandes actividades. Datos originales base 2008, serie convertida en forma directa –línea roja– e indirecta –línea azul–



Fuente: Elaboración propia.

En las gráficas de la figura 8 se nota cierta subestimación al usar el método directo en las GA 1 y 2, lo cual hace que el total de actividad económica de la Ciudad de México se subestime respecto al resultado obtenido en forma indirecta.

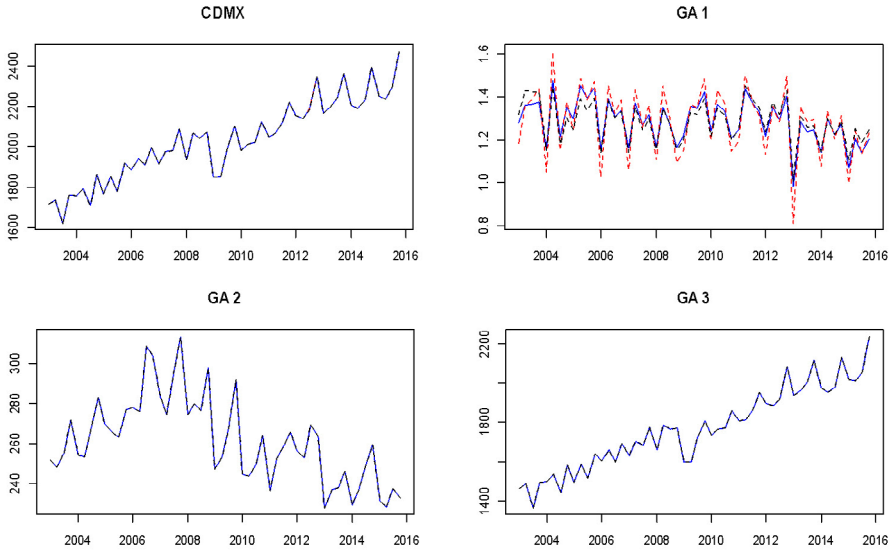
Desagregación temporal univariada de cada GA y del total

La desagregación a nivel de GA se realizó con los métodos de Nieto (1998) y de Denton-Cholette (véase Dagum y Cholette, 2006) con el fin de comparar empíricamente los resultados de los dos métodos en esta aplicación específica. Las gráficas de la figura 9 no muestran discrepancias notorias en las series desagregadas de GA 2 y GA 3 con uno y otro método, por lo que usar cualquiera de ellos es aparentemente indistinto, pero no ocurre lo mismo con GA 1, donde sí se observan discrepancias. Lo que se busca es que la serie desagregada cumpla con las restricciones temporales impuestas, o sea que al sumar los resultados de las tres GA se obtiene la desagregación estatal, lo cual está garantizado con los dos métodos empleados, y que el patrón observado en la serie desagregada sea lo más parecido posible a la serie preliminar, para que se cumpla la preservación del movimiento de esa serie. Esto último también ocurre, pero el criterio se satisface mejor con el método de Nieto, ya que en general la línea azul muestra mayor cercanía con la negra, que la línea roja.

Por otro lado, aunque los resultados se aprecien visualmente y numéricamente iguales, la validez de los mismos debe juzgarse por el cumplimiento de los supuestos que respaldan a los

Figura 9.

Desagregación por gran actividad. Preliminar –línea negra–, Denton-Cholette –línea roja– y Nieto –azul–. Millardos de pesos a precios constantes de 2008



Fuente: Elaboración propia.

métodos. En el caso del método de Denton-Cholette, la verificación de supuestos no es claro cómo realizarla, porque algunas decisiones se toman arbitrariamente y no se puede verificar su adecuación a la serie en estudio. En cambio, con el método de Nieto se usan criterios estadísticos y se pueden verificar los supuestos con los datos disponibles de cada serie. En la tabla 2 se muestran los resultados de la verificación del modelo AR para cada serie preliminar, en lo que toca a la estacionariedad de los errores del modelo y a la ausencia de autocorrelación residual, de manera que el modelo captura las regularidades empíricas de la serie respectiva. La estacionariedad es verificable con las raíces del polinomio AR, que deben estar fuera del círculo unitario y la no autocorrelación del error se verifica con la prueba de Ljung-Box. Otro supuesto deseable que se cumpla es la no-correlación cruzada entre residuos del modelo para la serie preliminar con las discrepancias entre serie preliminar y serie desagregada. De no cumplirse este último supuesto, lo que se debe hacer es buscar un mejor modelo para la serie preliminar o sustituir la serie preliminar. Sin embargo, si ya se verificó que el modelo para la serie preliminar se justifica estadísticamente, se debe buscar una serie preliminar alternativa, lo cual en este caso no es factible, porque únicamente existe la serie que proporciona la base de datos oficial.

En la tabla 2 se aprecia que se cumple la estacionariedad, pues todas las raíces de los modelos AR están fuera del círculo unitario. Asimismo, se cumple la no-autocorrelación del error y todos los coeficientes son significativamente distintos de cero, al compararlos con sus respectivos errores estándar. De esta forma, los modelos son razonablemente válidos –desde el punto de vista de la teoría estadística– y por ende los resultados obtenidos tienen soporte en los datos. La validación del método de Nieto se complementa al verificar ausencia de correlación entre residuos del modelo para las series preliminar y de discrepancias entre serie desagregada y preliminar: GA 1, Q = 7.43, valor-p = 0.96; GA 2, Q = 5.49, valor-p = 0.99; GA 3, Q = 7.48, valor-p = 0.96, que distan mucho de indicar significancia y brindan soporte empírico al método.

Tabla 2.

Validación de modelos AR para las series preliminares

GA 1	$\hat{W}_{1,t} = 1.32 + 0.34W_{1,t-2} - 0.14d_{1,t} + 0.05d_{2,t} - 0.02d_{3,t}$ (0.02) (0.14) (0.03) (0.02) (0.03) con $\hat{\sigma}^2_{W,1} = 0.005$, raíces: 1.72, -1.72 Q(16) = 16.20, valor-p = 0.44
GA 2	$\hat{W}_{2,t} = 271.48 + 0.78W_{2,t-1} - 24.21d_{1,t} - 22.16d_{2,t} - 10.67d_{3,t}$ (9.88) (0.07) (2.71) (3.06) (2.64) con $\hat{\sigma}^2_{W,2} = 104.8$, raíz: 1.15 Q(16) = 24.24, valor-p = 0.08
GA 3	$\hat{W}_{3,t} = 1870.66 + 0.84W_{3,t-1} + 0.56W_{3,t-4} - 0.41W_{3,t-5} - 98.30d_{1,t}$ $-84.16d_{2,t} \quad -108.20d_{3,t}$ (280.49) (0.09) (0.13) (0.14) (24.03) (27.23) (23.70) con $\hat{\sigma}^2_{W,3} = 2062$, raíces: 1.01, -1.17, 0.02-1.17i, 1.49, 0.02+1.71i Q(16) = 11.14, valor-p = 0.80

Retropolación restringida hasta 1993

Los datos desagregados se usan para construir modelos VAR trivariados que permitan generar pronósticos hacia atrás, del cuarto trimestre de 2002 al primer trimestre de 1993, con origen en el primer trimestre de 2003, para las tres GA simultáneamente. Primero se transforman los datos con el logaritmo natural de cada una de las series, para evitar que los pronósticos tomen valores negativos, aunque al transformar de esta forma también mejora la estabilidad de la varianza del error involucrado. Luego, los pronósticos en logaritmos se retransforman a la escala original de las GA al exponenciarlos. El orden del modelo VAR se elige con el criterio de Schwarz y se especifica la parte determinística del modelo, es decir, si debe llevar constante, tendencia, ambas o ninguna, y si se deben usar variables artificiales para capturar la estacionalidad de los datos. Para realizar esta labor se utilizó el paquete desarrollado por Pfaff (2008). Una vez estimado el modelo VAR, se le somete a una etapa de verificación de los supuestos de estacionariedad y de no-autocorrelación residual.

Después de verificar la validez del modelo, se puede confiar en los resultados que produce y así se deduce la credibilidad de los pronósticos que surgen de manera irrestricta. Sin embargo, no sólo se cuenta con la serie "observada" –la serie desagregada temporalmente–, sino con los datos de una serie anual que permite imponer restricciones temporales a los pronósticos de cada una de las GA, con lo cual se obtiene la retropolación restringida. En esta etapa no hay restricciones contemporáneas y es en la fase de reconciliación donde se incorporan las restricciones de que la suma de los valores de una GA, sea el total nacional de esa GA. Los resultados de la retropolación restringida multivariada se muestran a continuación.

La tabla 3 presenta los resultados de la estimación del modelo VAR de orden 1 con el que se obtienen los pronósticos irrestrictos. Este modelo contiene intercepto, sin tendencia, y

Tabla 3.

Resultados de la estimación del modelo VAR(1) usado para generar pronósticos irrestrictos
(Estimaciones preliminares)

Variable explicada	*Significancia de pruebas F para explicar variabilidad*				R^2 *ajustada*
	GA 1	*GA 2*	*GA 3*	*Estacionalidad*	
GA 1	0.46	0.24	0.01	0.00	0.95
GA 2	0.39	0.00	0.63	0.00	1.00
GA 3	0.48	0.69	0.31	0.00	1.00

Raíces de la ecuación determinante para estacionariedad: 0.75, 0.34, 0.08
Prueba de no-autocorrelación: Ji-cuadrada (135) = 142.12, valor-p = 0.32

variables artificiales para capturar la estacionalidad, las cuales son significativas para explicar a cada una de las tres GA. El efecto estacional es básicamente lo que hace que el coeficiente R^2 ajustado por grados de libertad sea tan alto en las tres ecuaciones, porque a GA 1 adicionalmente la explica GA 3, mientras que GA 2 es explicada por sí misma, y GA 3 no es explicada por ninguna de las otras GA, o sea que se comporta de manera prácticamente exógena. Por otro lado, los supuestos de estacionariedad y no-autocorrelación del error tienen soporte empírico y el modelo puede considerarse estadísticamente válido.

La prueba de si la serie múltiple de discrepancias entre valores estimados y preliminares se comporta como ruido blanco, brindó como resultado el valor de la Ji-Cuadrada = 362.61, con 135 grados de libertad, lo que produjo el valor-p = 0.00, por lo que se decidió aplicar la segunda etapa. En la segunda etapa se estimó un modelo VAR(1) para las discrepancias, cuyos resultados se muestran en la tabla 4.

Tabla 4.

Resultados de la estimación del modelo VAR(1) para la serie múltiple de discrepancias
(Segunda fase de MCG)

Discrepancia explicada	*Pruebas F de significancia para explicar variabilidad*				*R^2 ajustada*
	Disc (GA 1)	*Disc (GA 2)*	*Disc (GA 3)*	*Estac.*	
Disc (GA 1)	0.01	0.38	0.33	0.44	0.48
Disc (GA 2)	0.14	0.01	0.52	0.20	0.92
Disc (GA 3)	0.25	0.37	0.41	0.60	0.64

Raíces de la ecuación determinante para estacionariedad: 0.81, 0.81, 0.65
Prueba de no-autocorrelación: Ji-cuadrada (135) = 102.55, valor-p = 0.98

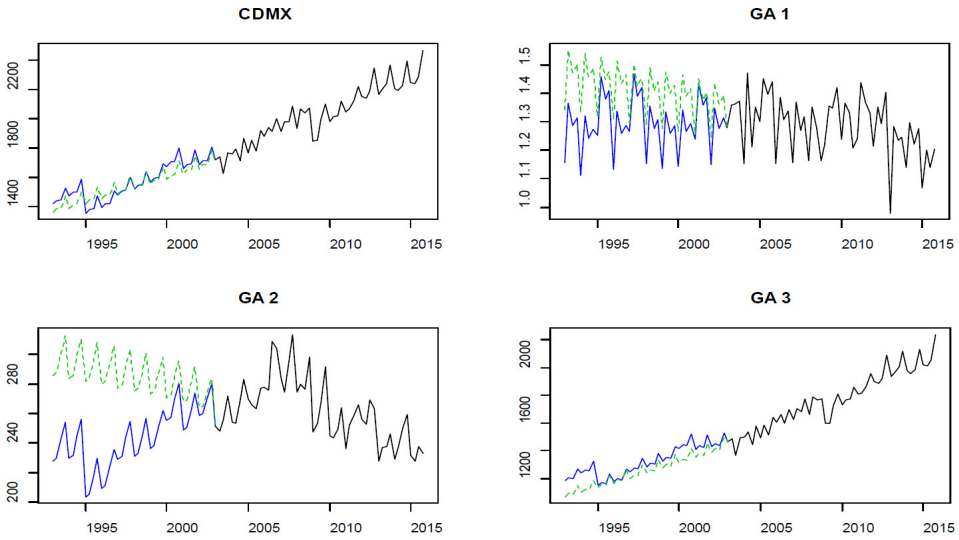
El modelo resumido en la tabla 4 cumple con los supuestos de estacionariedad y de no autocorrelación del error, por lo que se considera estadísticamente válido y se obtienen las matrices de varianzas para los errores de pronóstico un periodo hacia atrás, que se usan al generar las retropolaciones restringidas, o sea,

$$\hat{\Sigma}_e = \begin{pmatrix} 0.0032 & -0.0002 & 0.0001 \\ -0.0002 & 0.0002 & 0.0002 \\ 0.0001 & 0.0002 & 0.0007 \end{pmatrix} \text{ y } \hat{\Sigma}_\varepsilon = \begin{pmatrix} 0.0047 & 0 & 0 \\ 0 & 51.2412 & 0 \\ 0 & 0 & 1528.4680 \end{pmatrix}.$$

Los pronósticos irrestrictos y restringidos que surgen de esta aplicación se muestran en la figura 10, que permite apreciar el beneficio de incorporar las restricciones, puesto que

Figura 10.

Retropolación restringida multivariada: GA y total. Desagregada –línea negra–, pronósticos irrestrictos -línea verde- y restringidos –azul–
(Millardos de pesos de 2008)



Fuente: Elaboración propia.

los pronósticos irrestrictos simplemente marcan la tendencia y la estacionalidad de las GA, mientras que los pronósticos restringidos incorporan información acerca del nivel de las series.

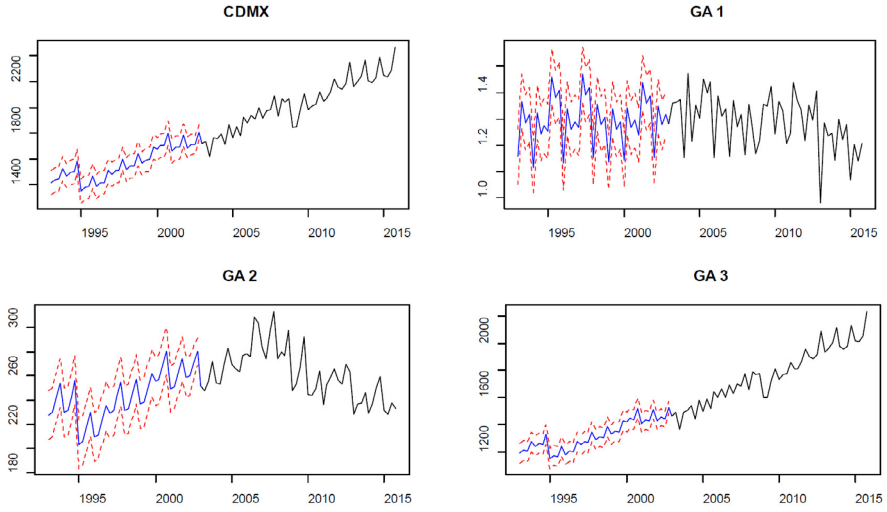
Reconciliación con las cifras trimestrales nacionales 1993-2002

La reconciliación de las bases de datos estimadas mediante retropolación restringida se aplica a cada una de las GA, para incorporar la información de la base de datos nacional. Como resultado se obtiene un ajuste de los datos retropolados que cumplen con la restricción de que la suma de valores de cada trimestre brinda el total de las tres GA del estado. De igual manera se obtiene el promedio de los valores trimestrales para cada una de las GA. Lo importante es que el patrón de las series obtenidas con la retropolación restringida sufre algunas modificaciones, como puede apreciarse en las gráficas de la figura 11.

Las gráficas muestran series más creíbles, ya que su patrón dinámico no es extraño en algún sentido que pudiera contradecir al conocimiento que se tiene del fenómeno económico subyacente, además de que satisfacen la restricción contable de que la suma de todos los estados es el total nacional. Por otro lado, cabe destacar el hecho de que la GA 1 se ve alterada muy poco –en términos relativos– al aplicar la reconciliación a la estimación que surge de

Figura 11.

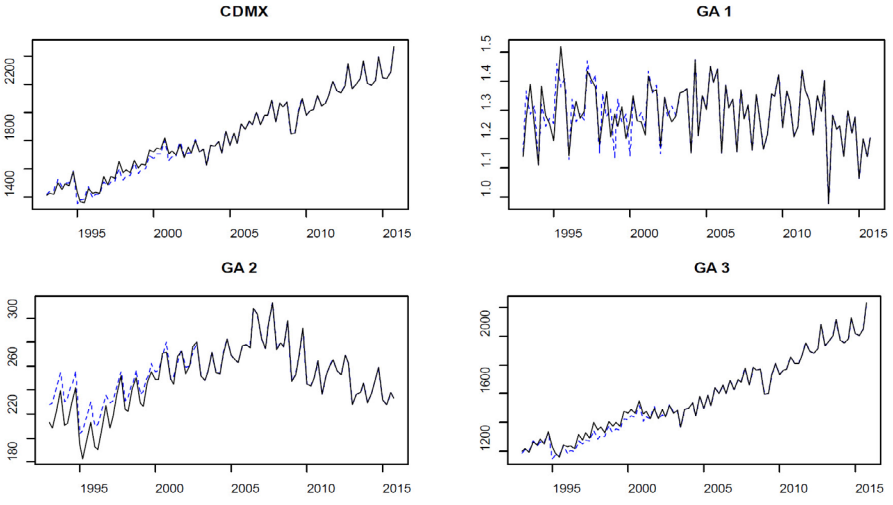
Retropolación restringida multivariada e intervalos de predicción: GA y total
(Millardos de pesos a precios constantes de 2008)



Fuente: Elaboración propia.

Figura 12.

Reconciliación de cifras estatales y nacionales: GAs y total. Retropolada –azul punteada–, reconciliada –negra–
(Millardos de pesos de 2008)



Fuente: Elaboración propia.

la retropolación restringida. En cambio, la GA 2 es la que –en términos relativos– se ve más afectada al reconciliar los datos de los estados con los datos de la base nacional. La tabla 5 enfatiza el hecho de que las cifras reconciliadas cumplen con las restricciones contables de que la suma de valores de las tres GA, brinda cada trimestre el PIB total del estado –excepto por redondeos en la presentación de las cifras–. Algo que debe resaltarse es que el promedio anual sigue la dinámica del PIB estatal anual convertido de año base 1993 a base 2008. Además, la suma de valores de las GA por estado es igual al total de la base de datos nacional.

Tabla 5.

Resultados de la reconciliación de cifras trimestrales por GA, para años seleccionados de 1993 a 2002
(Millardos de pesos de 2008)

Actividad	*Trimestre de 1993*					...	*Trimestre de 2002*				
	1	*2*	*3*	*4*	*Prom*		*1*	*2*	*3*	*4*	*Prom*
GA 1	1.2	1.3	1.4	1.3	1.3		1.2	1.4	1.3	1.3	1.3
GA 2	212.8	208.4	222.3	239.4	220.7		253.9	261.0	176.1	280.6	167.9
GA 3	1.195	1.215	1.192	1.260	1.216		1.425	1.490	1.436	1.516	1.467
Total	1.409	1.425	1.416	1.501	1.438		1.680	1.752	1.714	1.798	1.736

En las tablas 6, 7 y 8, se ilustra que el total nacional es la suma de los resultados estatales, de las respectivas GA, excepto para la GA 1. En consulta directa con los funcionarios encargados de calcular el PIB oficial en México, se mencionó que: "En los datos base 2008 coinciden 733 de las 734 actividades económicas con las que se integran los cálculos del PIB, a excepción de la agricultura, que en el ITAEE se mide por 'año calendario'; en el PIB por entidad federativa se mide por 'año agrícola' y en el PIB trimestral por 'año calendario', por ello no coincide el ITAEE con el PIB del estado" (Lourdes Mosqueda, 2017, comunicación personal).

Tabla 6.

Resultados de la reconciliación de cifras trimestrales de la GA 1, para la Ciudad de México y años seleccionados de 1993 a 2002
(Millardos de pesos de 2008)

Estado	*Trimestre de 1993*					...	*Trimestre de 2002*				
	1	*2*	*3*	*4*	*Prom*		*1*	*2*	*3*	*4*	*Prom*
CDMX	1.1	1.3	1.4	1.3	1.3		1.2	1.3	1.3	1.3	1.3
...											
Total[a]	281.8	289.3	265.9	331.2	292.0		336.2	359.1	297.4	389.7	345.6
Total[c]	285.4	292.9	269.4	334.8	295.6		339.7	362.7	300.9	393.3	349.2

Nota: Total[a] surge al sumar los valores trimestrales de todos los estados y corresponde al concepto de año agrícola, mientras que Total[c] es el valor del PIB trimestral nacional por año calendario.

El "año agrícola" que se usa en el subsector de agricultura se refiere al hecho de que la producción –desde la preparación de la tierra hasta el levantamiento de la cosecha– abarca más de un año calendario y el valor agregado de los cultivos se considera proporcional al costo de los insumos empleados en la producción, como son la fuerza de trabajo y los insumos intermedios, lo cual conduce a distribuir el valor total de la producción en proporción a los costos incurridos cada trimestre (véase INEGI, 2013). Por ello, las series reconciliadas de la GA 1 para cada estado se ajustan para satisfacer el criterio usado en el INEGI también durante el periodo 1993-2002 y así se obtienen valores referidos al "año agrícola".

Tabla 7.

Resultados de la reconciliación de cifras trimestrales de la GA 2, para la Ciudad de México y años seleccionados de 1993 a 2002
(Millardos de pesos de 2008)

Estado	*Trimestre de 1993*					...	*Trimestre de 2002*				
	1	*2*	*3*	*4*	*Prom*		*1*	*2*	*3*	*4*	*Prom*
CDMX	212.8	208.4	222.3	239.4	220.7		253.9	261.0	276.1	280.6	267.9
...											
Total	2.296	2.946	3.022	3.105	3.018		3.689	3.821	3.939	3.882	3.833

Tabla 7.

Resultados de la reconciliación de cifras trimestrales de la GA 3, para la Ciudad de México y años seleccionados de 1993 a 2002
(Millardos de pesos de 2008)

Estado	*Trimestre de 1993*					...	*Trimestre de 2002*				
	1	*2*	*3*	*4*	*Prom*		*1*	*2*	*3*	*4*	*Prom*
CDMX	1.195	1.215	1.192	1.260	1.216		1.425	1.490	1.436	1.516	1.467
...											
Total	4.457	4.608	4.562	4.715	4.611		5.654	5.881	5.735	5.910	5.795

4. CONCLUSIONES

Este trabajo presenta dos aplicaciones de metodologías que, a primera vista, podrían considerarse ajenas entre sí. No obstante, el lazo unificador de dichas técnicas se encuentra en las ideas propuestas dentro del contexto de big data, ya que en ambas aplicaciones se hace uso de algunas de las Vs características de este contexto. Adicionalmente, en las dos ilustraciones que se presentan, el interés radica en el PIB y la contabilidad nacional, que son fuentes de datos que proveen información macroeconómica de la mayor importancia para analizar el estado de la economía de un país y para elaborar políticas económicas.

La elaboración de la contabilidad nacional es un proceso que resulta complejo de cuantificar, sin embargo es posible tomar ventaja de los distintos frentes que se han desarrollado

de manera acelerada en los últimos años para incrementar la disponibilidad de la información que, como consecuencia, conlleva a mejorar las mediciones económicas. De manera particular, en este trabajo se emplea el desarrollo de grandes volúmenes de datos, así como una variedad alternativa de fuentes de información de los mismos, para proveer instrumentos adicionales en la medición de una de las variables fundamentales en el entorno macroeconómico: el PIB.

En los últimos años, el acelerado crecimiento de herramientas para procesar información, debido a los avances tecnológicos, ha permitido el acceso a fuentes de información que anteriormente se encontraban inaccesibles. Por esta razón, la búsqueda de métodos alternativos para cuantificar variables relacionadas con la actividad económica se ha vuelto una tarea recurrente entre los analistas. Las imágenes satelitales se pueden considerar como un elemento que ha comenzado a cobrar relevancia en el mundo, para medir la actividad económica. Algunas de las grandes ventajas del uso adecuado de tales imágenes es que permiten incorporar en la medición, factores que en ocasiones se pierden en la contabilidad nacional, tal es el caso de la economía no observada, así como la posibilidad de obtener un instrumento que permita llevar a cabo comparaciones transversales entre las distintas economías. En este sentido se evidenció que, para el caso de México, las cifras presentadas por el INEGI muestran un crecimiento menor a lo estimado con esta metodología alternativa.

Por otro lado, la retropolación de las series de cuentas nacionales, permite trazar un puente para homogeneizar la información, con lo cual se permite llevar a cabo análisis históricos, que son útiles para identificar patrones en las distintas series de tiempo que se estimaron en este ejemplo. Adicionalmente, esta metodología permite llevar a cabo vinculaciones en la información a través de distintos canales, como por ejemplo, las diversas coberturas que pueden alcanzarse en los ámbitos geográfico, sectorial e incluso espacial. Es por ello que el análisis desarrollado en el presente artículo, con base en la implementación de metodologías que confluyen con los criterios de big data para la economía mexicana, muestra la relevancia de la aplicación de tales metodologías para incentivar un mayor desarrollo en distintos frentes de medición económica. De esta manera, es posible transitar hacia una mejora generalizada en la imprescindible tarea de hacer más eficiente la medición de variables que conduzcan a un mejor diseño de políticas públicas.

Referencias

Braaksma, B. y Zeelenberg, K. (2015). "Re-make/Re-model": Should big data change the modelling paradigm in official statistics? *Statistical Journal of the IAOS,* Vol. 31, pp. 193-202.

Corona, F. y López-Pérez, J. (2020). Una evaluación econométrica de la retropolación de la actividad económica estatal de México. *Estudios Económicos,* Vol. 35(2), pp. 193-212.

Correa, S. V., Escandón, A. A., Luengo, P. R. y Venegas, M. J. (2003). Empalme de series anuales y trimestrales del PIB. *Economía Chilena,* Vol. 6 (1), pp. 77-86.

Dagum, E. B. y Cholette, P. A. (2006). *Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series.* Lecture Notes in Statistics, 186. New York: Springer-Verlag.

Ghosh, T., Sutton, P., Powell, R., Anderson, S. y Elvidge, Ch. D. (2009). Estimation of Mexico´s Informal Economy and Remittances Using Nighttime Imagery. *Remote Sensing,* Vol. 1(3), pp. 418-444.

Guerrero, V. M. y Corona, F. (2017). Retropolación óptima de series de tiempo de las tres grandes actividades económicas de México, por estado y trimestre, a precios constantes de 2013, para 1980-2016 (Documento de Investigación). *DGAI-DGIAI,* 17-02. Ciudad de México: INEGI.

Guerrero, V. M. y Corona, F. (2018a). Retropolating some relevant series of Mexico's System of National Accounts at constant prices: The case of Mexico City's GDP. *Statistica Neerlandica.* Special Issue Article, Vol. 72 , pp. 495-519.

Guerrero, V. M. y Corona, F. (2018b). Retropolación hasta 1980 del PIB trimestral de México por entidad federativa y gran actividad económica. *Realidad, Datos y Espacio. Revista Internacional de Estadística y Geografía,* Vol. 9(3), pp. 98-119.

Guerrero, V. M. y Mendoza, J. A. (2019). On measuring economic growth from outer space: a single country approach. *Empirical Economics,* Vol. 57(3), pp. 971-990.

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

Guerrero V. M. y Peña, D. (2003). Combining multiple time series predictors: A useful inferential procedure. *Journal of Statistical Planning and Inference,* Vol. 116, pp. 249-276.

Gupta, S., Mateu, J., Degbelo, A. y Pebesma, E. (2018) Quality of life, big data and the power of statistics. *Statistics and Probability Letters,* Vol. 136, pp. 101-104.

Hellberg, O. (2010). Backcasting Swedish Industrial Production. *Paper presented at the Workshop on Survey Sampling Theory and Methodology.* Vilnius (Lithuania).

Henderson, J. V., Storeygard, A. y Weil, D. N. (2012). Measuring Economic Growth from Outer Space. *American Economic Review,* Vol. 102(2), pp. 994-1028.

INEGI (2013) *Sistema de Cuentas Nacionales de México. Indicador Trimestral de la Actividad Económica Estatal.* Fuentes y metodologías. Aguascalientes (México): Instituto Nacional de Estadística y Geografía.

Nieto, F. H. (1998). Ex-post and Ex-ante Prediction of Unobserved Economic Time Series: A Case Study. *Journal of Forecasting,* Vol. 17(1), pp. 35-58.

Nordhaus W. y Chen, X. (2015). A sharper image? Estimates of the precision of nighttime lights as a proxy for economic statistics. *Journal of Economic Geography,* Vol. 15, pp. 217-246.

Parrot, F. y McKenzie, R. (2003). Linking factors for gross and seasonally adjusted series. *Note, Short Term Economic Statistics Division.* OECD.

Pfaff, B. (2008). VAR, SVAR and SVEC models: Implementation within R package vars. *Journal of Statistical Software,* Vol. 27(4), pp. 1–32.

Roulin, E. y Eidmann, U. (2007). *Back Casting Handbook.* Luxembourg: Eurostat.

Yuskavage, R. E. (2007). Converting historical industry time series data from SIC to NAICS. Paper prepared for the *Federal Committee on Statistical Methodology,* 2007 Research Conference, Arlington, VA (EE. UU).