Nowcasting of macroeconomic aggregates with non-contemporary indicators: the case of EMAE

Frank, Luis

DNMyP. Secretaría de Política Económica. Ministerio de Economía, Universidad de Buenos Aires. Facultad de Agronomía

25 August 2022

Predicción anticipada de agregados macroeconómicos con indicadores no contemporáneos: el caso de EMAE

Luis Frank *†

Resumen

El método utilizado por el BCRA para anticipar el PIB trimestral se extiende a la predicción del Estimador Mensual de Actividad Económica (EMAE), pero con variables difundidas en los primeros 15 a 20 días del mes que interesa predecir. Este procedimiento se extiende luego a situaciones en las que se dispone de más de un conjunto de variables tempranas, aunque no todas contemporáneas. Se deducen las fórmulas necesarias para la predección puntual y para hallar diferencias significativas entre predicciones consecutivas con distintos conjuntos de variables tempranas.

Palabras clave: EMAE, predicción anticipada, componentes principales.
JEL: C820

Abstract

The method used by BCRA to anticipate quarterly GDP is extended to forecast of the Monthly Estimator of Economic Activity (EMAE) but with variables released in the first 15 to 20 days of the EMAE data to be predicted. This procedure is then extended ti situations in which more than one set of early variables although not contemporary. The necessary formulas for a point forecast and to find statistically significant differences between consecutive predictions with different sets of early variables are also deduced.

Keywords: EMAE, nowcasting, principal components.
JEL: C820

1 Introducción

El Estimador Mensual de Actividad Económica (EMAE) elaborado por INDEC es "un indicador provisorio de la evolución del PIB [...] difundido con un rezago de 50 a 60 días de concluido el mes de referencia" [7] y es posiblemente el indicador de actividad económica de mayor impacto mediático de nuestro país. Se trata de un indicador que anticipa cifras (trimestrales) que, de no existir, se conocerían como mínimo 80 a 90 días después de concluido el trimestre de referencia. Si bien el EMAE constituye un verdadero avance en la disponibilidad de información respecto de los informes trimetsrales del PIB, en momentos críticos como la crisis financiera de abril de 2018, o la crisis sanitaria de marzo de 2020, el plazo de 50 a 60 días puede resultar

*DNMyP. Secretaría de Política Económica. Ministerio de Economía. Av. Hipólito Yrigoyen 250, C1086AAB. Buenos Aires, Argentina.
†Universidad de Buenos Aires. Facultad de Agronomía. Av. San Martín 4453, C1417DSE. Buenos Aires, Argentina.

demasiado extenso para actualizar proyecciones de crecimiento como las que rutinariamente realiza la DNMyP.[1]

La elaboración de indicadores anticipados de alta frecuencia a partir de series de baja frecuencia no es nueva en la bibliografía econométrico. El BCRA, por ejemplo, lleva años trabajando en un estimador anticipado del PIB (ver e.g. [4, 2, 3]) que reproduce métodos del vasto cuerpo de conocimiento conocido como *nowcasting*, es decir, la predicción del estado actual de una variable de baja frecuencia a partir de otras variables relacionadas de alta frecuencia. Éstos y otros trabajos, e.g. [8], se basan en la compilación de indicadores de alta frecuencia del más variado origen (desde indicadores de producción hasta indicadores de confianza del consumidor), la selección de un subconjunto de indicadores relevantes para predecir la variable de interés, y la extracción de este subconjunto de un conjunto aún más reducido de variables *latentes* (variables subyacente inobservables) para estimar el estado de la variable a predecir. El modelo que vincula a las variables latentes con la variable de interés suele ser un modelo dinámico que incorpora a la predicción información provista por el comportamiento pasado de esta última. El criterio que prima en la selección de las variables de alta frecuencia es, en general, el grado de asociación con la variable a predecir más que su anticipación, de manera que el resultado suele ser con frecuencia un estimador con buen nivel de ajuste, pero con modesta capacidad anticipatoria.

El informe que sigue constituye un ejercicio de estimación anticipada del EMAE. No se trata de un ejercicio de *nowcasting* en sentido estricto, ya que no se intenta estimar en alta frecuencia una serie de baja frecuencia, sino anticipar la propia serie de alta frecuencia con indicadores de la misma frecuencia, pero disponibles a los 15 a 20 días de concluido el mes de referencia. Lógicamente, un requisito tan estricto de anticipación limita enormemente la cantidad de series utilizables para la predicción. No obstante, esta aclaración, recurriremos al cuerpo teórico de *nowcasting* para explotar muchas de sus ventajas, como las técnicas de reducción de dimensión para hallar las variables latentes que guían la evolución del EMAE. Al final del ejercicio esperamos proponer un modelo que, sin ser el definitivo, permita proyectar el EMAE con cierta precisión, y discutir la posibilidad de ajustar la predicción a medida que nuevas variables predictoras se hallen disponibles.

2 Métodos

El procedimiento seguido para hallar un modelo que permita proyectar el EMAE prácticamente en tiempo real es el que sigue.

(a) Partimos de ocho series de coyuntura (IVA-DGI deflactado con IPIM, índice MERVAL deflactado con IPIM, EMBI, Índice de Confianza del Consumidor Nacional, ITCRM del BCRA, producción de automóviles en unidades, producción de utilitarios, y despachos de cemento Portland) elegidas por su disponibilidad en los primeros quince días del mes posterior al dato de EMAE que interesa estimar. Las series seleccionadas provinieron de [1], [9] y [10], y abarcaron el período agosto de 2005 a mayo de 2022.

(b) Las series de coyuntura fueron desestacionalizadas con el programa X13-ARIMA, consistidas con sus promedios anuales, y rescaladas a 100 en año 2012. Se chequeó estacionariedad mediante el test de Dickey y Fuller Aumentado o ADF.[2] De las ocho series de coyuntura se rechazó la presencia de raíz unitaria solamente en dos: utilitarios y despachos de

[1] Una metodología estándar de proyección de crecimiento puede consultarse, por ejemplo, en [5] y [6].

[2] El test se realizó con un programa propio escrito en lenguaje matricial `Euler Math Toolbox`. El modelo utilizado para la prueba consideró constante, tendencia y un rezago de la variable bajo análisis en diferencias.

cemento. No se rechazó la presencia de raíz unitaria en la serie de EMAE desestacionalizado. En consecuencia, todas las series fueron tomadas en diferencias, sabiendo que las de utilitarios y despachos de cemento podrían estar sobrediferenciadas, cuestión que quedó pendiente de revisión para más adelante.

(c) Se calcularon las componentes principales de las series en diferencias observándose que las primeras tres componentes acumulaban el 55 % de variabilidad total. Si bien ocho indicadores no es una cantidad tan grande que impida utilizarlos todos para proyectar el EMAE, quisimos realizar un ejercicio de reducción de dimensión compatible con colecciones más amplias de indicadores de coyuntura.

(d) A continuación se ajustó al ΔEMAE un modelo de rezagos distribuidos (ARDL) de segundo orden en el cual las variables exógenas fueron las tres primeras componentes principales de las ocho variables de coyuntura ya mencionadas. Formalmente, el modelo es el siguiente

$$\Delta y_t = \mu + \beta_1 \, z_{1t} + \ldots + \beta_m \, z_{mt} + \phi_1 \, \Delta y_{t-1} + \phi_2 \, \Delta y_{t-2} + \epsilon_t, \tag{1}$$

donde y_t es el EMAE desestacionalizado (publicado por INDEC) y rescalado a 100 en 2012, $z_{1t} \ldots z_{mt}$ son las primeras m componentes principales de las series de coyuntura (en nuestro caso $m = 3$), y ϵ_t es un término de error $i.i.d$ con media 0 y varianza σ^2. Nótese que la ecuación anterior no es una ecuación *bridge* como las que se utilizan para *nowcasting*. El modelo se ajustó con un programa propio escrito en `Euler Math Toolbox` al igual que el cálculo de las componentes principales. Los resultados del ajuste son los siguientes:

Variable	Estim.	Desvío	t	valor-p
Const.	0.1289	0.0965	1.3368	0.0914
Componente 1	-0.3866	0.0341	-11.3361	$< 0{,}0001$
Componente 2	0.2132	0.0466	4.5783	$< 0{,}0001$
Componente 3	-0.0122	0.0436	-0.2809	0.3895
$\Delta\text{EMAE}_{t\text{-}1}$	0.1686	0.0536	3.1429	0.0010
$\Delta\text{EMAE}_{t\text{-}2}$	-0.0870	0.0540	-1.6106	0.0544

El coeficiente de determinación ajustado de este modelo fue $R^2 = 0,45$; y, el valor-p del estadístico de Ljung-Box asociado resultó $\chi^2 = 0,0966$, por lo cual no hallamos evidencia de autocorrelación de los residuos.

(e) Se repitió el mismo ejercicio, pero calculando las componentes principales a partir de una matriz de datos que combinó de series no estacionarias en diferencias y las series estacionarias en niveles. Lógicamente, las tres componentes seleccionadas con esta variante fueron distintas a las anteriores, aunque acumularon aproximadamente la misma proporción de variabilidad total. Los resultados son los siguientes:

Variables	Estim.	Desvío	t	valor-p
Const.	0.1268	0.1105	1.1476	0.1263
Componente 1	0.2759	0.0385	7.1705	$< 0{,}0001$
Componente 2	0.1228	0.0473	2.5944	0.0051
Componente 3	0.1892	0.0679	2.7854	0.0029
$\Delta\text{EMAE}_{t\text{-}1}$	0.1542	0.0620	2.4870	0.0069
$\Delta\text{EMAE}_{t\text{-}2}$	-0.1326	0.0619	-2.1407	0.0168

(f) Por último, se replicó el ajuste del modelo (1) pero excluyendo los años 2020, 2021 y lo que va de 2022, a fin de chequear la significatividad de los parámetros estimados sin el efecto de la cuarentena. Los resultados fueron los siguientes:

Parámetro	$\hat{\beta}_j$	$s_{\hat{\beta}_j}$	t	valor-p
Const.	0.1053	0.0741	1.4216	0.0785
Componente 1	0.1453	0.0275	5.2838	$< 0{,}0001$
Componente 2	-0.1978	0.0447	-4.4233	$< 0{,}0001$
Componente 3	0.1615	0.0594	2.7163	0.0037
ΔEMAE$_{t\text{-}1}$	0.0395	0.067	0.5888	0.2784
ΔEMAE$_{t\text{-}2}$	0.1212	0.0672	1.8044	0.0365

Para evaluar cómo opera el modelo en la práctica simulamos la proyección del EMAE durante la pandemia de COVID. Para ello, repetimos el cálculo de las componentes principales, la estimación de parámetros y la proyección del EMAE mes a mes entre enero y julio de 2020, y comparamos las proyecciones con las diferencias ΔEMAE efectivamente observadas.[3] Se puede ver que el modelo anticipó razonablemente bien el derrumbe de la actividad económica en marzo y abril de 2020 y la recuperación de mayo y junio, aunque no captó la magnitud de la caída. Conviene tener en cuenta que en la muestra utilizada para estimar los parámetros no había antecedentes de caídas como la de marzo y abril de 2020.

Año 2020	ΔEMAE estim.	ΔEMAE obs.
Enero	1,6711	0,2756
Febrero	-0,9897	-0,3635
Marzo	-5,0039	-10,0258
Abril	-7,4028	-14,3871
Mayo	3,1593	7,9239
Junio	5,5609	5,4310
Julio	1,8713	1,7719

3 Discusión

El modelo (1) y las variable seleccionadas son adecuadas para anticipar el EMAE. Las tablas expuestas en los puntos (d), (e) y (f) muestran que al menos dos de las tres primeras componentes principales (calculadas a partir de ocho indicadores coincidentes) resultaron significativas y que estas componentes (más los rezagos de ΔEMAE) explican casi el 50 % de la variabilidad total observada en los cambios de nivel del EMAE de meses consecutivos. Esta proporción puede parecer modesta, pero debe ser contextualizada en una selección de indicadores de coyuntura muy reducida. Llama la atención que la tercera componente, que no había resultado significativa al incluir el período de pandemia y post-pandemia, resultara significativa al excluir este período; y que también resultara significativa al combinar variables de coyuntura en niveles y en diferencias. Sin embargo, esta contradicción es sólo aparente porque, por construcción, los coeficientes del modelo son condicionales a las componentes que ingresan al modelo. Para cada muestra las componentes que ingresan podrían ser distintas. En el extremo, por ejemplo, las primeras m componentes que ingresan al modelo cuando se dispone de T observaciones podrían ser distintas a las que ingresan al agregar un nuevo conjunto de observaciones $\{y_t, \mathbf{x}_t\}$. A medida que $T \to \infty$, sin embargo, las componentes convergerán a las componentes poblacionales

[3] El primer conjunto de parámetros estimados utilizado en la simulación en el de la tabla del punto (f).

si y sólo se cumple que $\text{lím}_{T\to\infty} \mathbf{X}'\mathbf{X}/T = \mathbf{Q}$.

La selección de indicadores de coyuntura tan tempranos nos permitió eludir el problema de determinar la cantidad "óptima" de indicadores a utilizar. La optimalidad a la que nos referimos no consiste en alcanzar el máximo nivel de ajuste con un mínimo de variables, sino con la mayor anticipación posible. Lógicamente, cuantos más indicadores se utilicen para proyectar una variable, más ajustada será la predicción al valor efectivamente observado, pero menor la anticipación con la que dicha variable pueda ser predicha, al menos con un nivel de confianza razonable. Para resolver esta contrariedad, [4] proponen un esquema de cálculo secuencial que consiste en agrupar los indicadores de coyuntura en intervalos temporales e incorporar cada grupo al proceso de proyección a medida que se encuentren disponibles.[4] Otros autores, como [2], proponen además combinar varias proyecciones en un solo estimador y presentan evidencia de que tal combinación supera a cualquiera de las predicciones individuales. El esquema secuencial propuesto conceptualmente por [4] no aparece formalizado como un procedimiento secuencial real, es decir, uno que aproveche la predicción anterior y la enriquezca con nueva información disponible, sino que simplemente replica el procedimiento original ampliando en cada oportunidad la matriz de variables relacionadas. Un procedimiento secuencial podría plantearse descomponiendo la estimación de parámetros del modelo (1) en forma de regresión particionada.

$$\mathbf{b}_{T_1} = (\mathbf{Z}'_{T_1}\mathbf{Z}_{T_1})^{-1}\mathbf{Z}'_{T_1}(\Delta\mathbf{y}_T - \Delta\mathbf{Y}_{T-p}\hat{\boldsymbol{\phi}}_T) \tag{2}$$

donde $\mathbf{Z}_{T_1}$ es la matriz de componentes principales de un primer conjunto de indicadores de coyuntura (presumiblemente contemporáneos entre sí) y $\mathbf{e}_T = \Delta\mathbf{y}_T - \Delta\mathbf{Y}_{T-p}\hat{\boldsymbol{\phi}}_T$ es el vector de residuales del modelo AR(p) ajustado hasta el mes T. La matriz $\Delta\mathbf{Y}_{T-p}$ es una matriz de rezagos de $\Delta\mathbf{y}_T$ con ordenada al origen. Para que $\mathbf{Z}_{T_1}$ sea conformable se deberán omitir las p primeras filas o agregar p elementos nulos al principio de $\mathbf{e}_T$, así como eliminar la última fila de $\mathbf{Z}_{T_1}$, la que quedará reservada para predecir Δy_{T+1}. Sin embargo, para evitar ambigëdades en la notación asumiremos que el subíndice T indica tanto el alcance como la dimensión de la serie, de manera que la series arrancan en $T-p$. La varianza de $\mathbf{b}_{T_1}$ es

$$var(\mathbf{b}_{T_1}) = \sigma^2(\mathbf{Z}'_{T_1}\mathbf{Z}_{T_1})^{-1}\mathbf{Z}'_{T_1}(\mathbf{I}_T - \mathbf{M})\mathbf{Z}_{T_1}(\mathbf{Z}'_{T_1}\mathbf{Z}_{T_1})^{-1} \tag{3}$$

donde $\mathbf{M}$ es la matriz de proyección de $T \times T$ del modelo autorregresivo utilizado para calcular los residuos $\mathbf{e}_T$. Cabe recordar que la matriz $\mathbf{Z}_{T_1}$ contiene m componentes principales, de manera que su dimensión es $T \times m$. No se le agrega una columna de $\mathbf{1}_T$ porque el vector de residuos es, por definición, centrado.

En la segunda etapa, una vez difundidos los indicadores del siguiente intervalo temporal, se calculan las m componentes principales y se optimiza la función

$$\mathcal{L} = (\mathbf{e}_T - \mathbf{Z}_{T_2}\boldsymbol{\beta}_{T_2})'(\mathbf{e}_T - \mathbf{Z}_{T_2}\boldsymbol{\beta}_{T_2}) + (\mathbf{b}_{T_1} - \boldsymbol{\beta}_{T_2})'(\mathbf{b}_{T_1} - \boldsymbol{\beta}_{T_2}). \tag{4}$$

En la función objetivo (4) se reemplaza $\mathbf{Z}_{T_1}$ por $\mathbf{Z}_{T_2}$ para actualizar $\mathbf{b}_{T_1}$ pero minimizando simultáneamente la distancia entre el nuevo estimador y el estimador anterior. Queda implícito que la cantidad de componentes tomadas en la segunda etapa es la misma que en la primera. El sentido de esta doble minimización es incorporar a la estimación previa la información aportada por $\mathbf{Z}_{T_1}$. Reiteramos que las componentes de $\mathbf{Z}_{T_2}$ no necesariamente coincidirán con las de $\mathbf{Z}_{T_1}$. La optimización de $\mathcal{L}$ conduce al estimador de $\boldsymbol{\beta}_{T_2}$

$$\mathbf{b}_{T_2} = (\mathbf{Z}'_{T_2}\mathbf{Z}_{T_2} + \mathbf{I}_m)^{-1}(\mathbf{Z}'_{T_2}\mathbf{e}_T + \mathbf{b}_{T_1}). \tag{5}$$

[4] Los mismos autores, sin embargo, no siguen este procedimiento, sino que realizan una preselección de las variables más correlacionadas con la de interés y extraen luego las componentes principales de este conjunto.

Llamando $\mathbf{A} = (\mathbf{Z}'_{T_2}\mathbf{Z}_{T_2} + \mathbf{I}_m)^{-1}$, la varianza de $\mathbf{b}_{T_2}$ es

$$var(\mathbf{b}_{T_2}) = \mathbf{A}\left[\sigma^2\mathbf{Z}'_{T_2}(\mathbf{I}_T - \mathbf{M})\mathbf{Z}_{T_2} + cov(\mathbf{Z}'_{T_2}\mathbf{e}_T, \mathbf{b}'_{T_1}) + cov(\mathbf{b}_{T_1}, \mathbf{e}'_T\mathbf{Z}_{T_2}) + var(\mathbf{b}_{T_1})\right]\mathbf{A}'$$

donde $cov(\mathbf{Z}'_{T_2}\mathbf{e}_T, \mathbf{b}'_{T_1}) = \sigma^2\mathbf{Z}'_{T_2}(\mathbf{I}_T - \mathbf{M})\mathbf{Z}_{T_1}(\mathbf{Z}'_{T_1}\mathbf{Z}_{T_1})^{-1}$ y $cov(\mathbf{b}_{T_1}, \mathbf{e}'_T\mathbf{Z}_{T_2})$ es su traspuesta. La expresión de la varianza se puede simplificar llamando $\mathbf{B} = (\mathbf{Z}'_{T_1}\mathbf{Z}_{T_1})^{-1}\mathbf{Z}'_{T_1}$

$$var(\mathbf{b}_{T_2}) = \sigma^2\mathbf{A}(\mathbf{Z}'_{T_2} + \mathbf{B})(\mathbf{I}_T - \mathbf{M})(\mathbf{Z}_{T_2} + \mathbf{B}')\mathbf{A}'. \tag{6}$$

El estimador $\mathbf{b}_{T_2}$ se puede extender a más de dos intervalos temporales simplemente reemplazando T_2 por $T_3, T_4, \ldots$ en la fórmula (5), aunque en la práctica sería poco probable superar los dos o tres intervalos. La cantidad "óptima" de intervalos temporales se alcanzaría, en teoría, cuando no haya diferencias estadísticamente significativas entre dos predicciones consecutivas de Δy_{T+1}. El intervalo de confianza para Δy_{T+1} en cada etapa es

$$\Delta\hat{y}_{T+1} \pm t_{(T-m+1;1-\alpha/2)}\sqrt{\mathbf{z}'_{T+1}var(\mathbf{b}_T)\mathbf{z}_{T+1}}, \tag{7}$$

donde $var(\mathbf{b}_T)$ se refiere la varianza de $\mathbf{b}_{T_1}, \mathbf{b}_{T_2}, \ldots$ según corresponda.

Referencias

[1] Ámbito Financiero, 2022. Riesgo País Histórico. Descargable libremente de `https://www.ambito.com/contenidos/riesgo-pais-historico.html`

[2] Blanco E., D'Amato L., Dogliolo Fiorella, y M. L. Garegnani, 2017. Nowcasting GDP in Argentina: Comparing the predictive ability of different models, Economic Research Working Papers, No. 74, Banco Central de la República Argentina (BCRA), Investigaciones Económicas (ie), Buenos Aires.

[3] Blanco E., D'Amato L., Dogliolo Fiorella, y M. L. Garegnani, 2022. Nowcasting durante la pandemia: Un apartado para Argentina 2022. Documentos de Trabajo, No. 2022/99, Banco Central de la República Argentina (BCRA), Investigaciones Económicas (ie), Buenos Aires.

[4] D'Amato L., Garegnani M. L. y E. Blanco, 2015. Nowcasting de PIB: evaluando las condiciones cíclicas de la economía argentina. Working Paper, No. 2015/69, Banco Central de la República Argentina (BCRA), Investigaciones Económicas (ie), Buenos Aires.

[5] Frank L., 2020. Revisión de modelos para la desestacionalización de series mensuales y trimestrales de actividad económica. MPRA Paper No. 111423. Disponible en `https://mpra.ub.uni-muenchen.de/111423/`

[6] Frank L., 2021. Revisión de modelos para ajuste estacional y proyección de sectores de actividad del EMAE. Año 2021. MPRA Paper No. 112285. Disponible en `https://mpra.ub.uni-muenchen.de/112285/`

[7] INDEC, 2016. Estimador Mensual de Actividad Económica (EMAE), Base año 2004. Metodología INDEC Nro. 20. Instituto Nacional de Estadística y Censos. ISSN 2545-7179.

[8] Índices de actividad económica provincial en base a un modelo factorial dinámico. Argentina 1997-2019. Cuadernos del CIMBAGE Nro. 22 (2 a ed. 2020) 145-173.

[9] Ministerio de Economía, Subsecretaría de Programación Macroeconómica, 2022. Información Económica al día. Descargable libremente de `https://www.argentina.gob.ar/economia/politicaeconomica/macroeconomica`

[10] Universidad Torcuato Di Tella, Centro de Investigación en Finanzas, 2022. Índice de Confianza del Consumidor. Descargable libremente de `https://www.utdt.edu`