BIG DATA

ANÁLISIS ECONOMÉTRICO Y BIG DATA

Daniel Peña
Pilar Poncela
Esther Ruiz
(editores)

funcas

BIG DATA

ANÁLISIS ECONOMÉTRICO Y BIG DATA

Daniel Peña
Pilar Poncela
Esther Ruiz
(editores)

funcas

Impreso en España
Edita: Funcas
Caballero de Gracia, 28, 28013 - Madrid

ISBN: 978-84-17609-54-2
Depósito legal: M-19102-2021
Maquetación: Funcas
Imprime: Cecabank

Contenido

Presentación

Este libro está dedicado a analizar cómo la presencia de datos masivos, denominados con frecuencia por su nombre en inglés, *big data,* ofrece nuevas oportunidades de aprendizaje en muchas áreas del Análisis Econométrico. Complementa, por tanto, el trabajo de los mismos editores publicado recientemente por Funcas sobre nuevos métodos de predicción económica con datos masivos[1]. En esta ocasión, se revisan distintas técnicas de Aprendizaje Automático (*Machine Learning, ML* por sus siglas en inglés) que se aplican a grandes conjuntos de datos para resolver distintas cuestiones en análisis econométrico. El volumen consta de nueve capítulos escritos por expertos en "big data" y aprendizaje automático y/o análisis económico, que se organizan de la siguiente manera.

En el capítulo 1, **García Montalvo** nos presenta una revisión crítica sobre la aplicación de distintas técnicas de aprendizaje automático a conjuntos de datos masivos para resolver problemas tales como la calificación crediticia, el seguimiento de la economía a muy alta frecuencia y la construcción de indicadores de crecimiento y desigualdad usando imágenes de satélites. El trabajo incluye también aplicaciones predictivas para el precio de la vivienda o, incluso, la predicción electoral. El autor alerta sobre los peligros de la utilización indiscriminada de este tipo de técnicas y, en particular, nos llama la atención sobre el peligro de la pérdida de privacidad, la replicación de situaciones de discriminación, la confianza en métodos que se comprenden mal y funcionan como una "caja negra" o la necesidad de actualizar los procedimientos cuando varían las condiciones en las que se recogen los datos. **Cuerpo** y **Morales** presentan otra aplicación interesante en el capítulo 2, dedicada a comprender la evolución del consumo público utilizando datos de la plataforma de contratación del sector público. Su trabajo utiliza una herramienta nueva en econometría y que se usa cada vez con más frecuencia en el análisis económico, los bosques aleatorios, o *random forest,* que son conjuntos de árboles de decisión muy útiles para relacionar variables de forma no lineal. En el capítulo 3, **Dolado** nos ofrece una panorámica sobre los métodos de regularización, que se utilizan cuando el número de parámetros es elevado, como ocurre con problemas de análisis económico cuantitativo con muchas variables. Estos métodos establecen restricciones en la función de estimación de los parámetros para reducir su número y tamaño, permitiendo una estimación más eficiente. Aquí se aplican a problemas de economía laboral, donde el objetivo es evaluar el efecto de determinadas políticas sobre variables del mercado de trabajo. En el

[1] https://www.funcas.es/libro/nuevos-metodos-de-prediccion-economica-con-datos-masivos/

capítulo 4, **Guerrero, Corona** y **Mendoza** ilustran cómo los nuevos datos pueden mejorar la estimación del producto interior bruto (PIB) de México. Proponen un método que utiliza de datos de luminosidad nocturna, recogidos por satélite, como indicador de la actividad económica y los combina con mediciones tradicionales de contabilidad nacional para mejorar las estimaciones de PIB. Su trabajo ilustra de forma efectiva y convincente las oportunidades que ofrecen los nuevos datos en problemas clásicos. Finalmente, cierra este primer bloque de aplicaciones el capítulo 5, donde **Poncela** y **Senra** analizan cuestiones de integración financiera, *nowcasting* (o predicción de la actividad económica en tiempo real) y de construcción de nuevos indicadores utilizando datos masivos, señalando algunas oportunidades que big data proporciona y apuntando algunos retos que quedan por resolver.

Un rasgo importante de muchas series económicas es la estacionalidad y en los capítulos 6 y 7 se presentan nuevas herramientas para su estudio. En el primero de ellos, **Nieto, Peña** y **Bolívar** introducen el modelo factorial estacional para series no estacionarias. El modelo factorial es una de las principales herramientas con las que se cuenta hoy en día para análisis macroeconómico y la práctica habitual es desestacionalizar las series antes de introducirlas en el modelo. La generalización que presentan Nieto, Peña y Bolívar permite aplicar esta técnica sin necesidad de preprocesar los datos. Los autores ilustran el funcionamiento de su modelo en análisis económico estudiando el efectivo en circulación, como un agregado monetario, en una muestra de 15 países de América Latina. En el capítulo 7, **Quilis** examina la relación entre la estacionalidad y el ciclo. Para ello, utiliza el análisis de conglomerados y el modelo factorial para identificar pautas comunes en los componentes estacional y cíclico que han sido extraídos previamente. La metodología propuesta es aplicada a una base de datos territorial de la economía española cuya cobertura es muy amplia, tanto temporal (1974-2019) como espacial (nivel provincial).

Finalmente, los dos últimos capítulos revisan otras técnicas para datos masivos. En el capítulo 8, **Galeano** presenta un nuevo enfoque para series temporales con pautas estables suaves de variación, el análisis de series temporales funcionales, y aplica esta metodología para analizar las curvas de rendimientos de los precios horarios de la electricidad en el mercado ibérico. En el capítulo 9, **Comas** y **Alonso** presentan una modificación del algoritmo de k-vecinos más cercanos en problemas de clasificación cuando se aplica a datos masivos y el coste computacional de calcular todas las distancias es tan alto que no lo hace factible. Aplican este método a varios problemas, entre ellos, el análisis de series temporales de oferta de electricidad horaria en el mercado secundario en España en el período 2014 a 2019.

Estos trabajos fueron presentados en una Jornada organizada por los editores de este libro y celebrada en Funcas el 24 de noviembre de 2020 bajo la denominación de *Análisis Económico y Big Data.* Hemos sustituido en el título de este volumen económico por econométrico para ser más precisos en las aportaciones que contiene. El lector interesado puede encontrar en la web (YouTube[2]) la grabación de las presentaciones que se hicieron en las Jornadas.

[2] https://www.youtube.com/watch?v=0gVekpBan30

Los editores queremos agradecer a Funcas su apoyo en la realización de este libro que esperamos contribuya a difundir la aplicación de técnicas de datos masivos en problemas económicos abordados desde un enfoque econométrico.

Abril, 2021

CAPÍTULO I

Data Science y sus aplicaciones económicas: una perspectiva personal

José García Montalvo*

En este trabajo se realiza un recorrido por diversas aplicaciones económicas basadas en *Big Data* y la aplicación de técnicas de *Machine Learning.* Las aplicaciones incluyen la ingeniería reversa de procedimientos de calificación crediticia, el seguimiento de la economía a muy alta frecuencia durante la pandemia de la COVID-19, la predicción del precio de la vivienda a nivel de código postal, la construcción de indicadores de crecimiento y desigualdad usando imágenes de satélites y la predicción electoral. El objetivo es destacar, utilizando estas aplicaciones, los aspectos que aportan más posibilidades en la utilización de dichas técnicas en el campo de la economía, así como matizar las excesivas expectativas que estos procedimientos puedan haber generado.

Palabras clave: árboles de decisión, imágenes de satélite, *scoring* crediticio.

* El autor agradece la colaboración, en las aplicaciones con datos bancarios, de Miguel Barcino, Estrella García, Lupina Iturriaga, Olga Pascual, José Carlos Pla y Guillermo Soler. Especial agradecimiento a los coautores de los estudios sobre los que se basa este trabajo: Oriol Aspach, Ruben Durante, Alberto Graziano, Josep Mestres, Omiros Papaspiliopoulos, J. M. Raya, Marta Reynal-Querol, y Timothée Stumpf-Fétizon.

1. INTRODUCCIÓN

La introducción de la ciencia de los datos *(Data Science)* en la economía, la empresa y las finanzas ha avanzado enormemente en los últimos años. En García-Montalvo (2014) se introducen los elementos fundamentales que caracterizan los proyectos de *Data Science,* entendido como la fusión del big data y las técnicas de *Machine Learning*[1], especialmente en contraste con la metodología habitual en econometría. Asimismo se analizan las primeras aplicaciones exitosas de técnicas de big data (BD) y *Machine Learning* (*ML*) en el campo de la economía (el proyecto Billion Prices de MIT o las predicciones híbridas utilizando indicadores adaptados de Google Trends) y, particularmente, en el sector financiero. García-Montalvo (2014) analiza la aplicación de técnicas de ML a las calificaciones crediticias de familias y empresas, tanto en su versión tradicional como en sus versiones basadas en la reputación social *(credit score social)*. Entre otras aplicaciones al ámbito financiero y bancario se incluían la protección contra el fraude en tarjetas de crédito, la aplicación al *peer-to-peer lending,* la segmentación de los clientes bancarios, el cumplimiento de la normativa regulatoria, la optimización del capital bancario o la creciente personalización de los precios de los seguros. El artículo acaba con algunos peligros de la utilización indiscriminada de algoritmos de ML a bases de datos cada vez más masivas. En particular, el peligro de pérdida de privacidad, la replicación de la discriminación observada en la realidad, la confianza en una caja negra o la necesidad de actualizar los procedimientos cuando varían las condiciones en las que se recogen los datos como demuestra el caso de la predicción de la gripe con el Global Flu Trends.

Este artículo es una continuación de aquel primer trabajo. Con la perspectiva del tiempo transcurrido he podido comprender mejor las dificultades de los alumnos de econometría para entender la aproximación adoptada en los proyectos de BD y las técnicas de ML. También he tenido multitud de experiencias en proyectos, cada vez más intensivos en la utilización de datos, que me han permitido mejorar la evaluación de qué cosas tienen sentido y qué cosas son meros fuegos de artificio sin ninguna implicación de mejora de nuestra comprensión de los fenómenos analizados. Por último, algunos de los peligros potenciales comentados en el artículo de 2014 se han manifestado como inconvenientes efectivos. El objetivo de este artículo es desarrollar estos tres temas actualizando la visión del artículo original y desarrollando nuevos temas.

2. *MACHINE LEARNING* PARA ECONÓMETRAS

En García-Montalvo (2014) se desarrollan los conceptos fundamentales que diferencian la aproximación basada en el ML de la aproximación de la econometría moderna. Fundamentalmente se enfatizaba el elevado nivel de ruido frente a señal del BD comparativamente

[1] En este trabajo se usa el término aprendizaje en sentido amplio. Estrictamente hablando se debe distinguir entre la aproximación basada en el aprendizaje y la basada en el diseño. En la visión basada en el aprendizaje los datos son necesarios para derivar la función objetivo mientras que en la basada en el diseño se postula una función objetivo con anterioridad a disponer de los datos.

con los datos utilizados normalmente en econometría así como la diferencia en perspectivas: mientras la econometría transitaba hacia técnicas que permitieran identificar y explicar los efectos causales, el ML solo está interesado en producir modelos predictivos y, por tanto, correlaciones. En econometría siempre tenemos un componente aleatorio que permite obtener bandas de confianza y realizar inferencias. Este no es el caso, en general, de las técnicas de ML. Estas diferencias no son complejas para un estudiante que ha superado un primer curso de econometría[2]. Los problemas, como ya sucedió en el pasado cuando otras técnicas estadísticas saltaron al trabajo con datos económicos, empiezan por la nomenclatura. Los libros de texto sobre machine learning hablan de *features* en lugar de variables explicativas, *labels* en lugar de variables endógenas, de muestras de entrenamiento y testeo en lugar de simplemente referirse a una muestra, etc. De la misma manera hablan obsesivamente de *sobreajuste* y el uso de *cross validation* sin referirse la mayor parte de las veces a R^2 o el R^2 ajustado. Finalmente se habla de LASSO, concepto que un estudiante de econometría de hace 30 años habría identificado como una generalización de la regresión *ridge* pero al que la mayoría de los estudiantes más recientes no han sido expuestos.

A estas diferencias terminológicas se unen profundas diferencias conceptuales. Un primer punto son los procedimientos de aprendizaje no supervisado que chocan con la visión tradicional de la econometría que trata con procedimientos supervisados. La teoría de la generalización y sus implicaciones respecto a la existencia de muestra de entrenamiento y muestras de testeo supone otra novedad conceptual. El estudiante de econometría está acostumbrado a la descomposición del error cuadrático medio de un estimador entre sesgo al cuadrado y varianza. Cuando hablas de los componentes del error esperado en la muestra de testeo tiene tres componentes (sesgo cuadrado, varianza y ruido) la influencia de la visión econométrica tradicional dificulta la comprensión del *trade-off* entre varianza y sesgo derivado de la complejidad creciente del algoritmo. Es más, cuesta explicar que el efecto no tiene relación con N (el tamaño de la muestra) sino con K (la complejidad del algoritmo). Si el sesgo es elevado hay que usar modelos más complejos (kernelizar), añadir features o utilizar *boosting* para algoritmos que producen resultados con sesgo. Si el problema es el contrario (varianza elevada o *sobreajuste*) la opción es añadir más datos, reducir la complejidad del modelo, por ejemplo, usando técnicas de regularización, *bagging*, etc. Otros aspectos difíciles de transmitir son la factibilidad del aprendizaje o, en la teoría de la generalización, la importancia de la dimensión de Vapnik por contraposición al simple número de parámetros a estimar.

3. CALIFICACIONES CREDITICIAS, RECOMENDACIONES E INGENIERÍA INVERSA

Montalvo y Reynal-Querol (2020) analizan la influencia del género del gestor encargado de la decisión sobre la concesión de un crédito sobre la posterior probabilidad de impago. Se dispone de información sobre más de un millón de solicitudes de créditos con las

[2] Pongamos un estudiante que entiende con claridad el funcionamiento del modelo lineal, incluida la predicción, y que sabe trabajar con variables dicotómicas tanto explicativas como endógenas.

características de la solicitud, destino, situación financiera del solicitante, colateral, etcétera. así como la calificación crediticia y la recomendación generada por el sistema experto. Los resultados muestran que las mujeres concedieron los créditos con recomendación negativa muchas menos veces que los hombres durante el periodo 2000-2008 y que este mayor seguimiento de la recomendación produjo una proporción menor de impagos cuando comenzó la crisis financiera. Para profundizar en este resultado nos planteamos realizar un análisis de discontinuidad en torno a la puntuación que marcaba la recomendación de no concesión. Sin embargo, la situación no era sencilla, puesto que la visualización de la relación entre *score* y recomendación mostraba con claridad que no existían unos límites claramente definidos en función de la puntuación. La figura 1 muestra un ejemplo con los resultados de la tabla referida a créditos hipotecarios para no clientes[3]. El eje de las abscisas representa el momento de concesión del crédito y el de ordenadas la puntuación asignada. La escala de grises representa el año de concesión. El primer panel contiene los créditos que el sistema de recomendación considera de riesgo bajo o muy bajo (VLR y LR), el segundo los créditos de riesgo neutral (NR), el tercero los de riesgo alto (HR) y el cuarto los de riesgo muy alto (VHR). Como se puede comprobar las puntuaciones para diferentes tablas se solapan a pesar de que las recomendaciones de riesgo bajo o neutro suponían concesión, mientras que las de riesgo alto o muy alto implicaban que, por lo general, debían rechazarse.

De esta forma es evidente que no existe una relación lineal entre puntuaciones y recomendaciones que permita construir un diseño de discontinuidad. Además, se comprueba que las tablas cambian en el tiempo, lo que supone una dificultad adicional.

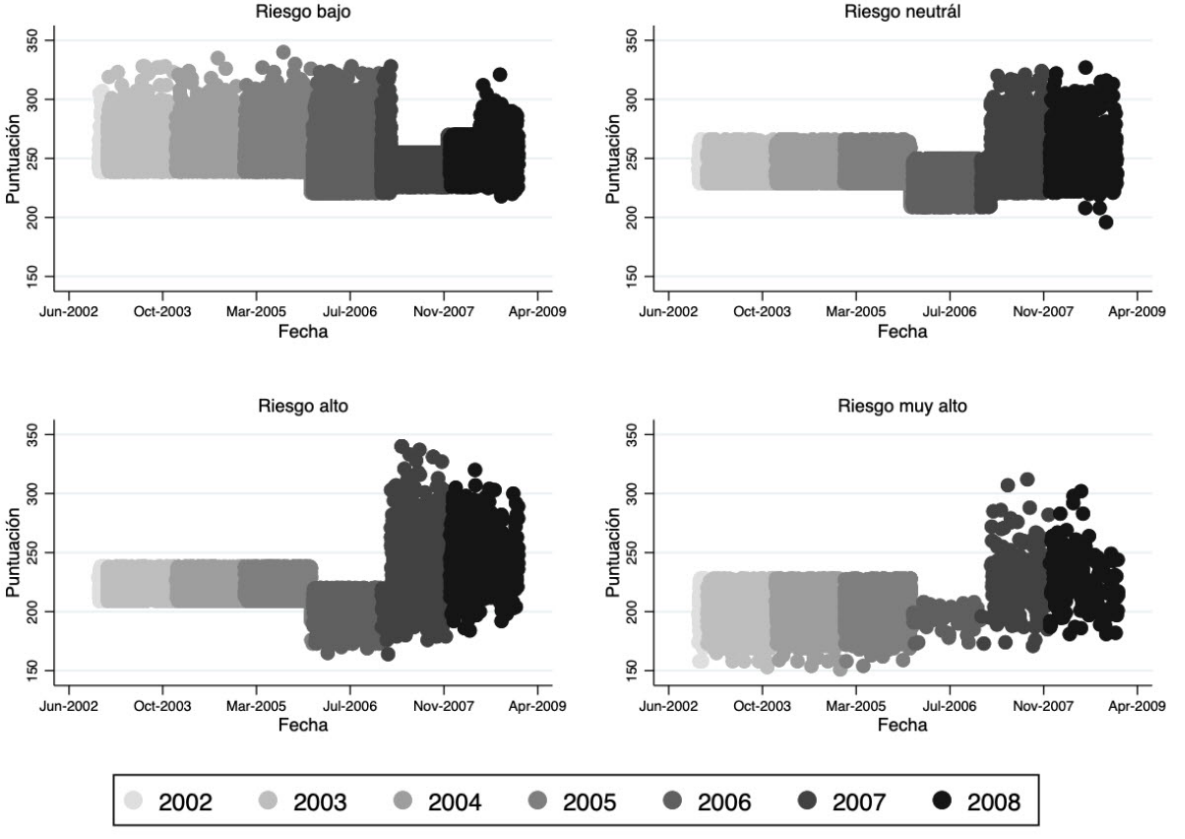
Como los datos se referían a un periodo bastante antiguo y la entidad que había proporcionado la información había desaparecido fue imposible encontrar los algoritmos que habían generado las recomendaciones[4]. De esta forma, la única opción viable fue utilizar ingeniería inversa: descubrir los algoritmos a partir de las recomendaciones y el resto de variables disponibles. Las variables consideradas fueron el *scoring,* la relación préstamo/valor *(loan to value, LTV)*, la relación préstamo/renta *(loan to income, LTI),* el tamaño del préstamo, la extensión temporal, el crédito sobre el patrimonio, la edad, el número de años como cliente, el balance medio en los últimos 6 meses, la ratio entre el balance medio a 6 y 12 meses, el tipo de contrato y su duración, el destino del crédito y el estado civil, entre otras. En principio se probaron diversos algoritmos, pero el que generalizaba mejor era un simple árbol de decisión. De hecho, no hizo falta ningún tipo de *bagging* o versiones tipo *random forest.* El grado de impureza de las ramas finales era 0 y así sucedía en todas las tablas. Por ejemplo, la figura 2 muestra el ajuste del árbol de decisión para las recomendaciones de los créditos hipotecarios a no clientes donde las predicciones de todos los tipos de recomendaciones de la última hoja se corresponden con precisión con las recomendaciones efectivamente observadas. Las únicas variables que tienen capacidad predictiva son la puntuación y el LTI.

[3] El banco mantenía 13 tablas de puntuación en función del producto (crédito hipotecario, consumo, subrogación hipotecaria, etc.) y de la relación del solicitante con el banco (cliente, no cliente).

[4] Se intentó contactar con los antiguos gestores de riesgos, pero obviamente no recordaban con precisión los algoritmos utilizados para cada producto y no tenían los manuales que explicaban los procedimientos. También contactamos con la empresa que colaboró en la implementación de los algoritmos que tampoco nos pudo dar información.

Figura 1.

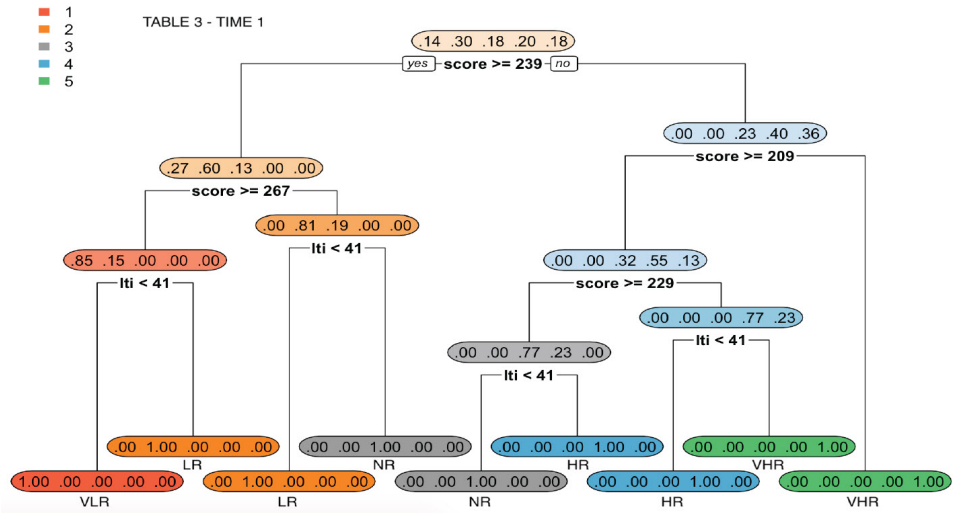
Relación entre puntuaciones y recomendaciones para créditos hipotecarios a no clientes



Fuente: Montalvo y Reynal-Querol (2020).

Figura 2.

Árbol de decisión para la tabla 3 (primer periodo)



Fuente: Elaboración propia.

Este ejercicio muestra un ejemplo claro de utilización adecuada y efectiva de técnicas de ML en una situación donde la única solución es la reingeniería del algoritmo que generó los datos originales.

4. *BIG DATA,* COVID-19 Y ECONOMÍA EN TIEMPO REAL

La propagación de COVID-19 ha afectado enormemente la actividad económica en todo el mundo. Los gobiernos han tomado muchas medidas para responder a la pandemia. Existe un alto grado de incertidumbre sobre el efecto de las respuestas de política y la idoneidad del monto total de apoyo público a la economía. Sin embargo, la mayoría de los indicadores oficiales y las estadísticas macroeconómicas tienen una frecuencia baja y se producen con grandes retrasos. Este es un desafío importante para los responsables de la formulación de políticas en sus esfuerzos por adaptar sus respuestas para aplanar la curva de recesión (Gourinchas, 2020) después de aplanar la curva de infección.

La necesidad de tener información actualizada con alta frecuencia sobre las principales macromagnitudes y el impacto de las medidas de política económica ha generado diversas iniciativas internacionales para rastrear en tiempo real (o con indicadores de alta frecuencia) la evolución de la actividad económica. Por ejemplo, Cicala (2020) utiliza el uso de electricidad de la red europea para representar la evolución de la actividad económica, ya que el consumo de electricidad está altamente correlacionado con su uso empresarial. Bick y Blandin

(2020) recurren a una encuesta en tiempo real *(RPS),* siguiendo la estructura de *CPS (Current Population Survey)* para construir estimaciones de alta frecuencia de empleo, horas trabajadas e ingresos. Chen *et al.* (2020) muestran que estas medidas de alta frecuencia de consumo de energía y horas trabajadas están fuertemente correlacionadas con los indicadores de movilidad del *Informe de Movilidad de la Comunidad de Google.* Chetty *et al.* (2020) han creado un rastreador económico para medir la actividad económica con alta frecuencia en los EE. UU. El objetivo es utilizar información de empresas privadas para construir agregados que tengan utilidad para el seguimiento de la economía en tiempo real y la medición del impacto de las medidas de política económica. La ventaja de esta información es su inmediatez, la disponibilidad con muy alta frecuencia (incluso diaria) y la posibilidad de agregarla en función de muy diferentes criterios. Chetty *et al.* (2020) utilizan información económica anónima de empresas privadas para medir el gasto de los consumidores (transacciones basadas en tarjetas de Affinity Solutions); la tasa de cierre de pequeñas empresas (negocios que realizan transacciones en un día determinado desde Womply); el tiempo dedicado al trabajo (datos GPS proporcionados por Google); o las horas trabajadas en pequeñas empresas (proporcionadas por Homebase).

En los últimos años, la investigación económica ha comenzado también a aprovechar la ingente cantidad de datos disponibles en las aplicaciones de gestión de finanzas personales o en las cuentas bancarias de las entidades financieras, para analizar diversas hipótesis. Gelman *et al.* (2014) utilizan datos diarios de un agregador financiero para estudiar la hipótesis del ingreso permanente (PIH), encontrando un exceso de sensibilidad del consumo al ingreso. Gelman *et al.* (2014) encuentran que, alrededor del período de recepción del salario y los pagos de la Seguridad Social, hay un significativo aumento del consumo. Este resultado se produce tanto en el gasto total, como en el gasto no recurrente y el gasto no esencial (como, por ejemplo, comida rápida y cafeterías). Olafsson y Page (2018) estudian también la PIH utilizando datos del agregador de *software* financiero de Islandia Meniga. Como Gelman *et al.* (2014) encuentran una respuesta significativa del consumo en el día de pago que es consistente en todas las categorías de ingresos y gastos. Baker (2018) también utiliza datos de un sitio web de finanzas personales en línea y encuentra que la heterogeneidad en la elasticidad del consumo se puede explicar por el crédito y la liquidez.

La disponibilidad de información económica muy granular y casi en tiempo real permite analizar el impacto de las políticas públicas con muy pequeños retardos respecto al momento de su ejecución. Por ejemplo, diversas investigaciones ya han comenzado a utilizar datos de alta frecuencia para analizar el impacto de los paquetes de estímulo económico para mitigar el efecto de la epidemia de COVID-19 en la actividad económica. Dos ejemplos para el caso de Estados Unidos son el efecto sobre el empleo agregado del Programa de Protección de Pagos (PPP) para pymes o el efecto sobre el consumo de los cheques de estímulo enviados por la Administración Federal. Autor *et al.* (2020) muestran que el empleo en las empresas elegibles aumentó entre el 2 y el 4,5 %, aunque los resultados son preliminares. Baker *et al.* (2020), utilizando datos de agregación financiera y aplicaciones de servicios, muestran que la propensión marginal a consumir a partir de estos cheques se sitúa entre el 0,26 y el 0,36 dependiendo de que la muestra tenga los pesos originales o se posestratifique para representar a la población de Estados Unidos. Cox *et al.* (2020) utilizan información de cuentas bancarias

anónimas de varios millones de clientes de JP Morgan Chase para estudiar el efecto heterogéneo de la pandemia de COVID-19 en el gasto y el ahorro. Sheridan *et al.* (2020) utilizan datos de un gran banco escandinavo para mostrar que las leyes de distanciamiento social tuvieron un pequeño impacto en la actividad económica. Sheridan *et al.* (2020) comparan los datos de Suecia, que no impuso un confinamiento estricto, con los datos de Dinamarca, que sí cerró la economía. La diferencia como resultado del cierre son 4 puntos porcentuales adicionales de reducción del gasto en Dinamarca. Este resultado concuerda con los hallazgos en Chetty *et al.* (2020) que muestran que la fuerte reducción en el gasto de los consumidores, el gasto de las pequeñas empresas y el tiempo dedicado al trabajo en la economía de EE. UU. comenzó semanas antes de la orden de quedarse en casa y el cierre del negocio no esencial y no se recuperó inmediatamente después del levantamiento de la orden de permanecer en casa.

Desde el comienzo de la pandemia de COVID-19, un número creciente de investigaciones han utilizado datos de agregadores financieros, TPV y tarjetas de crédito, y cuentas bancarias para analizar con datos semanales o diarios, la evolución del gasto total, por tipologías y por nivel de renta de los individuos. Además de los ejemplos comentados de Estados Unidos o Dinamarca también existen trabajos para el caso de Francia (Bounie *et al.,* 2020), Reino Unido (Hacioglu, Kanzig y Surico, 2020), Japón (Kubota, Koichiro y Toyama, 2020), Países Bajos (Golec *et al.,* 2020), China (Chen, Quian y Wen, 2020) o Portugal (Dos Santos, Carvalho y Peralta, 2020).

En España, Carvalho *et al.* (2020) utilizan las transacciones con tarjeta de crédito y TPV de BBVA (1.300 millones de transacciones) para medir la evolución del consumo por categorías. CaixaBank Research publica notas semanales que analizan el uso de tarjetas de crédito en TPV, transacciones *online* y retiradas de efectivo para estimar el impacto de las medidas de confinamiento en el consumo. Este tipo de información también es explotada habitualmente por otros bancos como Abanca (Observatorio Abanca por IESIDE) o el Banc Sabadell (Pulso)[5].

Para el caso de España, García-Montalvo y Reynal-Querol (2020) utilizan datos de 300 millones de transacciones en cuentas bancarias vinculadas por el agregador financiero Fintonic, una de las Fintech de gestión de finanzas personales más grandes de España, para analizar el efecto distributivo de la pandemia en el gasto. En este trabajo se analizan tres dimensiones de la heterogeneidad: ingreso, saldo de la cuenta y edad. García-Montalvo y Reynal-Querol (2020) muestran que, a diferencia de los resultados en otros países como EE. UU. (Cox *et al.,* 2020; Chetty *et al.,* 2020) y Reino Unido (Hacioglu, Känzig y Surico, 2020), no se encuentran diferencias significativas en la evolución del gasto agregado por nivel de ingresos: la caída y la recuperación del gasto es similar para todos los cuartiles de la distribución de la renta. Sin embargo, sí aparece alguna diferencia en la evolución del gasto por edad y por el saldo medio de las cuentas bancarias.

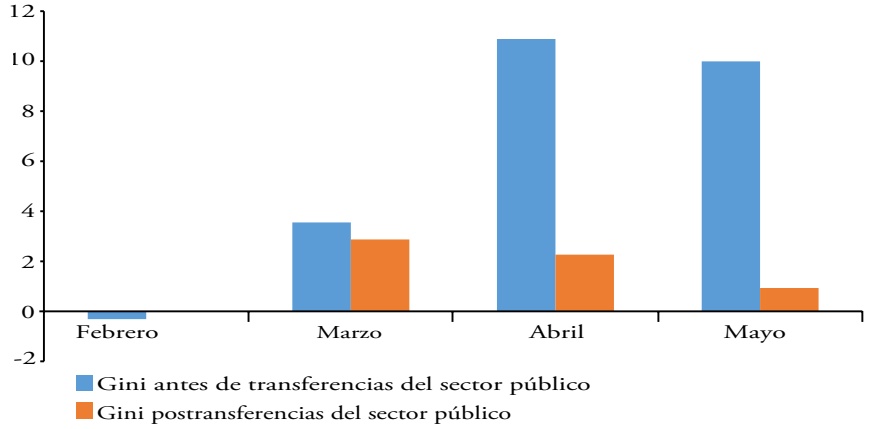
Una de las consecuencias habituales de las pandemias es el aumento de la desigualdad (Wade, 2020). En un trabajo muy diferente a los anteriores, Aspachs *et al.* (2020) describen

[5] El Banco de España también ha utilizado POS para rastrear el gasto durante la pandemia (Gonzalez, Urtasun y Perez, 2020).

una metodología para calcular la evolución de la desigualdad de los salarios utilizando datos de las cuentas bancarias. A diferencia de las muestras poco representativas que suelen ser habituales en los estudios que utilizan este tipo de datos[6], Aspachs *et al.* (2020) muestran que la distribución de los salarios de las nóminas depositadas en CaixaBank es prácticamente idéntica a la distribución de salarios de la *Encuesta de Estructura Salarial* del INE actualizada a 2020[7]. A partir de una base de datos de más de 3 millones de observaciones mensuales, Aspachs *et al.* (2020) calculan los movimientos mensuales en la distribución de los salarios (antes y después de considerar las políticas públicas) así como diferentes indicadores de desigualdad. La figura 3 muestra el enorme incremento del índice de Gini asociado al comienzo de la pandemia (más de 11 puntos) y la capacidad de los ERTE y la extensión del seguro de desempleo para moderar este incremento, aunque sin llegar a compensarlo totalmente. La granularidad de los datos permite analizar la evolución de la desigualdad por género, edad, nacionalidad, localización geográfica, etcétera.

Figura 3.

Cambio en la desigualdad salarial (Gini) en España entre febrero y mayo de 2020



Fuente: Aspachs *et al.* (2020).

5. PREDICIENDO PRECIOS DE INMUEBLES A NIVEL DE CÓDIGO POSTAL

Los cursos sobre Machine Learning suelen utilizar la predicción de precios residenciales como una de las aplicaciones más frecuentes. Supongamos que quisiéramos predecir los precios de distintos tipos de inmuebles (residenciales, comerciales, oficinas y suelo) en España a nivel de código postal a largo plazo. Contamos con información sobre los precios de

[6] Especialmente si se usan datos de agregadores financieros, como por ejemplo Gelman *et al.* (2014) o Baker (2018).

[7] Aspachs *et al.* (2020) también comprueban que la distribución de las variables demográficas (género y edad) son prácticamente idénticas a las observadas en las encuestas para los trabajadores por cuenta ajena.

compraventa trimestrales durante varios años en aquellos códigos donde se han producido transacciones así como de los precios de oferta y el número de transacciones en cada código postal. García-Montalvo *et al.* (2018) proponen un procedimiento basado en estimar la elasticidad respecto al crecimiento de la renta en cada código postal para cada tipo de inmueble[8]. Las predicciones pretendían utilizar la metodología más adecuada al problema sin perder de vista que la incertidumbre asociada a un ejercicio de esta naturaleza es enorme. García-Montalvo *et al.* (2018) aplican un procedimiento para hacer predicciones del PIB y los precios de la vivienda a largo plazo usando una combinación de VARs e inferencia a baja frecuencia (útil para predicciones a muy largo plazo). Las proyecciones de los precios agregados de la vivienda se supone que inciden en los cambios de los precios a nivel de código postal por analogía con un modelo financiero de valoración basado en un único factor.

Para la predicción a nivel de código postal se utiliza un modelo jerárquico (HML) que, posteriormente, permite agregar las predicciones a nivel municipal, provincia, regional e incluso nacional. Además de los precios agregados se incluye un factor idiosincrático asociado a la presión de la demanda en cada código postal. Se utilizan tres indicadores alternativos para cada código postal: la proporción de unidades en venta que disminuyeron su precio entre dos periodos consecutivos; la proporción de unidades que se mantienen de un periodo al siguiente; y la proporción de transacciones sobre las unidades ofertadas en los portales en el trimestre anterior.

En el modelo de precios residenciales tanto los parámetros del efecto del precio agregado como de la *proxy* de la presión de la demanda cambian para cada código postal. Esto implica que solo el modelo de precios residenciales supone estimar miles de parámetros.

5.1. Predicción agregada a largo plazo

El modelo básico es un vector autorregresivo bayesiano *(BVAR)* con variables $i \in 1:I$, periodos $i \in 1:T$ y retardos $l \in 1:L$)

$$y_t = \alpha + \sum_{l=1}^{L} \mathrm{B}_l y_{t-l} + \varepsilon_t, \quad \varepsilon_t \,|\, \Sigma_\varepsilon \sim \left(0, \Sigma_\varepsilon\right) \qquad [1]$$

donde B_l son $I \times I$ matrices de coeficientes Σ_ε y $I \times I$ es una matriz positiva definida con la covarianzas de los residuos[9]. Las variables incluidas en el vector son: PIB real, formación bruta de capital, consumo privado, tasa de desempleo, EONIA *(overnight),* IBEX 35 e índice de precios de la vivienda (IPV) del INE.

[8] El análisis incluye también predicciones para el número de transacciones y el precio de los alquileres aunque, en esta sección solo nos centraremos en el apartado de los precios de los inmuebles residenciales.

[9] Ver García-Montalvo *et al.* (2018) para los detalles sobre como se seleccionan las variables y los retardos a incluir en la especificación final usando un procedimiento de maximización de la distribución *a posteriori* (MAP).

El modelo anterior es útil para obtener predicciones de corto plazo, pero es menos apropiado para aprender la distribución estacionaria. Müller y Watson (2019) proponen un procedimiento para realizar predicciones a largo plazo del panel de procesos como una suma de un proceso Gausiano estacionario y un proceso de ruido idiosincrático, para luego estimar los momentos estacionarios del proceso Gausiano. Esta aproximación es efectivamente una descomposición espectral de un proceso Gausiano utilizando funciones trigonométricas.

La idea es proyectar las series temporales sobre un conjunto de basis functions $\psi_k(t) = \cos(\pi j\,(t\text{-}0{,}5)/T)$ considerando el modelo

$$y_t = \alpha + \mathrm{B}\psi_t + \varepsilon_t, \quad \psi_t = \sqrt{2}[\psi_1(t)\ \ldots\ \psi_K(t)]^\top, \quad \varepsilon_t \,|\, \Sigma_\varepsilon \sim (0, \Sigma_\varepsilon). \qquad [2]$$

Utilizamos $K = 12$ bases y la tasa de crecimiento anualizada entre trimestres, dado que las K bases capturan periodicidades mayores de $2T/K$ (Müller y Watson, 2019). Este modelo no pretende predecir la evolución temporal sino filtrar los componentes de baja frecuencia $\mathrm{B}\psi_t$ y usarlos en la estimación de las características estacionarias. Para simplificar el análisis, fijamos la prior conjugada en las columnas β_j de la matriz de correlación B:

$$\beta_j \,|\, \Sigma_\varepsilon \sim (0, g\Sigma_\varepsilon) \qquad [3]$$

Para predecir el proceso estacionario, $\tilde{\boldsymbol{y}}_t$, García-Montalvo *et al.* (2018) usan un modelo de error de dos componentes. El primer término, el error η_t, proviene de la variación a baja frecuencia explicada por las *basis functions,* mientras el segundo término de error ε_t representa la variación residual no estructurada.

$$\tilde{\boldsymbol{y}}_t = \alpha + \eta_t + \varepsilon_t, \qquad \varepsilon_t |\, \Sigma_\varepsilon \sim (0, \Sigma_\varepsilon) \qquad [4]$$

Podemos inferir la covarianza de η_t de $\mathrm{B}\psi_t$ utilizando las siguientes expresiones:

$$\eta_t|\mathrm{B} \sim (0, \Sigma_\eta\,(\mathrm{B})), \qquad \Sigma_\eta\,(\mathrm{B}) = T^{-1}\mathrm{B}\,(\textstyle\sum_t\,(\psi_t))\,\mathrm{B}^\top \approx T^{-1}\mathrm{B}\Psi\Psi^\top\,\mathrm{B}^\top. \qquad [5]$$

Por tanto, la distribución a largo plazo de $\tilde{\boldsymbol{y}}_t$ es

$$\tilde{\boldsymbol{y}}_t|(\alpha, \mathrm{B}, \Sigma_\varepsilon) \sim (\alpha, \Sigma_\varepsilon + T^{-1}\,\mathrm{B}\Psi\Psi^\top\,\mathrm{B}^\top) \qquad [6]$$

García Montalvo *et al.* (2018) explican cómo formular una versión Bayesiana de este modelo para combinar posteriormente con el BVAR. Esta combinación depende de los supuestos que se adopten sobre el panel VAR. Para evitar tener que hacer supuestos fuertes se puede adoptar una aproximación *ad-hoc* que directamente interpola la predicciones de corto y largo plazo. Específicamente se puede definir una combinación convexa de la distribución predictiva $f(\tilde{\boldsymbol{y}}_{t+s}|\boldsymbol{y}_t)$ dadas las observaciones en t y una distribución estacionaria $f(\tilde{\boldsymbol{y}}_{t+s})$:

$$g(\tilde{\boldsymbol{y}}_{t+s}|y_t) = e^{-\delta s} f(\tilde{\boldsymbol{y}}_{t+s}|y_t) + (1-e^{-\delta s}) f(\tilde{\boldsymbol{y}}_{t+s}) \quad [7]$$

donde el parámetro $\delta > 0$ determina la intensidad de la dinámica transicional. En este ejercicio se utiliza un $\delta = 0{,}1$ lo que implica que la dinámica de largo plazo empieza a tener relevancia a partir de aproximadamente el segundo periodo.

5.2. Modelo multinivel jerárquico

Antes de plantarse la estimación del modelo HML es necesario resolver la *sparcity* de información a nivel de código postal. Para producir un panel denso de tasas de crecimiento de precios de vivienda se utilizan dos procedimientos para agrupar observaciones de códigos postales: en primer lugar se usa un algoritmo basado en la proximidad de los códigos postales y un número mínimo de observaciones por clúster. En segundo lugar se utiliza una aproximación más formal que usa todas las variables disponibles a nivel de código postal (distancia entre códigos postales de una misma provincia, tensión del mercado, volumen de transacciones, etc.). Con estas variables se construyen 50 random forest (RF) de profundidad máxima 3 (para evitar sobreajuste) y se entrenan estos RF para predecir el precio por metro cuadrado. Se usan las 20 variables más importantes en media (entre los 50 RF) para entrenar un algoritmo K-means con el número de clústeres siendo función del número de transacciones en cada provincia.

Una vez clusterizados los códigos postales se estima un modelo de regresión lineal jerárquico,

$$y_{zt} = \boldsymbol{\beta}_z^{\top} \boldsymbol{x}_{zt} + \epsilon_{zt}, \qquad \epsilon_{zt} \sim (0, \tau^2) \quad [8]$$

donde y_{zt} es la respuesta en el código postal z en el periodo t, β_z es un J-vector de coeficientes pertenecientes al código postal z y x_{zt} es un vector de factores en el tiempo t. Dado que existe un gran número de parámetros, lo que podría producir sobreajuste y disminuir la predicción en la muestra de testeo, especialmente para códigos postales con pocas transacciones, se mezcla información de varios códigos postales via *priors* jerárquicas:

$$\beta_z \sim (\mu, \Sigma) \quad [9]$$

El modelo se completa con las siguientes *priors* para los *nuisance parameters:*

$$\tau^2 \sim (0{,}01, 0{,}01), \qquad p(\mu) \propto 1, \qquad \Sigma \sim (J, J\Pi) \quad [10]$$

En particular, la modelización de los precios residenciales incluye la variable comentada anteriormente (tensión del mercado), generando la siguiente especificación,

$$\Delta \log p_{zt} = \alpha_z + \beta_z \Delta \log h_t + \gamma_z s_z + \varepsilon_{zt} \quad [11]$$

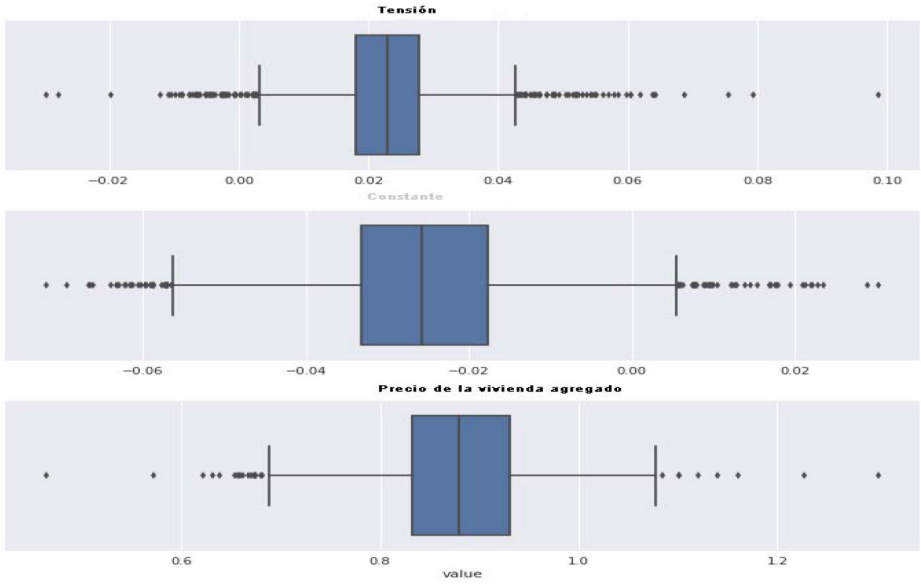
donde p_{zt} es el precio medio del metro cuadrado en el código postal z en t, h_t corresponde al índice IPV del INE en t, y s_z es la media de la tensión del mercado en cada código postal z[10]. Agregando los datos de los códigos postales se comprueba que el supuesto de linealidad respecto a la evolución del precio de la vivienda agregado es apropiado en este caso (ver García-Montalvo *et al.,* 2018).

Los resultados de la distribución *a posteriori* de los coeficientes, aparecen en el *box plot* de la figura 4.

La figura 5 muestra la distribución de los residuos versus los valores ajustados donde no parece haber desviaciones claras del supuesto de linealidad.

Figura 4.

Distribución *a posteriori* de los parámetros

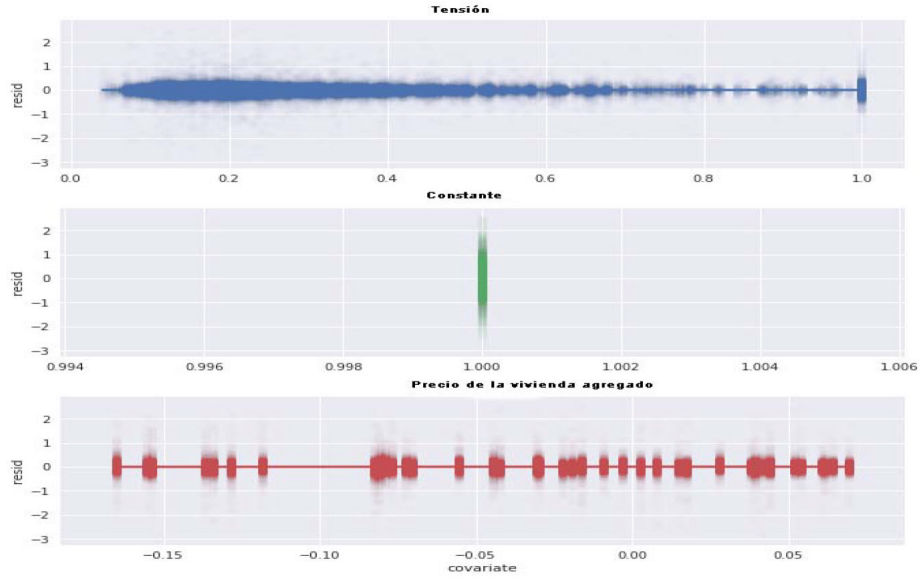


Fuente: Elaboración propia.

[10] La estimación del modelo utiliza un *Gibbs sampler* que proporciona resultados satisfactorios con muestras efectivas de unas 10.000 iteraciones. También se plantean tres escenarios en función de las bandas de confianza del 90 % alrededor de la predicción del crecimiento del precio de la vivienda agregado.

Figura 5.

Residuos



Fuente: Elaboración propia.

6. OTRAS APLICACIONES: LUCES Y ELECCIONES

Otras dos áreas donde se están multiplicando las aplicaciones son la utilización de imágenes de satélite como indicadores para el nivel de desarrollo o pobreza por áreas pequeñas y la predicción de resultados electorales.

6.1. Imágenes de satélite

La disponibilidad de gran cantidad de datos obtenida a partir de sensores e imágenes de satélites ha abierto también una gran ventana de oportunidad para utilizar estos datos en la medida del nivel de desarrollo, pobreza o desigualdad. Los primeros trabajos que utilizan la luz nocturna captada por satélites para medir el nivel de desarrollo o el crecimiento aparecieron unos años después de que se publicaran los primeros estudios de este tipo en las revistas sobre sensores e ingeniería de señales. Chen y Nordhaus (2011) y Henderson, Storeygard y Weil (2012) utilizan modelos muy similares. En Henderson, Storeygard y Weil (2012) la señal de la luz nocturna se utiliza como medida de actividad económica en un modelo de error de medida clásico donde tanto el crecimiento de la intensidad de la luz como el crecimiento del PIB están relacionados con el crecimiento de la renta verdadera no observable. A partir

de este modelo se derivan los pesos óptimos para minimizar el error de medida. Utilizando este procedimiento, Henderson, Storeygard y Weil (2012) muestran que la intensidad de la luz por km^2 tiene capacidad predictiva sobre la tasa de crecimiento real del PIB. Utilizando unos supuestos similares, Pinkovskiy y Sala-i-Martin (2016) rechazan la hipótesis de que el peso óptimo de las encuestas utilizadas para aproximar la renta per cápita sea superior al de las cuentas nacionales. Asimismo, muestran que las encuestas son peores que la contabilidad nacional para aproximar los resultados que importan para la población con menor nivel de renta en países en vías de desarrollo[11]. García-Montalvo y Reynal-Querol (2020) utilizan la luz nocturna para medir índices de desigualdad con el objetivo de encontrar un procedimiento que pueda permitir calcular índices de Gini en zonas donde, por su nivel de desarrollo o por la existencia de conflictos, no existen datos de encuestas ni contabilidad nacional que permitan hacer un cálculo mínimamente fiable del nivel de desigualdad. El procedimiento consiste en dividir la extensión del área deseada en pequeños píxeles y dividir la intensidad de la luz por el número de habitantes después de alienar las tramas. Este procedimiento genera suficientes observaciones para, con posterioridad, realizar un *ranking* de luz media por habitante de cada pixel. Por ejemplo, a nivel de país, la media de píxeles de los países es de 4,4 millones. Se aplican diferentes correcciones a los datos de luz nocturna para evitar los problemas del *top-coding* y de baja intensidad. Una vez calculados los cocientes y ordenados se calcula el índice de Gini. Para la validación cruzada de este procedimiento se utilizan los datos de países disponibles a partir de obtener un factor común a los estimadores disponibles en las diferentes bases de datos existentes. A partir de la correlación con dicho factor se calculan los pesos del Gini medido a partir de la luz.

6.2. Predicción de elecciones con nuevos partidos políticos

García-Montalvo, Papaspiliopoulos y Strumpf-Fetizon (2019) proponen una nueva metodología para predecir los resultados de elecciones en presencia de nuevos partidos políticos sin historial electoral, o con un historial muy limitado. Esta situación ha sido muy común en los países europeos desde 2008 con el avance de nuevos partidos de orientación populista. El procedimiento combina un modelo fundamental con todas las encuestas disponibles a través de un mecanismo bayesiano de síntesis de evidencia.

El modelo fundamental predice, a nivel nacional, la proporción de votos y escaños utilizando microdatos de las encuestas electorales del CIS así como datos demográficos. El modelo para las encuestas frecuentes tiene en cuenta no solo su nivel de incertidumbre sino también la calidad de la empresa/institución que produce los resultados de cada encuesta. El procedimiento utiliza todas las encuestas disponibles durante todas las elecciones que se producen con anterioridad al periodo para el cual se busca la generalización. La especificación incluye un sesgo invariante en el tiempo para cada empresa/institución que realiza encuestas (*house effect*), un sesgo invariante para cada empresa que se estima para cada elección sepa-

[11] Los análisis recientes no solo utilizan la luz nocturna. Por ejemplo Jean *et al.* (2016) complementan la luz nocturna con imágenes diarias sobre el tipo de tejados de las viviendas, por ejemplo, para complementar la información en el cálculo de la pobreza por áreas pequeñas.

radamente (*election effect*), un tendencia lineal con un coeficiente que puede cambiar para distintas elecciones pero es común a todas las empresas demoscópicas; y, finalmente, un error idiosincrático que es específico de cada encuesta debido a diferencias en la metodología utilizada y la variabilidad muestral. La hibridación de ambos modelos se produce a partir de sus probabilidades condicionadas. García-Montalvo, Papaspiliopoulos y Strumpf-Fetizon (2019) aplican este procedimiento a las elecciones españolas en 2015 y muestran que el procedimiento genera mejores predicciones, especialmente en términos de escaños, que especificaciones alternativas.

7. CONCLUSIONES: LECCIONES APRENDIDAS

Las técnicas de ML y la creciente disponibilidad de grandes bases de datos generan enormes posibilidades para el avance de la ciencia económica. Permiten analizar el impacto de las políticas económicas prácticamente en tiempo real y avanzar en modelizaciones cada vez más complejas apoyadas en millones de datos. Pero no son ni mucho menos una panacea. El *hype* actual del big data en economía tiene similitudes con lo sucedido a mitad de los años noventa con la revolución de los experimentos randomizados. La extensión de esta metodología llevó a un desprecio creciente de otros procedimientos estadísticos. Sin embargo, las técnicas empíricas dependen crucialmente del tipo de aplicación, la cantidad de datos disponibles y la calidad de los mismos. Además las técnicas no son excluyentes: por ejemplo, en la actualidad existen diversas modelizaciones que permiten combinar resultados utilizando big data o, en general, datos observacionales, y datos de experimentos[12].

En algunas ocasiones las técnicas de ML son esenciales como en el caso de la ingeniería reversa comentado en la tercera sección puesto que los datos han sido generados por algún tipo de algoritmo habitual en ML. También es importante cuando se intenta obtener un nuevo indicador que, bien por lo caro que puede resultar realizar encuestas o lo difícil que puede ser hacerlo para áreas pequeñas, es mejor aproximar utilizando otras fuentes de datos. Este es el caso del uso de las imágenes de satélite para aproximar el nivel de desarrollo o la desigualdad para áreas pequeñas o no delimitadas por fronteras administrativas. En el seguimiento de la economía en tiempo real, la utilización de datos administrativos es la mejor opción para avanzar por mucho que se puedan realizar encuestas por Internet con elevada periodicidad.

Sin embargo, en otros casos se plantean ejercicios que los datos disponibles difícilmente podrán responder. Este sería el caso de la predicción del crecimiento de los precios de la vivienda por código postal. Es cierto que la predicción es mejor que otros procedimientos alternativos pero, en cualquier caso, el nivel de error en la generalización es elevado. Por tanto, si el objetivo es mejorar otros procedimientos, o buscar una metodología con una buena justificación, es un ejercicio razonable pero, si el objetivo es tener una buena predicción, la conclusión es diferente.

[12] Imbens, Chetty y Athley (2020) presentan una posibilidad de realizar dicha combinación utilizando un supuesto de *latent unconfoundedness* para los datos observacionales.

Una de las características más importantes de las técnicas de ML es la insistencia en la comparación de diferentes hipótesis (especificaciones, modelos o técnicas) para determinar cuál es la que tiene mejores propiedades predictivas. Desafortunadamente en muchas aplicaciones económicas este ejercicio no se realiza y se toman las técnicas más complejas como las más apropiadas, sin tener en cuenta la importancia que el campo del ML otorga al problema del sobreajuste. Otro problema importante es la escasa tendencia de los economistas a mirar y limpiar los datos antes de realizar ningún análisis. Teniendo en cuenta la baja señal sobre ruido de muchas de las bases de datos masivas este defecto disciplinario se hace particularmente relevante. A pesar de mi insistencia resulta casi imposible que un ayudante de investigación mire los datos antes de empezar a hacer análisis. No es infrecuente encontrar que ni siquiera la carga de los datos se produjo de la forma esperada y ya se han realizado multitud de análisis.

Otra característica relevante del big data es la reutilización de datos creados con otros fines, y la fusión de estas bases de datos con el objetivo de realizar investigación. La posibilidad de fusionar bases de datos reutilizadas permite investigar temas que, de otra forma, no podrían ser tratados. García-Montalvo y Raya (2012) fusionan datos de diversas fuentes para conseguir una base de datos que contiene, para cada inmueble residencial, el precio de venta de mercado, el precio registral y el valor de tasación. De esta forma comprueban que a un crédito valor del 83 % le correspondían, durante la gran expansión inmobiliaria de los años 2000, un crédito sobre precio de compra medio en el entorno del 110 %. Esto, lógicamente, tiene implicaciones significativas en términos de política macroprudencial (García-Montalvo y Raya, 2018). Estos mismos datos permiten calcular la proporción de operaciones que implicaron algún tipo de pago opaco a la Hacienda Pública (García-Montalvo, Piolatto y Raya, 2020).

Asimismo, las técnicas de ML insisten en la capacidad predictiva de los procedimientos lo que implica que no importa si el método es una caja negra mientras el error de generalización sea pequeño. Sin embargo, como economistas nos interesa entender los motivos que explican los resultados empíricos que obtenemos más allá de la simple predicción. Esta es una área de investigación que será cada vez más importante en el futuro del data science, pues los sesgos de los algoritmos de aprendizaje en aplicaciones reales requieren cada vez más de auditorías de los procedimientos para comprender su origen e intentar resolverlo. Este es el caso de la discriminación en los sistemas de puntuación crediticia aunque se evite considerar la variable que describe las diferencias entre los grupos (por ejemplo, afroamericanos, inmigrantes o mujeres).

En el caso español la utilización de datos administrativos continúa en un claro retraso frente a otros países, especialmente los escandinavos. De hecho existe una clara animosidad hacia la investigación basada en big data como demostró la enorme polémica que generó el estudio del INE sobre movilidad utilizando el posicionamiento de teléfonos móviles. A pesar de que los datos se iban a agregar geográficamente y estaban anonimizados la opinión pública puso el grito en el cielo. Lo sorprendente es que la mayoría de los ciudadanos estén dispuestos a ceder todos sus datos por bajarse una aplicación de linterna al teléfono móvil y se escandalicen por la utilización de datos anonimizados para la consecución de un objetivo socialmente deseable a partir de un estudio de una institución pública. Es

necesaria una estrategia común para explicar la importancia de este tipo de estudios y todas las salvaguardias sobre privacidad que se aplican. Por ejemplo la AIReF ha realizado varias reuniones de expertos sobre estos temas y ha presentado un documento de opinión, AIReF (2020), con una propuesta para una estrategia nacional para avanzar en la disponibilidad de datos administrativos para la investigación sin ánimo de lucro que redunde en beneficio del interés público. En este sentido el anuncio del INE, la Agencia Tributaria, la Seguridad Social y el Banco de España de comenzar a trabajar conjuntamente en un sistema de acceso a sus bases de datos con fines científicos es una excelente noticia.

Finalmente, la creciente utilización de técnicas de ML en economía y finanzas ha multiplicado el número de revisiones de la literatura sobre técnicas de ML escritos por económetras. Desgraciadamente la gran mayoría son simplemente malos resúmenes de libros de texto clásicos en el ámbito del ML. Seguramente, como ya sucediera con cierta frecuencia en el pasado, los economistas acabaremos redescubriendo muchos de los resultados clásicos de las técnicas de ML.

En resumen, las técnicas de aprendizaje automático requieren de una serie de condiciones para que dicho aprendizaje sea factible. Esta situación no está garantizada para cualquier aplicación o conjunto de datos disponible por muy grande que sea el mismo.

Referencias

AIReF (2020). Opinión para una estrategia de acceso a datos administrativos. *Opinión,* 1/20.

Aspachs, O., Durante, R., Graziano, A., Mestres, J., García-Montalvo, J. y Reynal-Querol, M. (2021). Real-time inequality and the welfare state in motion: evidence from COVID-19 in Spain.

Athey, S., Chetty, R. e Imbens, G. (2020). *Using Experiments to Correct for Selection in Observational Studies.* Mimeo.

Autor, D., Cho, D., Crane, L., Goldar, M., Lutz, B., Montes, J., Peterman, W., Ratner, D., Villar, D. y Yildirmaz, A. (2020). An evaluation of the Paycheck Protection Program using administrative payroll microdata. *MIT Working Paper.*

Baker, S. (2018). Debt and the response of household income to shocks: validation and application of linked financial account data. *Journal of Political Economy,* 126(4), pp. 1504-1556.

Baker, S., Farrokhnia, R., Meyer, S., Pagel, M. y Yannelis, C. (2020). Income, liquidity, and the consumption response to the 2020 economic stimulus payments. De próxima aparición en *Review of Asset Pricing Studies.*

Bick, A. y Blandin, A. (2020). *Real time labor market estimates during the 2020 coronavirus outbreak.* Mimeo, Arizona State University.

Bounie, D., Camara, Y., Fize, E., Galbraith, J., Landais, C., Pazem, T. y Savatier, B. (2020). *Consumption dynamics and COVID fiscal stimulus strategy: evidence from France.* Mimeo.

Carvalho, V. M., Hansen, S., Ortiz, A. Garcia, J. R., Rodrigo, T., Rodriguez Mora, S. y Ruiz de Aguirre, J. (2020). Tracking the Covid-19 crisis with high-resolution transaction data. *CEPR Discussion Paper,* 14642

Chen, S., Igan, D. Pierri, N. y Presbitero, A. F. (2020). Tracking the Economic Impact of COVID-19 and Mitigation Policies in Europe and the United States. *IMF WP/20/125.*

Chen, X. y Nordhaus, W. D. (2011). Using Luminosity Data as a Proxy for Economic Statistics. *Proceedings of the National Academy of Sciences,* 108(21), pp. 8589-8594.

Chen, H., Quian, W. y Wen, Q. (2020). The impact of the COVID-19 pandemic on consumption: learning from high frequency transaction data. Mimeo.

Chetty, R., Friedman, J. N., Hendren, N., Stepner, M. y el equipo The Oppotunity Insights (2020). How did Covid-19 and stabilization policies affect spending and employment? A new real-time economic tracker based on private sector data. *NBER Working Paper,* 27.431.

Cicala, S. (2020). *Early Economic Impacts of COVID-19 in Europe: A View from the Grid.* Mimeo.

Cox, N., Ganong, P., Noel, P., Vara, J., Wong, A., Farrell, D. y Greig, F. (2020). Initial impact of the pandemic on consumer behavior: evidence from linked income, spending, and savings data. De próxima aparición en *Brookings Papers on Economic Activity.*

Dos Santos, J., Carvalho, B. y Peralta, S. (2020). *What and how did people buy during the Great lockdown? Evidence from electronic payments.* Mimeo.

García-Montalvo, J. (2014). El impacto del Big Data en los servicios financieros. *Papeles de Economía Española,* 114, pp. 43-59.

García-Montalvo, J., Muñoz, J. C. y Reynal-Querol, M. (2020). *Measuring inequality from above.* Mimeo.

García-Montalvo, J., Papaspiliopoulos, P., Rossell, D. O. y Strumpf-Fetizon, T. (2018). *Predicting real estate prices at zip code.* Mimeo.

García-Montalvo, J., Papaspiliopoulos, P. y Strumpf-Fetizon, T. (2019). Forecasting electoral outcomes with new parties' competition. *European Journal of Political Economy,* 59, pp. 52-70.

García-Montalvo, J., Piolatto, A. y Raya, J. M. (2020). Transaction tax evasion in the housing market. *Regional Science and Urban Economics,* 81, pp. 1-17.

García-Montalvo, J. y Raya, J. M. (2012). What is the right price of Spanish residential real estate? *Spanish Economic and Financial Outlook,* 1(3), September, pp. 22-29.

García-Montalvo, J. y Raya, J. M. (2018). Constraints on LTV as a macroprudential tool: a precautionary tale. *Oxford Economic Papers,* 70(3), pp. 821-845.

García-Montalvo, J. y Reynal-Querol, M. (2020). Distributional Effects of COVID-19 on Spending: A First Look at the Evidence from Spain. *Barcelona GSE Working Paper,* 1.201 September, 2020.

Gelman, M., Kariv, S., Shapiro, M. D., Silverman, D. y Tadelis, S. (2014). Harnessing naturally occurring data to measure the response of spending to income. *Science,* 345, 6193, pp. 212-215.

Golec, P., Kapetanios, G., Neuteboom, N., Ritsema, F. y Ventouri, V. (2020). *Consumtion during the COVID-19 pandemic: loackdown or fear? Evidence from transaction data for the Netherland.* Mimeo.

González, J., Urtasum, A. y Perez, M. (2020). Evolución del consumo en España durante la vigencia del estado de alarma: un análisis a paritr del gasto con tarjetas de pago. *Articulos Analiticos,* 3/2020.

Gourinchas, P. O. (2020). Flattening the pandemic and recession curves. *Mitigating the COVID Economic Crisis: Act Fast and Do Whatever, pp.* 31-40.

Hacioglu, S., Kanzig, D. y Surico, P. (2020). The distributional impact of the pandemic. *CEPR Discussion Paper,* 15.101.

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

Jean, N., Burke, M., Xie, .M., Davis, W. M., Lobell, D. B. y Ermon, S. (2016). Combining satellite imagery and Machine Learning to predict poverty. *Science,* 353(6301), pp. 790-794.

Kobota, S., Koichiro, O. y Toyama, Y. (2020). Consumption responses to COVID-19 payments: evidence from a natural experiment and bank account data. *Covid Economics,* 62, pp. 90-123.

Müller, U. K. y Watson, M. W. (2019). Low-Frequency Analysis of Economic Time Series. De próxima aparición en S. Durlauf, L. P. Hansen, J. J. Heckman y R. Matzkin (eds.), *Handbook of Economics,* Vol. 7. North Holland.

Olafsson, A. y Pagel, M. (2018). The liquid hand-to-mouth: evidence from personal finance management software. *Review of Financial Economics,* 31(11), pp. 4398-4446.

Pinkovskiy, M. y Sala-i-Martin, X. (2016). Lights, camera... income! Illuminating the national account-household surveys debate. *Quarterly Journal of Economics,* 131(2), pp. 579-631.

Sheridan, A., Andersen, A., Hansen, E. y Johannesen, N. (2020). Social distancing laws cause only small losses of economic activity during the COVID-19 pandemic in Scandinavia, *Proceeding of the National Academy of Sciences,* August, 1-6.

Wade, L. (2020). An unequal blow. *Science, 368,* 6492, pp. 700-703.

Xiong, T., Bao, Y., Hu, Z. y Zhang, L. (2015). A combination method for interval forecasting of agricultural commodity futures prices. *Knowledge-Based Systems,* 77, pp. 92-102.

CAPÍTULO II

Adelantando el consumo de las administraciones públicas: *big data* a través del *BOE*

Carlos Cuerpo Caballero*
Teresa Morales Gómez-Luengo

Los avances en las técnicas de análisis de big data, junto con la creciente disponibilidad de grandes repositorios de datos, están permitiendo novedosas aplicaciones en el campo de la macroeconomía, en especial en la previsión de los principales agregados macroeconómicos. Este artículo presenta una posible aplicación, a través de la previsión del consumo público para España utilizando la plataforma de contratación del sector público. El análisis recoge la práctica totalidad de la actividad contractual del sector público, cubriendo más de 1.185.337 licitaciones, realizadas desde 2018 por más de 15.000 órganos de contratación distintos.

Palabras clave: big data, consumo público, machine learning, contratación pública.

* Los autores agradecen las ideas y los comentarios de Israel Arroyo Martínez y Raquel Losada Muñoz. Las ideas reflejadas en el artículo corresponden a los autores y no representan a las instituciones en las que trabajan.

1. BASE DE DATOS: PLATAFORMA DE CONTRATACIÓN DEL SECTOR PÚBLICO

1.1. Definición y características generales

Los datos utilizados en este capítulo proceden de la plataforma de contratación del sector público, mantenida y puesta a disposición de los usuarios de forma gratuita por el Ministerio de Hacienda[1]. La plataforma representa la puerta de entrada a la actividad contractual del sector público, facilitando información relevante sobre las licitaciones públicas, desde su convocatoria, hasta los resultados de las mismas.

Su primera gran ventaja, además de ser pública, reside en su completitud, puesto que recoge las licitaciones del sector público entendido en sentido amplio[2], incluyendo, entre otras:

- La Administración General del Estado, las administraciones de las comunidades autónomas y las entidades que integran la Administración Local.
- Las entidades gestoras y los servicios comunes de la Seguridad Social.
- Los organismos autónomos, las entidades públicas empresariales, las universidades públicas y las agencias estatales, entre otras entidades de derecho público vinculadas a o dependientes del sector público.
- Las sociedades mercantiles en cuyo capital social la participación pública, directa o indirecta, sea superior al 50 %.
- Las mutuas de accidentes de trabajo y enfermedades profesionales de la Seguridad Social.

Las licitaciones registradas recopilan información sobre al menos seis dimensiones de interés. En primer lugar, sobre la duración del contrato, incluyendo fechas de publicación, adjudicación y número de meses de duración. En segundo lugar, sobre la tipología de los contratos, destacando las categorías de servicios, suministros, obras y concesión de obras, entre otras. En tercer lugar, información geográfica y sectorial sobre el órgano contratante. En cuarto lugar, información sobre el procedimiento mediante el cual se lleva a cabo la contratación, ya sea Abierto, Restringido, Negociado con y sin publicidad, Acuerdo Marco, Simplificado, etc. En quinto lugar, el tipo de tramitación, separando la Ordinaria de la Urgente y de la de Emergencia. En sexto lugar, información sobre la competencia entre licitadores, destacando el número de concurrentes y el precio de licitación y adjudicación final, que permite analizar las bajas o ahorros en el precio debidas al proceso de competencia entre empresas licitadoras.

[1] Puede accederse a través de su página web: https://contrataciondelestado.es/wps/portal

[2] Tal y como recoge el artículo 3.1 del Real Decreto Legislativo 3/2011, Texto Refundido de la Ley de Contratos del Sector Público.

La base de datos original incluye información desde 2012 si bien se ha seleccionado un periodo de análisis de 2018 a 2020[3]. Con esta muestra se dispone de información sobre 1.185.337 licitaciones, realizadas por más de 15.000 contratantes distintos. Para llegar a este número se ha realizado un proceso de depuración previa de la base de datos original, eliminando las entradas duplicadas[4], los valores extremos y errores detectados en los precios de adjudicación , duración y número de licitadores.

1.2. Análisis descriptivo

1.2.1. Tipo de contrato

La plataforma clasifica los contratos en función de su objeto en las siguientes categorías: (a) Servicios, (b) Suministro, (c) Gestión de servicios públicos, (d) Concesión de servicios, (e) Obras, (f) Concesión de obras públicas, (g) Privado, (h) Administrativo especial, (i) Colaboración entre sector público y privado, (j) Patrimonial y (k) Otros. Para facilitar la interpretación, las once categorías se agregan en tres más genéricas: (i) Consumo público, que incluye las cuatro primeras, (ii) Inversión pública, que incluye Obras y Concesión de obras públicas; y la categoría (iii) Otros, que incluye al resto.

Tal y como refleja la figura 1, los contratos cuyo objeto es el consumo público son mayoritarios, tanto en número como en presupuesto, seguido de los contratos de inversión y, en último lugar y con una presencia residual, del resto. En concreto, los contratos de consumo representan en torno al 90 % del total de los contratos, por un 8 % de los contratos de inversión. Sin embargo, la cuantía promedio de los contratos de inversión es mayor y esto hace que, en términos del presupuesto total anual (panel b) su peso relativo supere el 20 %.

1.2.2. Tipo de procedimiento

La plataforma clasifica los contratos también en función del procedimiento utilizado para su adjudicación. Se incluyen las siguientes categorías: (a) Abierto, (b) Abierto simplificado, (c) Restringido, (d) Negociado sin publicidad, (e) Negociado con publicidad, (f) Diálogo competitivo, (g) Normas internas, (h) Acuerdo marco, (i) Concurso de proyectos, (j) Asoc. para la innovación, (k) Sist. dinámico adquisición, (l) Licitación con negociación, (m) Otros, y (n) Menores.

---

[3] Los contratos menores están únicamente disponibles desde 2018. Para el resto de contratos se seleccionan licitaciones publicadas desde 2017 pero adjudicadas como pronto en 2018.

[4] Para cada licitación se encuentran varias entradas pues la información va actualizándose conforme se suceden las distintas fases en el proceso, desde la publicación original hasta la adjudicación. El número de entradas previo a esta depuración era de más de 2.150.000.

Como norma general, y con arreglo a la Ley 9/2017 de Contratos del Sector Público, los contratos que celebren las administraciones públicas se adjudicarán utilizando (a) el procedimiento abierto o (c) el procedimiento restringido.

En el caso de los procedimientos abiertos, toda empresa interesada podrá presentar una propuesta, sin posibilidad de negociación de los términos del contrato. Existe la posibilidad de recurrir a la modalidad (b) abierta simplificada en los contratos de obras, suministro y servicios cuando su valor estimado no supere unas cantidades determinadas legalmente y no haya criterios de adjudicación evaluables mediante juicio de valor que superen el 25 % del total[5]. Esta modalidad permite realizar las adjudicaciones en el plazo de un mes desde la licitación, agilizando el proceso.

En cuanto a los procedimientos (c) restringidos, solo podrán presentar proposiciones aquellas empresas que sean seleccionadas por el órgano de contratación en atención a su solvencia. Tal y como especifica la Ley 9/2017, este procedimiento está particularmente indicado para servicios intelectuales de especial complejidad.

El resto de procedimientos solo pueden darse en los casos previstos en la ley. Entre ellos, destacan los contratos (n) menores y los negociados sin (d) y con (e) publicidad. Los menores permiten realizar adjudicaciones de forma directa, sin previa licitación, hasta un importe máximo (15.000 euros en caso de servicios y suministros). En cuanto a los negociados, su principal característica radica en que las condiciones del contrato se negocian previamente con uno o varios licitadores. La ley habilita esta modalidad para los contratos de obras, suministros, servicios, concesión de obras y concesión de servicios cuando se cumplan ciertos supuestos tasados en los artículos 167 y 168, como que la prestación incluya soluciones innovadoras o así lo exija la complejidad de la prestación, entre otras. Por defecto, los procedimientos negociados se harán con publicidad de la licitación, excepto en los casos tasados en el artículo 168 de la Ley 9/2017, como por ejemplo aquellos procesos en los que no se hubiera presentado ninguna oferta[6].

Los menores son los contratos más comunes representando el 62 % del total, tal y como puede verse en la figura 2, panel a. Esta característica se mantiene tanto para los contratos de consumo público como para los contratos de inversión, representando el 66 % y el 40 %, respectivamente. La segunda categoría más importante la constituyen los contratos abiertos (18 %) y abierto simplificado (10 %). Si añadimos el procedimiento negociado sin publicidad, que representa el 5 % del total, quedarían recogidos el 95 % de los contratos en estas cuatro categorías.

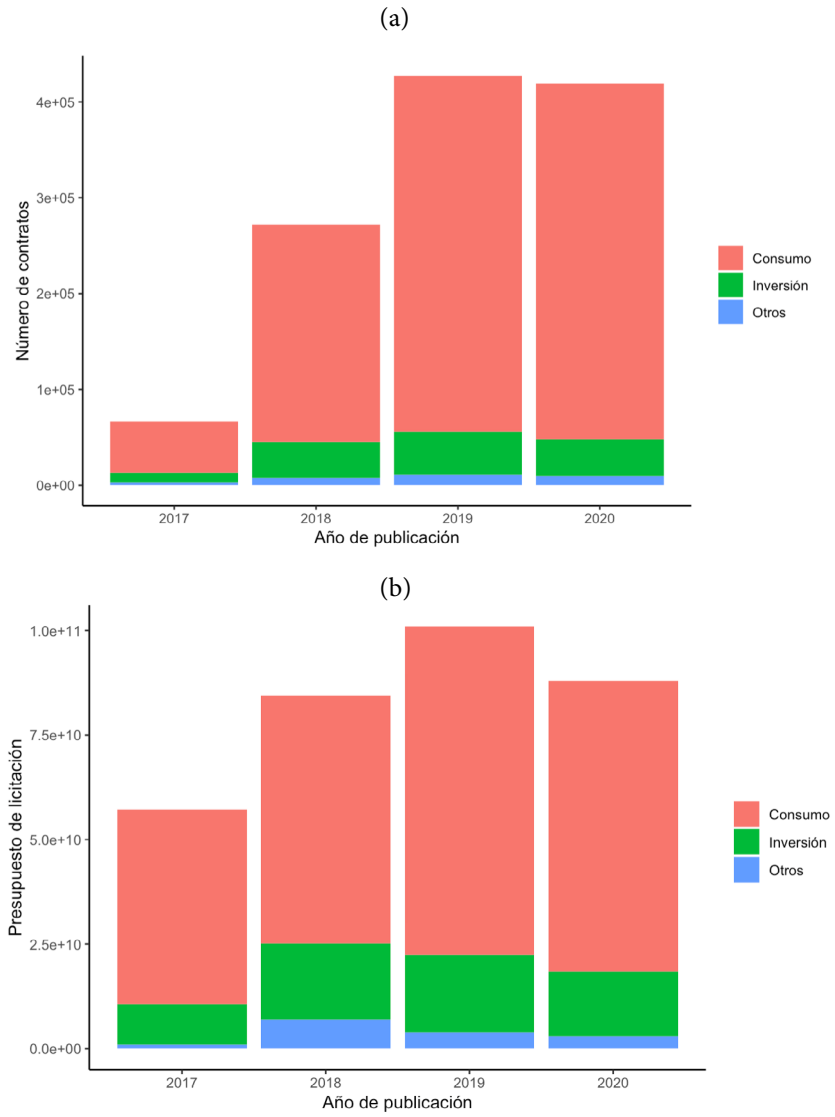
Pese a ser los contratos más comunes, los menores apenas suponen el 1 % del presupuesto movilizado, como refleja el panel b de la figura 2. Los contratos abiertos reinan en esta clasificación, alcanzando el 69 % de la cantidad total movilizada. Este liderazgo se mantiene

[5] Salvo para los contratos de prestaciones de carácter intelectual, en que su ponderación no podrá superar el cuarenta y cinco por ciento del total.

[6] Para más detalles sobre los distintos procedimientos, ver Royo (2018).

Figura 1.

Número y presupuesto de los contratos, por tipología y por año de publicación

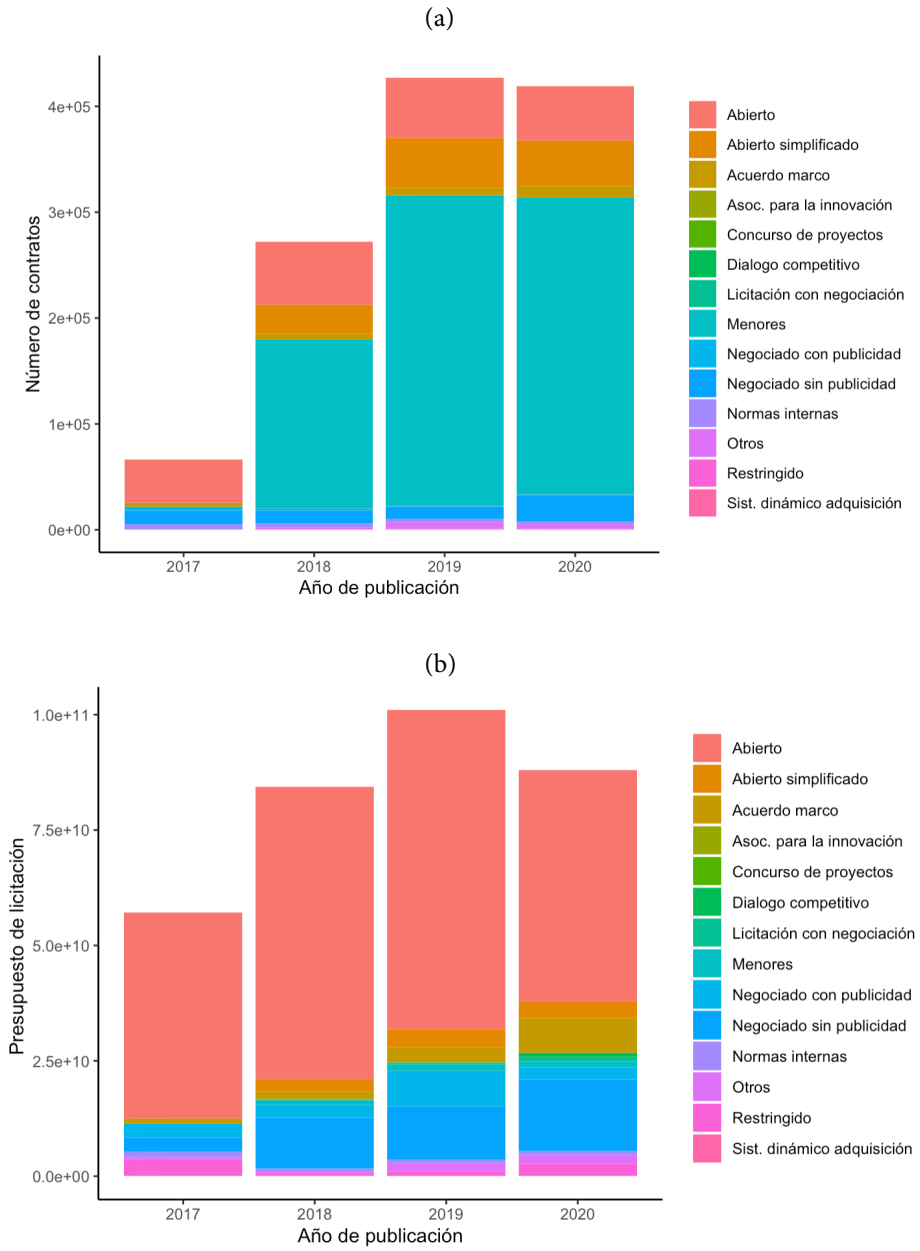


Fuente: Elaboración propia y plataforma de contratación del sector público.

tanto para los contratos de consumo como para los de inversión. Los contratos negociados se encuentran en segundo lugar de importancia en esta clasificación, con un 17 % del total movilizado. Cabe destacar la importancia de las licitaciones abiertas simplificadas en el caso de los contratos de inversión, llegando a representar un 12 % de los mismos.

Figura 2.

Número y presupuesto de los contratos, por procedimiento y por año de publicación



Fuente: Elaboración propia y plataforma de contratación del sector público.

1.2.3. Tipo de tramitación

La Ley 9/2017 prevé, más allá de (a) la tramitación ordinaria, dos mecanismos para acelerar las licitaciones: la tramitación de urgencia y (c) la tramitación de emergencia.

La tramitación de urgencia, regulada por el artículo 119 de la Ley 9/2017 supone la reducción de los plazos de licitación, adjudicación y formalización a la mitad. Solo será de aplicación a aquellos contratos cuya celebración responda a una necesidad inaplazable o que sea preciso adjudicar de forma acelerada por causa de interés público.

La tramitación de emergencia, regulada en el artículo 120, se limita excepcionalmente para aquellas situaciones en que las administraciones públicas tengan que actuar de manera inmediata como consecuencia de acontecimientos catastróficos, de situaciones que supongan grave peligro o de necesidades que afecten a la defensa nacional. Ante estas circunstancias de extrema gravedad, el órgano de contratación podrá ordenar la ejecución sin obligación de tramitación del expediente.

Por norma general, el procedimiento ordinario es el más común superando el 70 % del total en términos de número de contratos y el 60 % en términos de presupuesto movilizado. La tramitación de emergencia se ha limitado a menos del 1 % del total en años ante-

Figura 3.

Número y presupuesto de los contratos, por tramitación y por año de publicación

(a)

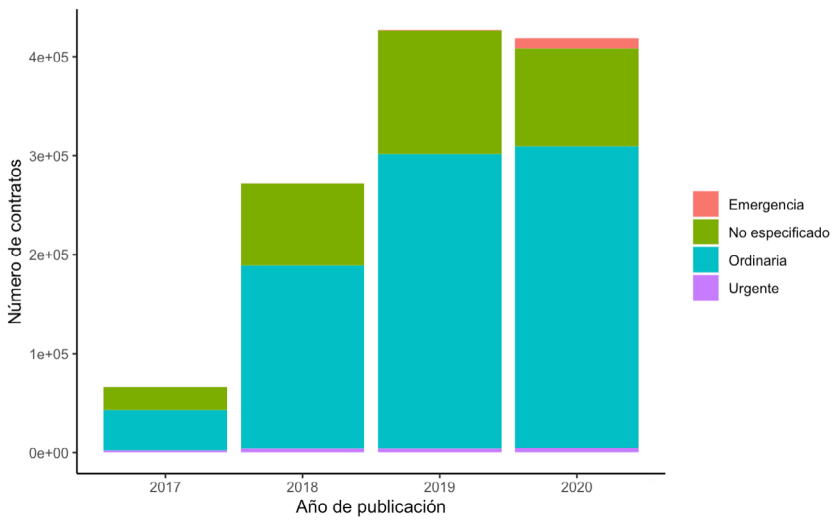
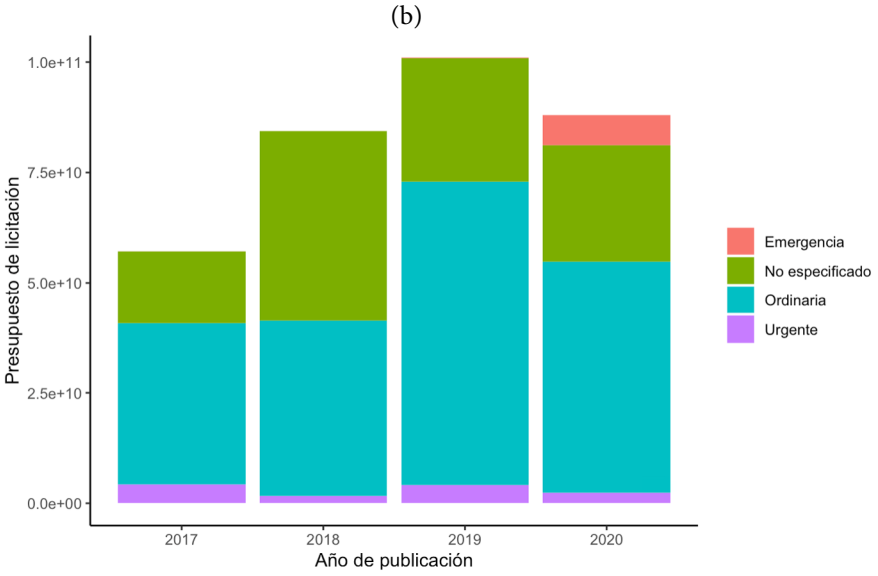


Figura 3. *(continuación)*

Número y presupuesto de los contratos, por tramitación y por año de publicación



Fuente: Elaboración propia y plataforma de contratación del sector público.

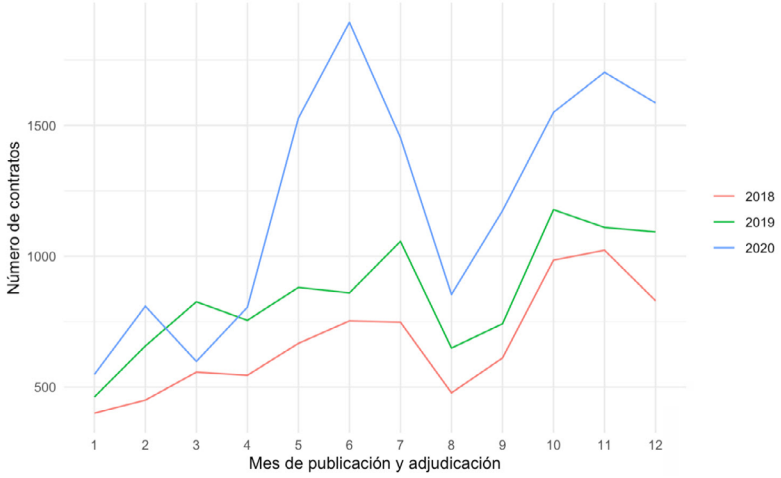
riores a 2020. Sin embargo, la declaración del estado de alarma como consecuencia de la COVID-19 ha supuesto un incremento sustancial del peso de los contratos de emergencia en 2020, pasando a representar el 3 % del total de contratos de consumo y el 9 % del presupuesto de los mismos.

Esta circunstancia puede verse también en la figura 4, que recoge el número de contratos adjudicados y publicados en el mismo mes en 2018, 2019 y 2020. Puede observarse como en el año 2020 en los meses de abril, mayo y junio, existe un repunte en los contratos publicados y adjudicados dentro del mismo mes (principalmente a través de la utilización de los procedimientos de emergencia).

La importancia de la COVID-19 en la contratación de emergencia puede verse a través de un análisis de la frecuencia de las palabras más repetidas en el objeto del contrato. Si miramos la totalidad de los contratos de consumo en 2020 (panel a, figura 5), destaca la importancia de los contratos de mantenimiento, o de equipamiento informático, tanto *hardware* como *software*. Si por el contrario nos centramos en los procedimientos de emergencia, aparece la motivación de la pandemia con los suministros relacionados con el material de protección frente al virus como principal objeto.

Figura 4.

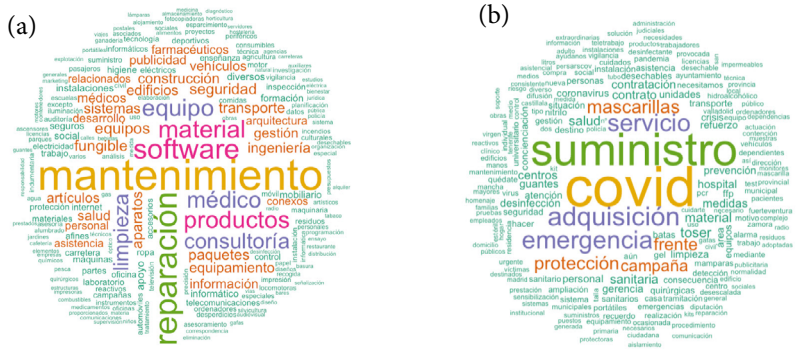
Evolución mensual del número de contratos adjudicados en el mismo mes
(Tiempo de adjudicación = 0)



Fuente: Elaboración propia y plataforma de contratación del sector público.

Figura 5.

Frecuencia de palabras en contratos de consumo y contratos de consumo de emergencia en 2020



Fuente: Elaboración propia y plataforma de contratación del sector público.

1.2.4. Concurrencia, baja de precio en la adjudicación y duración

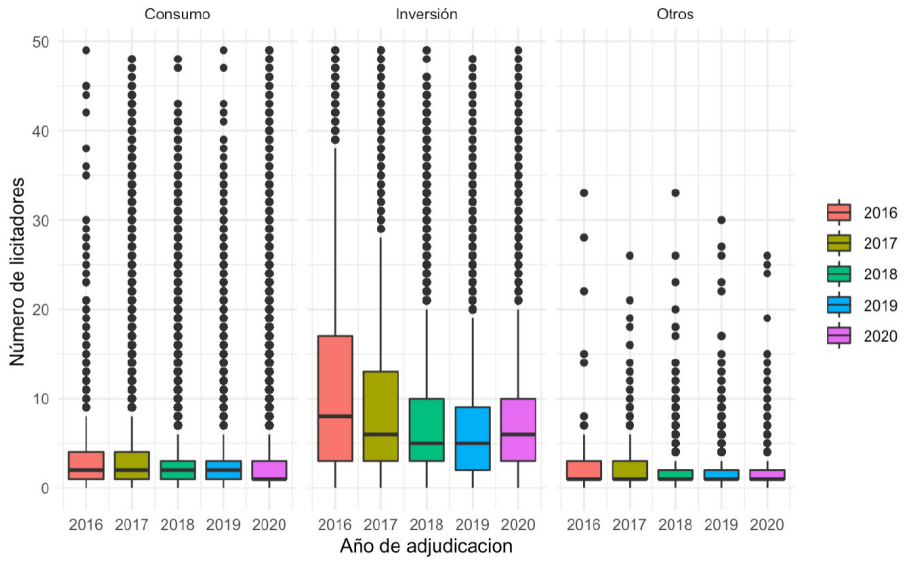
La competencia en los procesos de licitación permite optimizar la eficiencia en el uso de los recursos públicos, disminuyendo el riesgo de prácticas lesivas como la colusión entre operadores o la corrupción[7].

[7] Ver CNMC (2011) para una visión más profunda de la importancia de la competencia como principio inspirador de la normativa en materia de contratación pública.

La figura 6 refleja la situación en términos de concurrencia por tipo de contrato, excluyendo los contratos menores, menos relevantes en términos de cuantía presupuestada. Destacan los contratos de inversión con un promedio de más de siete licitadores, muy por encima de los de consumo, que en su mayoría se deciden entre tres empresas.

Figura 6.

Número de licitadores por año y tipo de contrato
(Contratos no menores)



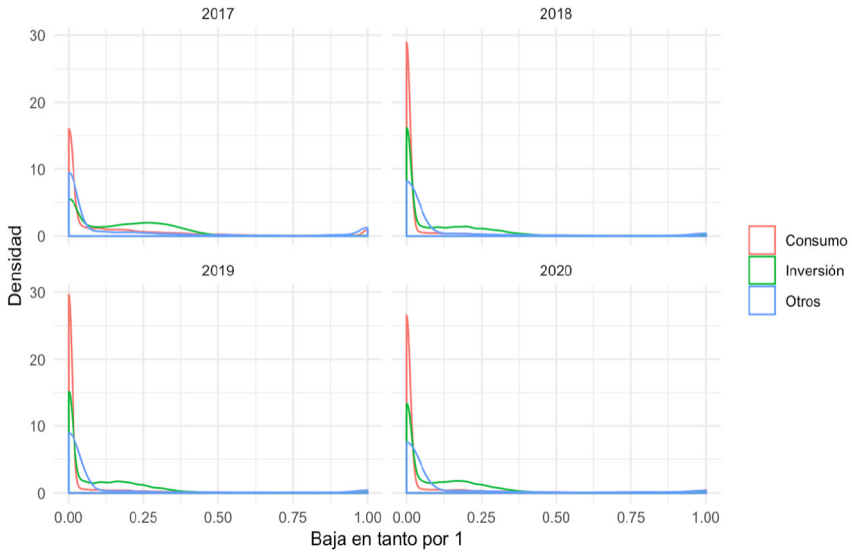
Fuente: Elaboración propia y plataforma de contratación del sector público.

La baja de precio desde el precio presupuestado o licitado hasta el precio final de adjudicación, depende también del tipo de contrato, como puede verse en la figura 7 y, por lo tanto, de la concurrencia en la licitación. Los contratos de inversión, de mayor concurrencia, presentan también una mayor baja porcentual en el precio, del 10 %, frente a los de consumo, que apenas llegan al 5 % en promedio.

La duración de los contratos sigue ciertos patrones comunes a lo largo de los años. En primer lugar, se observa una concentración de los contratos de obras por cortos períodos de tiempo (menores de seis meses). Los contratos de consumo se concentran en periodos más largos y tienen una periodicidad marcada por la frecuencia anual, fecha clave también para la renovación de los mismos (ver figura 8). Finalmente, se observan algunos cambios en 2020 en comparación con el resto de años, como un aumento de la densidad en los contratos menores de tres meses, sobre todo para las de consumo.

Figura 7.

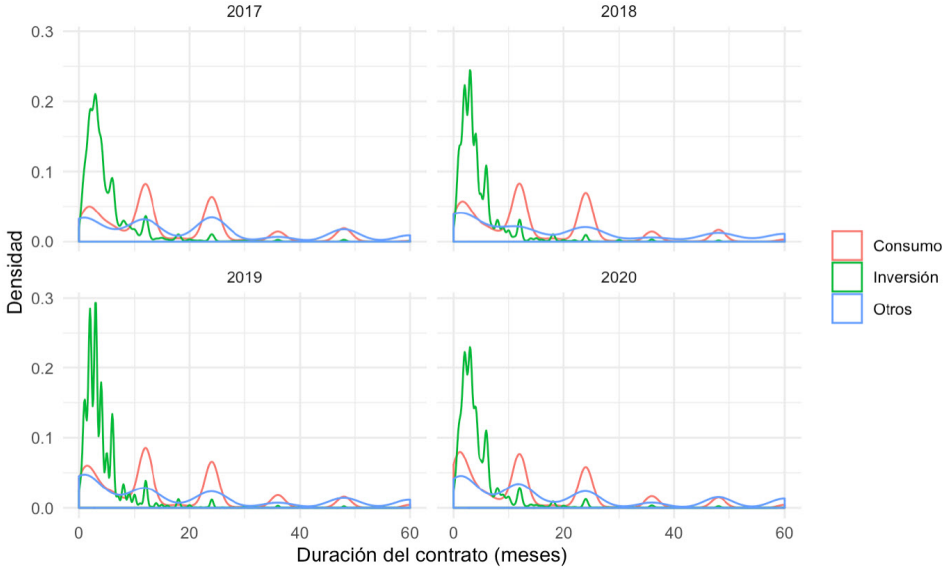
Densidad de las bajas de todos los contratos



Fuente: Elaboración propia y plataforma de contratación del sector público.

Figura 8.

Densidad de la duración de contratos no menores



Fuente: Elaboración propia y plataforma de contratación del sector público.

2. EJERCICIO DE PREDICCIÓN DEL CONSUMO PÚBLICO

2.1. Planteamiento del ejercicio

El objetivo final de este ejercicio es utilizar la plataforma de contratación descrita en el apartado anterior para realizar una previsión del consumo público en tiempo real.

El consumo público es uno de los componentes más importantes del producto interior bruto (PIB), suponiendo en torno a un 20 % del mismo en promedio. Su provisión se encuentra altamente descentralizada, correspondiendo un 60 % a las comunidades autónomas (educación y salud) y apenas un 20 % a la Administración Central. La remuneración de asalariados supone en torno al 60 % del total, seguido de los consumos intermedios (un 25 %), y de otros elementos como las ventas, el consumo de capital fijo, otros impuestos sobre la producción y las transferencias sociales en especie (TSE), adquiridas en el mercado. Esta última categoría puede llegar a presentar hasta el 15 % del total[8]. La disparidad de conceptos se traduce en una elevada heterogeneidad en sus determinantes principales, que pueden ir desde factores tan estructurales como la población para el componente salarial, hasta variables de coyuntura o el *stance* de política fiscal.

Estas características dificultan la previsión del consumo público con modelos estructurales, particularmente de sus componentes más pequeños y volátiles, como las TSE adquiridas en el mercado.

Para sortear estas dificultades y poder realizar una previsión del gasto público en tiempo real se plantea la posibilidad de proyectar el gasto comprometido utilizando los datos disponibles en la Base de Datos del Sector Público en el momento de la publicación de cada licitación. Esto permitiría hacer una previsión mensual que incorpore los contratos licitados hasta ese mes, independientemente de que se hayan adjudicado o no. Como hemos visto, la información disponible acerca de cada licitación es muy amplia, incluyendo información sobre el tipo de procedimiento, el objeto del contrato, el sector, la geografía, etcétera.

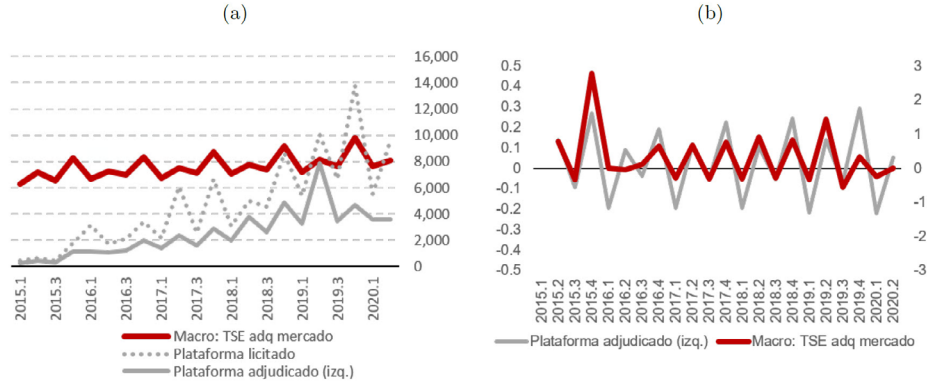
Sin embargo, si lo que se busca es proyectar el gasto en el momento de la licitación, existen dos datos fundamentales con los que no contamos: el precio de adjudicación, que aparecerá registrado en la mayoría de los casos en un momento posterior y el tiempo que tardará la licitación en adjudicarse, que puede ser muy variable en función de diversas circunstancias. Estas dos variables son fundamentales a la hora de conseguir un buen ajuste entre las licitaciones y el consumo público. En efecto, tal y como se observa en la figura 9 (panel a), la evolución de las TSE y la información de los contratos de consumo licitados sigue un perfil marcadamente distinto. Sin embargo, una vez que incorporamos la información final sobre la adjudicación, la dinámica de ambas series se vuelve altamente correlacionada, como puede verse en la evolución de sus tasas de crecimiento en el panel de la derecha.

[8] Para un análisis completo y pormenorizado del consumo público y sus componentes ver Losada (2017).

Figura 9.

Contratos de consumo como *proxy* de las transferencias sociales en especie adquiridas en mercado

(M€, panel a) y (% variación trimestral, panel b)



Fuentes: INE, elaboración propia y plataforma de contratación del sector público.

2.2. Metodología

Lo que planteamos en esta fase de nuestro estudio es analizar si dichos factores de incertidumbre (baja porcentual resultante en la adjudicación y tiempo de adjudicación) pueden ser estimados utilizando modelos de aprendizaje automático.

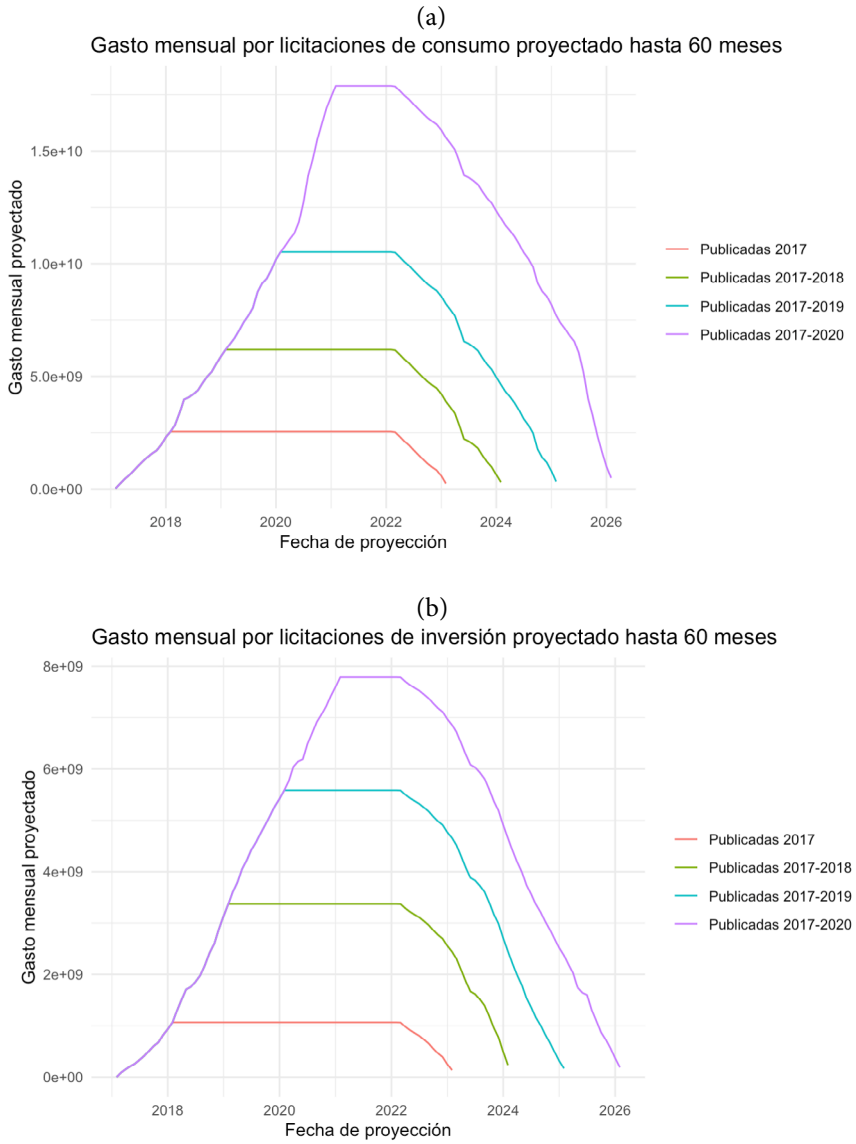
A continuación explicamos en primer lugar, los resultados que obtenemos cuando hacemos un ejercicio de proyección mensual de los montantes licitados que no tiene en cuenta los elementos de incertidumbre relacionados con la adjudicación. En segundo lugar explicamos como ir más allá, utilizando modelos de aprendizaje automático que nos permitan predecir el tiempo de adjudicación y la baja porcentual prevista cuando se publica cada licitación. En un ejercicio posterior habría que incorporar estas predicciones a la proyección de gasto público junto con otros elementos que pudieran incorporarse con datos como los de subvenciones, remuneraciones de los empleados públicos, etcétera.

2.2.1. Ejercicio de proyección del gasto

En la figura 10 se muestra una proyección del gasto mes a mes realizada para los contratos clasificados como consumo (panel a) y los contratos clasificados como inversión (panel b). Como vemos en los gráficos se realiza un ejercicio de proyección diferente, que incluye la información disponible desde 2017 hasta el año indicado. Es decir, la línea más alta muestra la proyección del gasto de todos los contratos publicados entre 2017 y 2020, la siguiente más baja los publicados entre 2017 y 2019 y así sucesivamente.

Figura 10.

Proyección del consumo con la información disponible sobre presupuesto de licitación y duración del contrato



Fuente: Elaboración propia y plataforma de contratación del sector público.

Para realizar esta proyección se calcula el gasto mensual asociado a cada contrato y se proyecta ese gasto mensual durante el plazo de duración del mismo. De este modo se está asu-

miendo que todos los contratos se ejecutarían mes a mes y que se empezaría a ejecutar el gasto comprometido al mes siguiente de la publicación de la licitación. Además, por simplicidad, se fija un límite temporal para la proyección de cada contrato de 60 meses.

Como podemos observar, la información disponible en la base de datos de contratación del sector público nos permite identificar una cuantía considerable de gasto ya comprometido en ejercicios futuros, tan lejanos como 2026. Además, con la información disponible a día de hoy, es posible prever un incremento del gasto mensual ya comprometido que alcanza su volumen máximo a finales de 2021 y no comienza a descender hasta finales de 2022. Esto es así tanto en contratos de consumo como en contratos de inversión.

Por otro lado, conforme se replica este ejercicio año a año se observan patrones comunes que parecen permitir identificar anomalías en la evolución del gasto. Así, la distancia entre las curvas de licitaciones de consumo publicadas hasta 2017, 2018 y 2019 es homogénea y contrasta con un salto que parece producirse en la curva de consumo de los contratos publicados hasta 2020. Este salto podría indicar un incremento considerable del gasto en consumo probablemente asociado, como hemos visto, a los contratos asociados a la crisis de la COVID-19.

2.2.2. Modelos de machine learning

Las proyecciones del montante licitado, sobre el que se ha practicado una reserva de crédito son sin duda informativas. Sin embargo, como decíamos, resultaría interesante poder contar con una predicción más cercana a la realidad del gasto que se va a ejecutar utilizando predicciones del tiempo de adjudicación y del montante de adjudicación (o baja presentada por la empresa adjudicataria). Para ello planteamos utilizar modelos de aprendizaje automático que, gracias al uso de una gran volumetría de datos y variables, pueden ser más precisos que los modelos de regresión estadística tradicionales.

En esta primera fase de nuestro análisis hemos utilizado el modelo de *Random Forest regression*. Se trata de un modelo relativamente sencillo en comparación a las redes neuronales o *deep learning,* pero requiere de menos datos y ha mostrado ser igual de efectivo en algunos casos, gracias a la agregación de muchos modelos sencillos entrenados con submuestras. En particular el Random Forest lleva a cabo una agregación de árboles de decisión que utilizan muestras distintas extraidas de la muestra principal por muestreo aleatorio con reemplazo (*bagging*)[9].

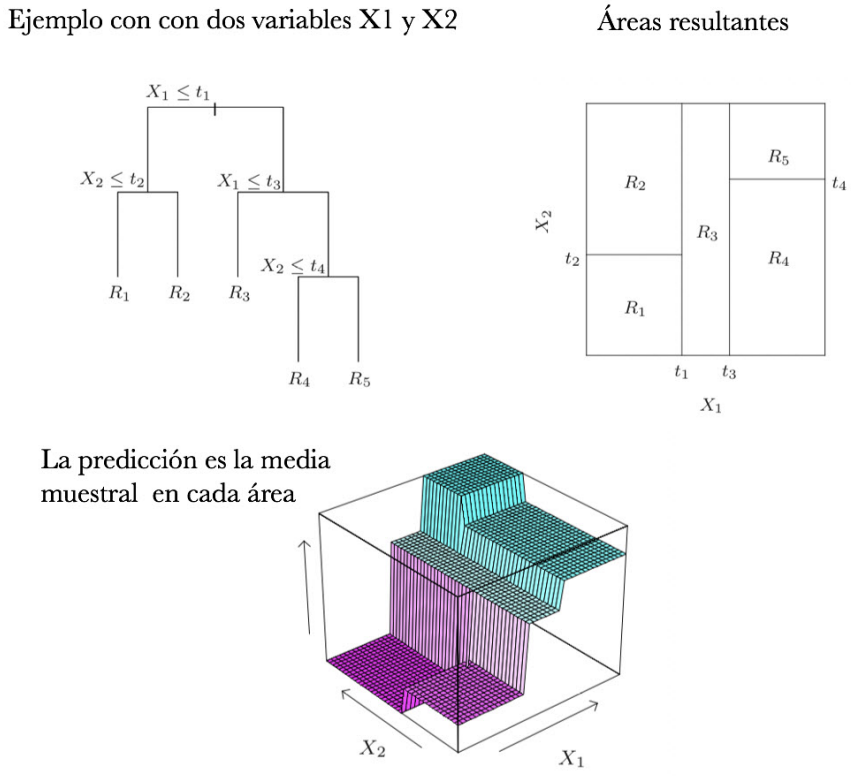
Los árboles de decisión son modelos simples que permiten ir clasificando la muestra en grupos basados en reglas de decisión binarias. Así, cada árbol de decisión elige la variable que permita dividir dos grupos de la muestra con la máxima diferencia en la variable objetivo

[9] Ver Breiman (1996) para una descripción detallada de por qué la técnica de *bagging* permite obtener buenos resultados.

posible. Para cada uno de esos grupos repite el ejercicio y así sucesivamente, hasta contar con una predicción basada en la media de una submuestra lo suficientemente pequeña de características similares. A continuación se presenta un ejemplo gráfico de cómo funciona un árbol de decisión para un caso de regresión:

Figura 11.

Representación de un árbol de decisión



Fuente: Hastie,Tibshirani y Friedman (2001).

Si se permite que el algoritmo persiga una predicción perfecta, el árbol de decisión extendería sus hojas hasta el límite de manera que, en cada extremo, solamente habría una observación y la predicción sería igual a la variable objetivo. Sin embargo, el árbol así entrenado no funcionaría para predecir muestras que no hubiera observado (muestra de validación o cualquier muestra que se encuentre cuando entre en producción). Para evitar este efecto, que se denomina *overfitting* se deben introducir elementos de regularización que permiten simplificar el algoritmo resultante para que sea generalizable a otras observaciones. Entre las técnicas que permiten regularizar los árboles de decisión se encuentran el denominado *tree*

prunning o limitación de la profundidad del árbol, la fijación de un número mínimo de muestras previo a la división en un nodo o la limitación del número de variables utilizadas para la división de las ramas (ver James *et al.,* 2014).

Como decíamos, el algoritmo de Random Forest, lo que hace es entrenar un número elevado de árboles de decisión. Como resultado de dicho entrenamiento con muestras diferentes, cada árbol de decisión habrá utilizado variables distintas y tendrá un número diferente de nodos de decisión. La predicción final es la media de las predicciones de todos los árboles que, por agregación, ha demostrado tener una precisión mucho más alta que un árbol de decisión simple individual.

Para construir un modelo de machine learning que permita realizar una predicción se utilizan técnicas de entrenamiento de los modelos con el objetivo de elegir los valores de sus parámetros e hiperparámetros. En el caso del Random Forest las elecciones que deben extraerse de este ejericio son la secuencia de variables y valores límite que permitirá ir dividiendo la muestra para cada árbol, el número de árboles que utilizará el algoritmo y las variables de regularización a las que nos hemos referido como la profundidad máxima del árbol o el tamaño mínimo de muestra de entrenamiento para la división.

Para entrenar el modelo se selecciona de forma aleatoria una muestra de entrenamiento con el 70 % de las observaciones y una muestra de validación con el 30 % restante. Adicionalmente, para asegurar que se seleccionan los parámetros e hiperparámetros que permiten generalizar mejor el modelo se utiliza la técnica de validación cruzada (*cross validation*). Esta técnica consiste en dividir la muestra de entrenamiento en *n* partes y entrenar el modelo en *n* iteraciones, dejando fuera cada vez una parte distinta. Después de entrenar cada modelo posible con las *n* iteraciones se elige el modelo que menor error cuadrático medio tiene en promedio. En nuestro caso la muestra de entrenamiento se ha dividido en tres partes por lo que cada uno de los modelos que se han probado se han entrenado tres veces, dejando fuera del proceso de entrenamiento cada vez una parte distinta, con la que se calcula el error cuadrático medio de predicción. Para cada modelo que se quiere considerar, se obtienen tres errores cuadráticos medios estimados, correspondientes a las tres iteraciones, con los que se calcula el promedio global para determinar qué modelo es el más adecuado.

Por último, otra técnica habitual en machine learning es considerar un amplio abanico de valores posibles para los hiperparámetros con el objetivo de elegir el modelo que mejor predicción haga cuando se enfrente a la realidad desconocida. Para ello, hemos utilizado una búsqueda aleatoria, que permite entrenar un número razonable de modelos al tiempo que se contemplan miles de combinaciones posibles ya que se van eligiendo hiperparámetros lo suficientemente diferentes de entre las opciones iniciales que se planteen.

2.3. Resultados

Se han construido dos modelos para predecir, por un lado, el tiempo de adjudicación previsto y por otro, la baja porcentual aplicada a una licitación. Las variables utilizadas son

el presupuesto de licitación, la duración del contrato y variables ficticias para cada categoría de código CPV, tipo de contrato (clasificación detallada y agregada), Comunidad Autónoma, tipo de procedimiento, tipo de tramitación y mes de publicación. Considerando cada variable ficticia por separado el número total de variables utilizadas es de 5.565.

El tamaño total de la muestra que excluye observaciones con valores omitidos en alguna de esas variables o en la variable objetivo es 75.403 para el modelo de tiempos de adjudicación y 194.941 para el modelo de bajas porcentuales. Se utiliza validación cruzada en bloques de tres y múltiples combinaciones de hiperparámetros.

Los modelos seleccionados a través del procedimiento de validación cruzada con búsqueda aleatoria de hiperparámetros tienen las siguientes características[10]:

a) Modelo de predicción ganador para tiempos de adjudicación

Número de árboles de decisión	100
Número mínimo de observaciones en un nodo para poder dividirlo	446
Número máximo de observaciones utilizado en cada árbol	25.000
Número máximo de variables consideradas para la división	1.000
Máxima profundidad de cada árbol (niveles)	70

b) Modelo de predicción ganador para baja porcentual

Número de árboles de decisión	2.000
Número mínimo de observaciones en un nodo para poder dividirlo	8.000
Número máximo de observaciones utilizado en cada árbol	72.500
Número máximo de variables consideradas para la división	4.000
Máxima profundidad de cada árbol (niveles)	200

Para evaluar estos modelos elegidos en el proceso de entrenamiento se utiliza el 30 % de la muestra reservada como muestra de validación. Por tanto, se trata de datos que el modelo nunca ha procesado por lo que permiten aproximar el error que se obtendría si se utilizara el modelo con datos nuevos, no disponibles hasta la fecha. Los resultados muestran que el modelo construido para la predicción del tiempo de adjudicación es muy útil, mientras que el modelo de predicción de las bajas porcentuales presenta una peor calidad.

El modelo de predicción de tiempos de adjudicción permite predecir cuándo será adjudicado un contrato con un error de en torno a un mes. En particular, la raíz del error cuadrático medio (*RMSE,* por sus siglas en inglés) es de 1,23 meses, muy por debajo del error que obtendríamos con un modelo de regresión lineal con regularización de *Ridge* (2,52

[10] Para la construcción de los modelos se ha utilizado la librería Scikit-learn de Python (ver Pedregosa *et al.,* 2011).

meses)[11]. El error cuadrático medio pondera más las predicciones muy alejadas de la realidad entre las que es posible que se encuentren valores atípicos. Para evitar que los atípicos pesen demasiado, se suele considerar también el error absoluto medio (*MAE,* por sus siglas en inglés) que da el mismo peso a todas las predicciones. El error absoluto medio para el modelo de predicción de tiempos de adjudicación es de tan solo 0,5 meses, frente a un valor de 0,7 para la regresión lineal con regularización de Ridge. Por lo tanto, en general se considera que este modelo de predicción de tiempos de adjudicación es bastante satisfactorio.

En cambio, las predicciones resultantes del modelo construido para predecir la baja porcentual asociada al precio de adjudicación son malas predicciones de la realidad. En particular, si bien el modelo exige mucho más tiempo de entrenamiento y mucha más capacidad computacional que el modelo de regresión lineal, se acaban obteniendo resultados muy similares. Ambos modelos (Random Forest y regresión lineal con regularización de Ridge) predicen la baja presentada por la empresa ganadora del concurso con una raíz del error cuadrático medio de 23 puntos porcentuales y un error absoluto medio de en torno a 15 puntos porcentuales, resultados que, como se observa en la figura 13, son bastante pobres.

Modelo	*Tiempo de adjudicación (meses)*		*Baja porcentual (tanto por 1)*	
Random Forest Regressor	RMSE = 1,23	MAE= 0,50	RMSE = 0,23	MAE = 0,16
Regresión lineal	RMSE = 2,52	MAE = 0,7	RMSE = 0,23	MAE = 0,15

En los siguientes gráficos se puede ver una muestra aleatoria de 100 observaciones de las muestras de validación ordenadas de menor a mayor por su valor real.

En el modelo de tiempos de adjudicación (a), una gran parte de las predicciones mostradas en rojo acierta el tiempo de adjudicación real observado, a diferencia de lo que sucede con el modelo lineal (b), en donde se observan discrepancias generalizadas entre lo que predice el modelo y lo que sucede en la realidad. Además el modelo de aprendizaje automático parece detectar saltos que podrían estarse produciendo debido a la existencia de distintos tipos de procedimiento, diferentes tipos de tramitaciones y otros aspectos contemplados en las variables incluidas. En cambio, el modelo de regresión lineal no parece ser capaz de detectar este tipo de patrón. Cuando se analiza qué variables están influyendo más en esta predicción aparece, en primer lugar, el tipo de contrato y, en segundo lugar, el código CPV, es decir el sector.

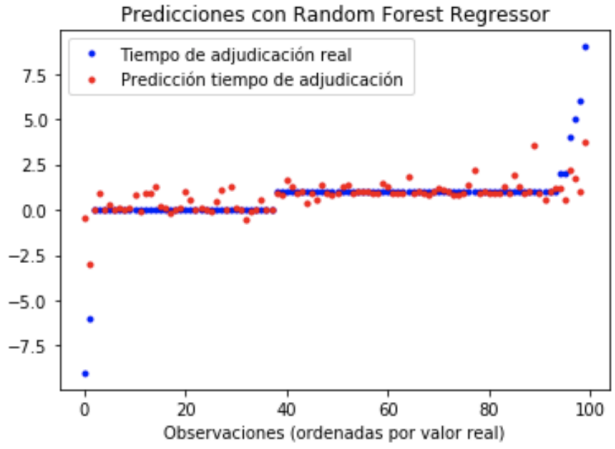
En la figura 13 se observa que el modelo de Random Forest de predicción de bajas porcentuales del precio de adjudicación es poco eficiente. Si bien parece comportarse algo mejor que un modelo de predicción lineal en los valores más reducidos, en términos generales el modelo no supone una mejora particularmente relevante respecto de un modelo lineal, tal y como confirman los valores de error cuadrático medio y error absoluto medio mostrados anteriormente. En su conjunto ninguno de los dos modelos utilizados para predecir la baja

[11] Se utiliza una regularización de Ridge para contar con un modelo de referencia básico que incluya, aunque sea con una influencia pequeña, todas las variables contempladas en el modelo de Random Forest

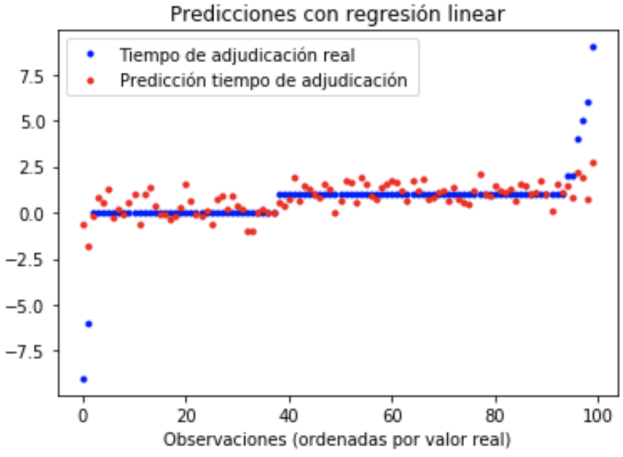
Figura 12.

Tiempos de adjudicación. Predicciones y observaciones de la muestra de validación
(Para 100 observaciones aleatorias)

(a)



(b)



Fuente: Elaboración propia.

porcentual parece útil, por lo que sería necesario buscar modelos alternativos que permitan mejorar estos resultados.

En el caso de la baja porcentual, las variables más relevantes son en primer lugar el sector (CPV) y en segundo lugar, el tipo de contrato. Una de las limitaciones importantes de este análisis es la imposibilidad de incluir el número de empresas concurrentes, variable que

probablemente sea muy significativa a la hora de determinar la baja ganadora. No obstante, no consideramos la posibilidad de incluir esta variable ya que lo que se pretende es predecir el gasto derivado de cada licitación en el momento en que esta se publica es decir, cuando no se cuenta con información sobre el número de empresas que concurrirán. Una alternativa que se podría explorar es construir primero un modelo que permita estimar el número de empresas concurrentes y, posteriormente, utilizar esa predicción como *input* de un segundo modelo que prediga la baja porcentual ganadora. Adicionalmente, se deberían explorar otro tipo de modelos más sofisticados e incorporar información no estructurada, como el texto descriptivo del objeto del contrato que se encuentra también disponible en esta base de datos.

Conviene apuntar que, al margen de los resultados obtenidos, los modelos de aprendizaje automático tienen ciertas limitaciones que deben tenerse en cuenta a la hora de hacer predicciones futuras sobre gasto público. En primer lugar, dado que estos modelos son muy intensivos en el uso de la evidencia empírica, es posible que se queden desactualizados a menudo. Así, si las relaciones entre variables cambian con el tiempo por circunstancias como cambios en la gestión pública de los contratos o situaciones excepcionales como la de la COVID-19, la representatividad de los datos utilizados para construir el modelo acaba siendo limitada. Por lo tanto, los modelos deben reentrenarse con cierta frecuencia. Por otro lado, estos modelos también sufren del problema estadístico tradicional de datos omitidos. Si las observaciones no consideradas por no tener toda la información no son aleatorias, podría haber un problema de generalización del modelo con datos futuros.

Figura 13.

Baja porcentual. Predicciones y observaciones de la muestra de validación
(Para 100 observaciones aleatorias)

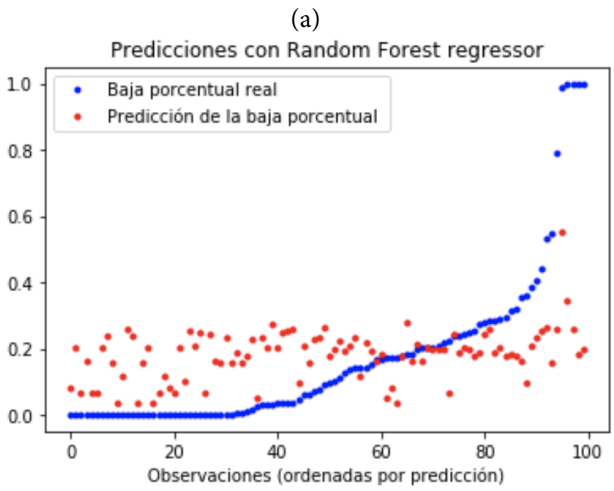
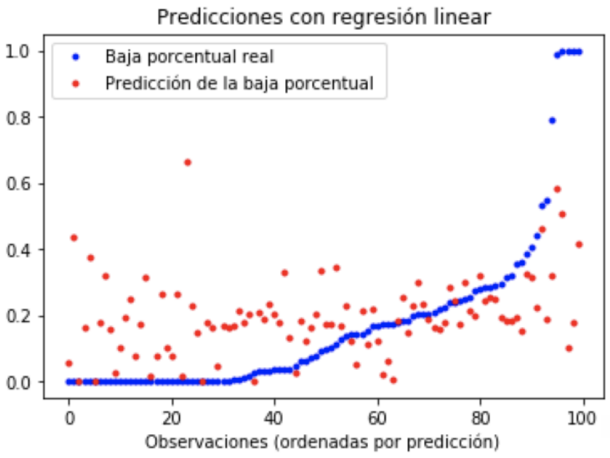


Figura 13. *(continuación)*

Baja porcentual. Predicciones y observaciones de la muestra de validación
(Para 100 observaciones aleatorias)

(a)



Fuente: Elaboración propia.

3. CONCLUSIÓN Y APLICACIONES ADICIONALES

La Base de Datos de Contratación del Sector Público tiene un amplio potencial para el análisis del gasto público tanto desde un punto de vista macroeconómico como desde un punto de vista microeconómico y de gestión pública. El análisis desarrollado ha mostrado el potencial de utilizar toda esa información para predecir el gasto público en tiempo real. No obstante, el potencial de uso de estos datos para otras finalidades es enorme. En particular consideramos que la información disponible permitiría por ejemplo detectar cuellos de botella en la absorción de fondos europeos, ayudar a mejorar la concurrencia de empresas en licitaciones públicas o detectar irregularidades en la contratación basadas en conexiones anómalas entre empresas y órganos de contratación o el fraccionamiento de contratos.

En lo que respecta a la posibilidad de contar con una previsión del gasto público actualizada con alta frecuencia, las avenidas abiertas en términos de investigación se centran en dos cuestiones fundamentales:

En primer lugar, mejorar los modelos que nos permitan predecir la baja porcentual de las licitaciones públicas. Para ello se recurrirá a modelos más sofisticados como modelos encadenados que permitan incorporar una previsión del número de empresas licitantes o redes neuronales. Sobre todo, se procurará explotar información adicional como la contenida en el objeto del contrato a través de técnicas de procesamiento del lenguaje natural.

En segundo lugar, se hará extensivo el análisis a otros conceptos de consumo e inversión públicas y a las subvenciones, utilizando toda la información pública disponible, incluida por ejemplo la Base de Datos Nacional de Subvenciones.

Una vez se desarrollen los necesarios modelos predictivos de los distintos factores que determinan el gasto público, se podrá contar con un indicador adelantado que agregue toda esta información y permita prever la evolución del gasto público con mayor precisión y anticipación.

Referencias

Breiman, L. (1996). Bagging predictors. *Machine Learning,* Volume 24, pp. 123–140.

CNMC (2011). *Guía sobre Contratación Pública y Competencia*. https://www.cnmc.es/file/123708/download

Hastie, T., Tibshirani, R. y Friedman, J. (2001). *The Elements of Statistical Learning*. New York Inc: Springer.

James, G., Witten, D., Hastie, T. y Tibshirani, R . (2014). *An Introduction to Statistical Learning: With Applications* in *R*. Springer Publishing Company.

Losada, R. (2017). ¿A qué nos referimos al hablar de consumo público? *AIReF, documentos de trabajo,* 2/2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. y Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *JMLR,* 12, pp. 2825-2830.

Royo, M. T. (2018). Los procedimientos de adjudicación de contratos públicos. *Monografías de la Revista Aragonesa de Administración Pública,* pp. 365-373

CAPÍTULO III

Economía laboral y *big data:* panorámica sobre técnicas de regularización en la evaluación de efectos causales

Juan J. Dolado*

En este trabajo se ofrece una panorámica sobre las técnicas de regularización existentes en la literatura de *machine learning* para modelos lineales y no lineales, con controles exógenos y tratamientos endógenos, destinados a evaluar los efecto de deteterminadas políticas sobre variables del mercado de trabajo. Una aplicación empírica de dichas técnicas al conocido estudio de Angrist y Krueger (1991) acerca de los efectos de la educación sobre los salarios sirve para ilustrar su uso creciente en economía laboral.

Palabras clave: big data, machine learning, postselección simple, postselección doble, lasso, efecto causal.

* Estoy agradecido a los editores, Juan Carlos Escanciano y participantes en un seminario interno de UC3M por sus comentarios. Este artículo fue presentado en las jornadas de Funcas sobre *Nuevos Enfoques de Problemas Económicos con "Big Data"*. Se agradece el apoyo financiero del Ministerio de Economía y Competitividad (proyecto ECO2017-86009-P).

1. INTRODUCCIÓN

La disponibilidad de fuentes estadísticas con un gran número de variables relativas al tamaño de la muestra (en el argot: datos masivos o *big data*) es cada vez más común en economía aplicada. Por ello, el uso de técnicas de aprendizaje automático (*machine learning* o *ML*) avanza a pasos agigantados en la investigación económica[1]. La economía laboral no es una excepción a esta regla ya que muchas de las bases de datos administrativas que se utilizan en esta disciplina poseen dimensiones inherentemente elevadas, con multitud de características para cada observación disponible. Por ejemplo, en el caso de España, tanto el *Censo de Población* como la *Encuesta de Población Activa,* la *Muestra Continua de Vidas Laborales,* la *Encuesta de Salarios,* la *Encuesta Financiera de las Familias* o los datos de la Agencia Tributaria, entre otras, proporcionan información sobre cientos de variables en relación con empresas o trabajadores. Su relevancia es fundamental para contrastar las hipótesis derivadas de los principales modelos acerca de, por ejemplo, el funcionamiento de los mercados de trabajo o sobre los efectos de la acumulación de capital humano sobre los salarios. Además, incluso cuando el número de variables relevantes fuera reducido, los investigadores rara vez conocen la forma funcional exacta con que aparecen en el modelo, lo que supone enfrentarse a un gran conjunto de interacciones y transformaciones potenciales de las variables subyacentes.

Sin embargo, como señalan Angrist y Frandsen (2020), *a priori* no está claro que el uso de técnicas de ML se adapte fácilmente a las necesidades de la economía laboral. Tradicionalmente la mayoría de las cuestiones relevantes en esta área de investigación se refieren a las características de las distribuciones de las variables aleatorias de interés (como puede ser la forma funcional de la media condicional), más que a la precisión de las predicciones fuera de la muestra. En efecto, gran parte de la agenda de investigación en economía laboral está dirigida tanto a la estimación de efectos causales –por ejemplo, el efecto de la educación sobre los salarios o qué tipo de tratamiento a los parados resulta más efectivo para mejorar su empleabilidad– como a proporcionar evidencia descriptiva sobre el efecto que tienen los cambios tecnológicos o políticas impositivas sobre la desigualdad de rentas y riqueza. Para ello, las herramientas estadísticas que se han venido utilizando tradicionalmente han sido los métodos de regresión habituales, incluyendo el uso extensivo de variables instrumentales (IV). Por tanto, ya sean causales o descriptivas, rara vez las cuestiones relevantes en economía laboral se han centrado en problemas de predicción pura. Como apuntan acertadamente Mullainathan y Spiess (2017) en su panorámica sobre el uso de ML en economía, la distinción entre estimación de parámetros y predicción individual es paralela a la diferencia entre las pendientes estimadas en un modelo de regresión ($\hat{\beta}$) y el R^2. El objetivo de las técnicas de ML es mejorar la precisión de los valores ajustados ($\hat{y}$), en lugar de optimizar las propiedades del estimador de un determinado coeficiente, aparentemente lo contrario de lo que interesa a los economistas laborales, quienes raramente prestan atención a $\hat{y}$ como objeto central de su investigación.

[1] Panorámicas recientes sobre el uso de ML en economía puede encontrarse en Belloni, Chernozhukov y Hansen (2014a), Mullainathan y Spiess (2017), y Athey e Imbers (2019).

No obstante, como señalan Belloni, Chernozhukov y Hansen (2013) y Chernovzhukov, Hansen y Spindler (2015), la conexión entre ambos tipos de objetivos aparece mucho mas evidente en presencia de big data. En efecto, en cualquier aplicación empírica con multitud de controles resulta necesario evitar un ajuste excesivo intramuestral (*data mining*) que impida extraer conclusiones a muestras diferentes de las que se usan para estimar los parámetros de interés. Igualmente, en presencia de un gran número de intrumentos, la precision de los estimadores de los efectos causales de una variable sobre otra a través de mínimos cuadrados bietápicos (MCB) mejora sustancialmente cuando la estimación de la primera etapa se enfoca como un problema de predicción en el que, de nuevo, se evite un sobreajuste de las variables intrumentadas.

A la vista de estas consideraciones, este trabajo se centra en estos dos dominios (uso de MCO e IV para especificar relaciones con big data) en los que el ML podría desempeñar un papel muy relevante en la búsqueda de efectos en la economía laboral. Para ello, se ofrece una panorámica de procedimientos recientes de regularización (esto es, selección de controles e instrumentos en la especificación de modelos lineales y no lineales con variables exógenas y endógenas).

El resto del artículo está organizado de la siguiente forma. En la sección segunda se revisa la relación existente entre efectos causales y modelos de regresión. Las secciones tercera y cuarta están dedicadas a repasar el uso de procedimientos de regularización en modelos lineales con variables exógenas y endógenas, respectivamente. La quinta sección resume las propiedades estadísticas de los principales métodos de regularización en ML. La sección sexta extiende los resultados anteriores a modelos no lineales. En la sección séptima se ofrece una aplicación empírica de estos procedimientos para la estimación de los rendimientos salariales de la educación. Finalmente, la octava sección concluye.

2. REGRESIÓN Y EFECTOS CAUSALES EN ECONOMÍA LABORAL

Como es bien sabido, la regresión utiliza modelos (generalmente) lineales para describir funciones de expectativas condicionales. La esperanza condicional de una variable aleatoria y_i para un individuo i ($i = 1,2,...,n$), como función de los datos de un conjunto de variables, x_i, se puede escribir $\mathbb{E}[y_i \mid x_i = x]$ o, en forma abreviada, $\mathbb{E}[y_i \mid x]$. El símbolo "$\mathbb{E}$" denota un promedio de población, mientras que $\mathbb{E}[y_i \mid x]$ representa el promedio de y_i para todos aquellos individuos que poseen características x_i iguales a un valor particular, x. Así, el interés de los economistas laborales se ha centrado tradicionalmente en estimar en cuánto aumentan los salarios en promedio con cada etapa educativa completada. Por ejemplo, se compara $\mathbb{E}[y_i \mid x_i = 16]$, el salario promedio de los graduados universitarios, con $\mathbb{E}[y_i \mid x_i = 12]$, el ingreso de los bachilleres. Debido a que $\mathbb{E}[y_i \mid x_i = x]$ toma tantos valores como x, a menudo los economistas aplicados aspiran a simplificar el modelo de esperanza condicional para resumir sus características más importantes, donde la regresión de y_i sobre x_i proporciona la mejor aproximación lineal. Volviendo al ejemplo anterior, la pregunta clave se centra en establecer (si existe) relación causal entre completar un grado universitario (en vez de simplemente el bachillerato) y los ingresos de un determinado individuo. El hecho de completar la educación

superior se denomina variable de tratamiento, denotada de aquí en adelante como d_i. Idealmente uno estaría interesado en computar la diferencia entre los resultados potenciales, $y_{1i} - y_{0i}$, donde y_{1i} e y_{0i} son los ingresos del individuo *i* si se graduara ($d_i = 1$) y si no lo hiciera ($d_i = 0$) pero la dificultad evidente es que, en función del valor que tome d_i solamente se observa una de las dos situaciones: bien y_{1i} o y_{0i}. Por tanto, el analista aspira a medir un efecto causal medio como $\mathbb{E}\,[y_{1i} - y_{0i}]$, denominado efecto promedio del tratamiento *(average treatment effect, ATE)* o bien el efecto promedio condicionado al tratamiento, $\mathbb{E}\,[y_{1i} - y_{0i} \mid d_i = 1]$ *(average treatment on the treated, ATT)* o a su ausencia, $\mathbb{E}\,[y_{1i} - y_{0i} \mid d_i = 0]$ *(average treatment on the non- treated, ATNT)*.

El vínculo entre inferencia causal y regresión se ve facilitado en un contexto de efectos causales homogéneos en el que se subraya el problema del sesgo de selección muestral pasando por alto la distinción entre diferentes tipos de promedios causales. El modelo subyacente se puede escribir en la forma siguiente:

$$\begin{aligned} y_{0i} &= \mu + \upsilon_i, \\ y_{1i} &= \alpha + y_{0i}, \end{aligned} \qquad [1]$$

donde, en la primera ecuación [1], μ es la media de y_{0i} mientras que υ_i representa su desviación individual respecto a dicha media. La segunda ecuación expresa que el efecto causal del tratamiento, $y_{1i} - y_{0i}$, es homogéneo e igual a α. Utilizando la relación existente entre los resultados observados y los contrafactuales a través de la identidad $y_{1i} \equiv y_{0i} + (y_{1i} - y_{0i})\, d_i$, dicho modelo puede reescribirse en términos de una única regresión como:

$$y_i = \mu + \alpha d_i + \upsilon_i \qquad [2]$$

La ecuación [2] plantea el problema del sesgo de selección en términos de υ_i, que se asemeja a un término de error de regresión. Sin embargo, a diferencia de una regresión, donde por definición los residuos no están correlacionados con los regresores, υ_i puede estar correlacionado con d_i excepto que el tratamiento se aplique de manera completamente aleatoria (ver abajo). Con datos observacionales, las soluciones al problema del sesgo de selección se basan en el supuesto clave de *independencia condicional en media* (ICM). En concreto, se supone la existencia de un amplio conjunto de características observables del individuo tales que:

$$\mathbb{E}\,(\upsilon_i \mid d_i = 1, \mathrm{x}_i = \mathrm{x}) = \mathbb{E}\,(\upsilon_i \mid d_i = 0, \mathrm{x}_i = \mathrm{x}), \qquad [3]$$

donde x_i es un vector de *p* controles que toman un valor particular igual a x. En otras palabras, en la población con $\mathrm{x}_i = \mathrm{x}$, la comparación de los ingresos de individuos con diferentes niveles de educación es un contraste de "manzanas con manzanas" en vez de "manzanas con peras". El supuesto clave es que $\mathbb{E}\,(\upsilon_i \mid d_i, \mathrm{x}_i) = \mathbb{E}\,(\upsilon_i \mid \mathrm{x}_i)$, de manera que, si la media condicional de y_{0i} es una función lineal de x_i, los controles han de ser "variables predeterminadas al tratamiento", es decir, no pueden ser resultados en sí mismos. Ello implica que $\mathbb{E}\,(\upsilon_i \mid \mathrm{x}_i = \mathrm{x}) = \beta' \mathrm{x}$, o $\upsilon_i = \beta' \mathrm{x} + \varepsilon_i$, con $\mathbb{E}\,(\varepsilon_i \mid \mathrm{x}) = 0$. Combinando estos supuestos se obtiene el tradicional modelo de regresión con interpretación causal:

$$y_i = \mu + \alpha d_i + \beta' \mathrm{x}_i + \varepsilon_i, \qquad [4]$$

que suele emplearse para obtener estimaciones insesgadas del efecto causal de interés, α. Nótese que, aunque generalmente el vector de coeficientes de los controles, β, no suele ser objeto de interés por parte del analista, resulta crucial incluir las variables x_i en la regresión como diagnóstico sobre la plausibilidad de [4]. Claramente, cuando p sea muy grande, será imprescindible aplicar métodos de regularización que permitan acotar el conjunto de controles relevantes para los que se cumple el supuesto de ICM.

No obstante, en la práctica, normalmente los modelos teóricos no especifican todas las variables que deben controlarse al estimar una relación, además de que puede resultar complicado medirlas con precisión incluso cuando la especificación es correcta. Por ejemplo, este es el caso de la habilidad intelectual innata de un individuo en una ecuación *minceriana* de salarios[2]. Una solución al problema de las variables omitidas es asignar de forma aleatoria a los participantes en los grupos de tratamiento y de control, de manera que la participación en el programa no esté correlacionada con los factores personales o sociales omitidos.

Sin embargo, los experimentos aleatorios no siempre son factibles, incluso condicionando en observables, como ocurre en [4]. No resultaría ético obligar a un grupo de personas a asistir a la escuela un año más al tiempo que se excluye a otro, de la misma forma que no parece razonable asignar el valor del salario mínimo al azar entre diferentes regiones de un país. Sin embargo, sí que resulta posible identificar un grado de variación exógena en variables como la escolaridad. Las variables instrumentales ofrecen una posible solución en el contexto de experimentos naturales.

En el caso de la educación, la teoría del capital humano sugiere que las personas eligen su nivel de educación comparando los costes y beneficios de las diferentes alternativas a que se enfrentan. Por tanto, una posible fuente de instrumentos podría estar en las diferencias en las políticas de préstamos, becas u otros subsidios que varían independientemente de la habilidad o el potencial de ingresos, o en la existencia de las limitaciones institucionales en la edad de acceso a la educación obligatoria. En este último caso, un famoso trabajo de Angrist y Krueger (1991), que analizaremos en detalle posteriormente, elige un amplio conjunto de variable instrumentales, denominadas z_i, basadas en la regla de que, en varios distritos escolares de EE. UU., los niños ingresan en la educación obligatoria en otoño del año en que cumplen 6 años, mientras que a todos se les permite abandonar la escuela en el momento de cumplir 16 años. Por ello, los alumnos nacidos a principios de año acceden a la escuela a una edad más avanzada que aquellos nacidos a finales de año, de manera que alcanzan la edad legal de abandono escolar tras haber obtenido menos años de educación. En esencia, la combinación de las políticas sobre la edad de inicio de la escuela y las leyes de escolarización obligatoria crean un experimento natural en el que los niños se ven obligados a asistir a la escuela durante

[2] Una ecuación minceriana capta el rendimiento de la educación en términos salariales como el coeficiente de los años de educación en un modelo de regresión donde la variable dependiente es el logaritmo de los salarios por hora y a la que se añaden otros regresores, como la antigüedad laboral, para capturar la formación en el puesto de trabajo; véase Heckman, Lochner y Todd (2006).

periodos de tiempo diferentes, de acuerdo con su fecha de nacimiento, lo que repercute en su educación y por tanto en sus ingresos futuros.

En el caso de IV, cuando el tratamiento no sea aleatorio y pueda depender del término de error, el modelo causal relevante sería el siguiente:

$$\begin{aligned} y_i &= \alpha d_i + \beta' \mathrm{x}_i + \varepsilon_i, \\ d_i &= \gamma' z_i + e_i, \\ \mathrm{z}_i &= \varphi' \mathrm{x}_i + \omega_i, \end{aligned} \qquad [5]$$

donde el supuesto de ICM implicaría $\mathbb{E}(\varepsilon_i \mid \mathrm{z}_i, x_i) = 0$, siendo z_i un conjunto de instrumentos válidos proporcionados por algún experimento natural, como el discutido en la sección séptima[3]. De nuevo, cuando las dimensiones de los vectores de parámetros β, γ y φ sean muy elevadas, la aplicación de métodos de regularización de ML resultará imprescindible para logran un buen estimador del parámetro de interés α.

3. SELECCIÓN DE CONTROLES EN MODELOS LINEALES CON VARIABLES EXÓGENAS (MCO)

Empezaremos considerando el caso del modelo [4] donde existen p controles incorrelacionados con el término de error ε_i pero correlacionados con la variable de interés, d_i. Lo relevante en presencia de big data es que la dimensión p puede ser muy elevada (incluso muy superior al tamaño muestral), por las razones que se apuntaban anteriormente: por una parte, la creciente disponibilidad de infinidad de características individuales cuyo efecto sobre la distribución de d_i no puede descartarse *a priori* y, por otra, el hecho de que la forma funcional de las esperanzas condicionales sea desconocida y muy flexible. Como punto de partida, supondremos que la media condicional es lineal, de manera que [4] corresponde a:

$$y_i = \alpha d_i + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i \mid \mathrm{x}_i, z_i) = 0, \qquad [6]$$

donde, para simplificar la notación, el término constante de la regresión se ha incluido en el conjunto de p controles.

El siguiente paso se centra en escoger un método de regularización para seleccionar los controles más relevantes en [6]. La práctica más habitual consiste en seleccionar dicho subconjunto de regresores en [6] usando lo que se ha denominado el método de *Post Selección (PS)*. Funciona de la siguiente manera. Primero, se incluye un determinado regresor x_{ij} si resulta ser un predictor significativo de y_i en la ecuación [6], habiendo excluido d_i. Para ello se puede utilizar un contraste conservador dentro del ambito clásico (tests t o F) cuando

[3] Nótese que en la segunda ecuación de [5] se pueden incluir el control x_i, de forma que $d_i = \gamma' \mathrm{z}_i + \phi' \mathrm{x}_i + e_i$. Se supone que $\phi = 0$ para simplificar los cálculos sin que cambien cualitativamente los resultados derivados en secciones posteriores.

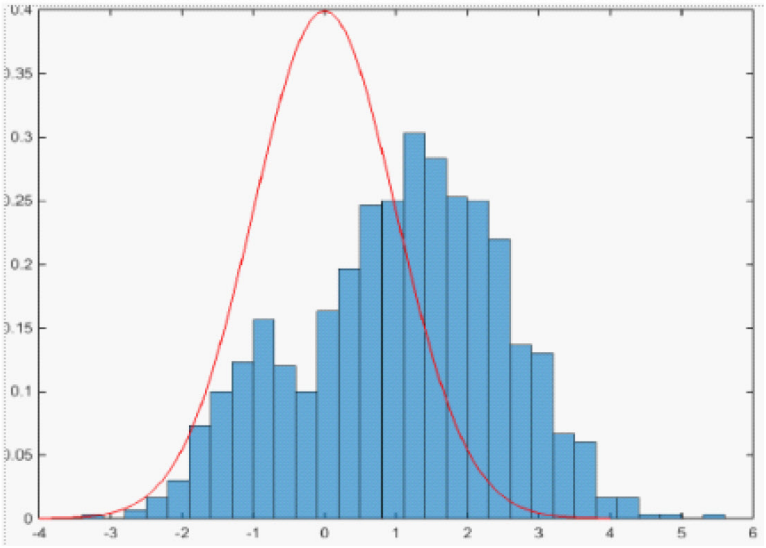
$p < n$ o alguna técnica más moderna dentro del conjunto de herramientas habituales de ML (Lasso y sus variantes, Random Trees, Random Forests, Boosting, Bagging, Neural Networks, etc.)[4]; en caso contrario, se excluyen dichos regresores. Una vez seleccionados los controles x_{ij} que son relevantes para predecir y_i, se vuelve a ajustar el modelo, esta vez incluyendo d_i en la regresión. De esta forma se estima el parámetro de interés α utilizando intervalos de confianza estándar. Sin embargo, como veremos a continuación, este procedimiento puede fallar si $|\beta|$ es cercano (pero no igual) a cero, esto es, formalmente cuando $|\beta| \approx 1/\sqrt{n}$. Para entender los problemas derivados de PS, es conveniente analizar un proceso de generación de datos (PGD en adelante) muy sencillo propuesto por Belloni *et al.* (2014a y b), donde hay un único control ($p = 1$) que está correlacionado con el tratamiento d_i pero no con la perturbación ε_i, es decir:

$$\begin{aligned} y_i &= \alpha d_i + \beta x_i + \varepsilon_i, \\ d_i &= \gamma x_i + \upsilon_i, \end{aligned} \qquad [7]$$

donde se simulan los datos para $n = 100$ considerando los siguientes valores de los parámetros: $\alpha = 0$, $\beta = 0{,}2$, $\gamma = 0{,}8$, $\varepsilon_i \sim N(0,1)$, $x_i \sim N(0,1)$, $\upsilon_i \sim N(0{,}0{,}32)$ y $\mathbb{E}(\varepsilon_i \upsilon_i) = 0$. Como se muestra en la figura 1, tras implementar el método PS en 1.000 simulaciones de Monte Carlo, se rechaza la hipotesis nula $H_0 : \alpha = 0$ en alrededor del 50 % de los casos con un nivel de significación nominal del 5 % para el constraste t, y lo mismo ocurre con Lasso (se puede demostrar que cuando $p \ll n$ ambos procedimientos propocionan resultados muy similares).

Figura 1.

(Post Selección (PS): t-ratio al 5 %)



Fuente: Elaboraciòn propia.

[4] Para un excelente compendio de técnicas de ML, véase Hastie, Tibshirani y Friedman (2009).

Otra posibilidad sería usar *Bootstrap*, pero tampoco funciona bien puesto que simplemente replica la distribución $N(0,1)$ del término de error ε_i.

Por contra, la aplicación de un procedimiento alternativo, denominado *Post Selección Doble (PSD),* sí que logra que el nivel de significación efectivo de los contrastes coincida con el nivel nominal (véase Belloni *et al.* (2014a y b). Dicho procedimiento consiste en los siguientes pasos:

(i) Se incluye x_{ij} como control en la regresión si resulta ser un predictor significativo tanto de y_i como de d_i con cualquiera de los procedimientos clásicos o de ML señalados anteriormente.

(ii) Una vez seleccionado o descartado el control, se ajusta el modelo utilizando intervalos de confianza estándar, o alternativamente, se regresa el residuo obtenido para y_i sobre el residuo de d_i, ambos obtenidos a partir de (i).

Nótese que el procedimiento PSD es equivalente al uso del teorema de parcialización de Frisch-Waugh-Lovell para el cómputo de los coeficientes de MCO en el modelo de regresión lineal estándar. En resumen, se incluye x_{ij} en la regresión siempre y cuando ayude a predecir tanto la variable dependiente como el tratamiento, a diferencia del método PS donde la inclusión de x_{ij} en la primera ecuación de [7] solo depende de si predice bien y_i.

El origen del problema de utilizar PS puede entenderse de manera intuitiva computando la forma reducida de y_i en el modelo ilustrativo [7], esto es:.

$$y_i = (\alpha\gamma + \beta)\, x_i + (\varepsilon_i + \alpha v_i) = \pi\, x_i + \eta_i, \qquad [8]$$

donde $\pi = \alpha\gamma + \beta$, y $\sigma_\eta^2 = \sigma_\varepsilon^2 + \alpha^2\sigma_v^2$. A partir de [8] se observa como el método PS solo tenderá a escoger x_i como regresor relevante en [7] cuando su coeficiente π alcance un valor suficientemente elevado, mientras que se descartará x_i cuando el tamaño de su coeficiente sea reducido. Sin embargo, en este último caso, el hecho de descartar un control que presenta un fuerte poder predictivo para d_i (p. ej. en la simulación anterior $\gamma = 0{,}8$), puede acarrear un importante *sesgo de variables omitidas* (SVO) en el estimador de α cuando el coeficiente β de esta variable en la ecuación a estimar sea pequeño (en el PGD previo, $\beta = 0{,}2$). Intuitivamente, el efecto de cualquier control con un impacto directo moderado sobre la variable y_i se atribuirá incorrectamente al efecto del tratamiento d_i, con el que está fuertemente correlacionado. En consecuencia, la variable x_i quedará excluida de la regresión. Igualmente, si se aplicara un método de selección de variables para predecir d_i en la segunda ecuación de [7], se excluiría x_i siempre que γ fuera reducido, lo cual sería incorrecto en caso de que el tamaño de β sea elevado. De nuevo, tal tipo de omisión de variables puede producir un SVO no despreciable.

En otras palabras, para aminorar el SVO en la estimacion de α, resulta crucial incluir en la primera regresión de [7] todos aquellos controles que sean que resulten útiles para predecir tanto y_i como d_i, procediendo a continuación a regresar (en este caso por MCO) y_i sobre d_i y la unión de todos los controles preseleccionados en la primera etapa.

Las consecuencias de omitir d_i en la regresión [7] cuando su impacto directo sobre y_i es reducido (*i.e.* β es reducido) pueden apreciarse formalmente derivando el SVO a partir de la expresión [8], esto es[5]:

$$\sqrt{n}\left(\hat{\alpha}-\alpha\right)=\left(\sum_{i=1}^{n}\frac{d_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{d_i\varepsilon_i}{\sqrt{n}}+\sqrt{n}\left(\sum_{i=1}^{n}\frac{d_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{x_i^2}{n}\beta\gamma:=\left(A\right)+\left(B\right).$$

Bajo el supuesto de exogeneidad de x_i, el término (A) en la expresión anterior es asintóticamente $N(0,E\left(d_i^2\right)^{-1})$. Nótese que el procedimiento PS descartará correctamente el control x_i siempre que su coeficiente β sea suficientemente pequeño, lo que formalmente ocurre cuando $\beta = O\left(\frac{1}{\sqrt{n}}\right)$. No obstante, incluso en dicho caso, el término (B) puede no anularse ya que asintóticamente se comporta como:

$$\sqrt{n}\beta\gamma\approx\sqrt{n}\,O\left(\frac{1}{\sqrt{n}}\right)\gamma\nrightarrow 0 \text{ si } \gamma\neq 0,$$

Por contra, el método de PSD solamente descartará x_i si no aparece como predictor descritivo relevante tanto para y_i como para d_i. Al igual que con β, ello ocurrirá si el tamaño del coeficiente γ es reducido, esto es, cuando $\gamma = O\left(\frac{1}{\sqrt{n}}\right)$. En dicho caso, el término (B) se convierte en :

$$\sqrt{n}\beta\gamma\approx\sqrt{n}O\left(\frac{1}{\sqrt{n}}\right)O\left(\frac{1}{\sqrt{n}}\right)\rightarrow 0,$$

por lo que esta vez el SVO desaparece.

4. SELECCIÓN DE CONTROLES EN MODELOS LINEALES CON TRATAMIENTO NO ALEATORIO (IV)

Adicionalmente, cabe señalar que los argumentos en el ejemplo anterior se generalizan fácilmente al caso de IV, cuando el tratamiento d_i no sea asignado aleatoriamente (condicional en x_i) sino que se administre en función de una variable instrumental z_i correlacionada con x_i pero ortogonal a la perturbación ε_i. En este caso, el PGD corresponde al siguiente sistema de tres ecuaciones:

$$\begin{aligned} y_i &= \alpha d_i + \beta x_i + \varepsilon_i, \\ d_i &= \gamma z_i + e_i, \\ z_i &= \varphi x_i + \omega_i, \end{aligned} \qquad [9]$$

donde ahora $\mathbb{E}(\varepsilon_i\, e_i) \neq 0$, $\mathbb{E}(\varepsilon_i\, \omega_i) = 0$ y $\mathbb{E}(\varepsilon_i \mid x_i) = \mathbb{E}(\omega_i \mid x_i) = 0$. Al igual que en [7], el sistema se puede reeescribir en forma reducida, de manera que el vector de variables (y, d, z) dependa

[5] Recuérdese que en una regresión lineal con variables exógenas estimada por MCO $y = \beta_1 x + \beta_2 z + u$, el sesgo de β_1 al excluir z es $\mathbb{E}\left(\hat{\beta}_1\right) - \beta_1 = \beta_2 \mathbb{E}(x,z) / \mathbb{E}\left(x^2\right)$.

del control exógeno x_i. En este caso, la aplicación del procedimiento PSD consistirá en escoger x_i siempre que sea un buen predictor de cada una de las tres variables en el sistema, aplicando posteriormente MCB a la primera ecuación de [9] con el fin de estimar α.

Pese a haber analizado inicialmente el caso sencillo con $p = 1$ a efectos ilustrativos, en la práctica el caso más realista en presencia de big data es aquel donde p es muy grande, dependiendo posiblemente del tamaño muestral n, de manera que $p = p_n$, donde $p \approx n$ o p » n. En estas circunstancias, la aplicación de MCO o MCB no es factible y, por ello, se requiere el uso de métodos de regularización basados en ML. Para cubrir los casos analizados previamente en términos de modelos predictivos, denotemos como $\mathbf{w}_i$ al vector de los datos $(y_i, d_i, z_i)'$, de manera que dichos modelos pueden representarse de la siguiente forma:

$$\mathbf{w}_i = \sum_{j=1}^{p} \phi_j x_{ij} + \xi_i;\ \mathbb{E}\left(\xi_i \mid \mathrm{x}_{ij}\right) = 0,\ j = 1,2,...,p. \qquad [10]$$

Para proceder a la regularización de los coeficientes en [10] se necesitan dos supuestos clave en ML: (I) *Parsimonia (approximate sparsity)* en el conjunto de parámetros ϕ en [10], lo que conlleva la existencia de un subconjunto de dichos coeficientes, de dimensión $s_n \ll p_n$, que son relevantes mientras que los restantes no lo son tanto; por ejemplo, tras ordenar los coeficientes por tamaño, esta condición se verifica si $|\phi_j| \leq Aj^{-a}$ para $j = 1,2,...,p_n$, siendo A una constante y $a > 1$, y (II) *Isometria Resgringida,* una propiedad de álgebra lineal aplicada a la matriz de covarianzas de los controles que implica la existencia de pequeños grupos de regresores que son cuasi-ortonormales, es decir, con dependencia muy reducida.

5. MÉTODOS DE REGULARIZACIÓN EN ML

Los supuestos descritos anteriormente subyacen a la mayoría de los procedimientos de selección de variables mediante ML, entre los que se encuentran los métodos de regularización más populares en econometría. La idea básica de estos procedimientos es que un predictor mejor fuera de la muestra puede conllevar un aumento de las sumas de los cuadrados de los residuos en [10]. En consecuencia, se añade un término de regularización que se encargue de la elimininación de los coeficientes más pequeños y, con ello, del diseño de modelos más parsimoniosos, esto es, con una menor dimensión que la contemplada en un ajuste por MCO del modelo sin restricciones. Idealmente, los estimadores minizan la siguiente función de pérdidas (donde, a efectos ilustrativos, nos centramos en la ecuación que determina y_i en [10]):

$$\min_b \left[\sum_{i=1}^{n} \frac{\left(y_1 - \Sigma_{j=1}^{p} b_j x_{ij}\right)^2}{n} + \frac{\lambda}{n} \sum_{j=1}^{p} 1\{b_j \neq 0\} \right], \qquad [11]$$

donde 1 {·}es una función indicador y el parámetro λ penaliza la dimensión del modelo. Dicho objetivo incluye a los criterios de información de Akaike y Schwartz. Desafortunadamente, este problema resulta prohibitivo en términos computacionales (problema NP) cuando p es muy grande, ya que requiere efectuar $\Sigma_{j \leq n} \binom{p}{j}$ regresiones. La solución propuesta por el

método Lasso es convexificar la función de pérdidas anterior sustituyendo el tamaño de los coeficientes por su valor absoluto, lo que da lugar a la siguiente función alternativa de pérdidas (Tibshirani, 1996):

$$\min_b \left[\sum_{i=1}^{n} \frac{\left(y_1 - \Sigma_{j=1}^{p} b_j x_{ij}\right)^2}{n} + \frac{\lambda}{n} \sum_{j=1}^{p} |b_j| \right], \qquad [12]$$

o en el caso de Lasso adaptativo

$$\min_b \left[\sum_{i=1}^{n} \frac{\left(y_1 - \Sigma_{j=1}^{p} b_j x_{ij}\right)^2}{n} + \frac{\lambda}{n} \sum_{j=1}^{p} \varpi_j |b_j| \right],$$

donde los pesos ϖ_j representan las penalizaciones heterogéneas de los coeficientes b_j que, junto a λ, han de elegirse antes de aplicar Lasso, bien mediante procedimientos de validación cruzada (CV) o a través de un estimador complentario (*plug-in*). Bajo los dos supuestos enunciados previamente, Bickel, Ritov y Tsybakov (2009) demuestran que una elección adecuada del parámetro *plug-in* λ es $\lambda = 2\sigma_\xi^2 \sqrt{2n1n(pn)}$ donde la varianza del ruido σ_ξ^2 puede calcularse de forma iterativa; p. ej., inicializando el proceso a través del cómputo de la varianza de los datos originales, σ_y^2 y procediendo de forma recursiva. Otra posibilidad es aplicar el procedimiento denominado *Root Lasso* propuesto por Belloni, Chernozhukov y Wang (2011), donde la función anterior de pérdidas pasa a ser:

$$\min_b \left[\sqrt{\sum_{i=1}^{n} \frac{\left(y_1 - \Sigma_{j=1}^{p} b_j x_{ij}\right)^2}{n}} + \frac{\lambda}{n} \sum_{j=1}^{p} |b_j| \right], \qquad [13]$$

de manera que el criterio a minimizar se convierte en pivotal respecto a σ_ξ^2, con $\lambda = \sqrt{2n1n(pn)}$. Cuando cualquiera de estos criterios se aplican para regularizar la regresión [10] bajo los dos supuestos anteriores (parsimonia en los parámetros e isometra en los regresores) se obtienen los siguientes resultados (véase Belloni *et al.,* 2014a y b) y Chernovzhukov, Hansen y Spindler (2015) para los detalles):

- Lasso y Root Lasso identifican modelos de tamaño óptimo s_n (véase la definición de *Parsimonia* arriba) que, en el caso de $|\phi_j| \leq A_j^{-a}$, resulta ser $s_n = n^{\frac{1}{2a}}$,
- El uso de ambos procedimientos como primera etapa de PSD en la regresión [10] de nuevo identifica modelos de tamaño óptimo que, en el caso de $|\phi_j| \leq A_j^{-a}$, resulta ser $s_n = \sqrt{\frac{s}{n}} \log(pn)$.
- En este último caso, se verifica que $\hat{\sigma}_n^{-1} \sqrt{\hat{\alpha} - \alpha} \rightsquigarrow N(0,1)$, donde σ_n es la fórmula convencional del estimador MCO de α en [6] y "$\rightsquigarrow$" denota convergencia débil en distribución.

Para ilustrar la utilidad de estos resultados, resulta conveniente considerar un PGD con controles exógenos similar al simulado previamente para $p = 1$, pero esta vez con $p \gg n$ y

donde se cumplen los dos supuestos clave señalados antes. En concreto, siguiendo a Belloni *et al.* (2014), consideremos la siguiente generalización del PGD anterior, ahora con un gran número de controles:

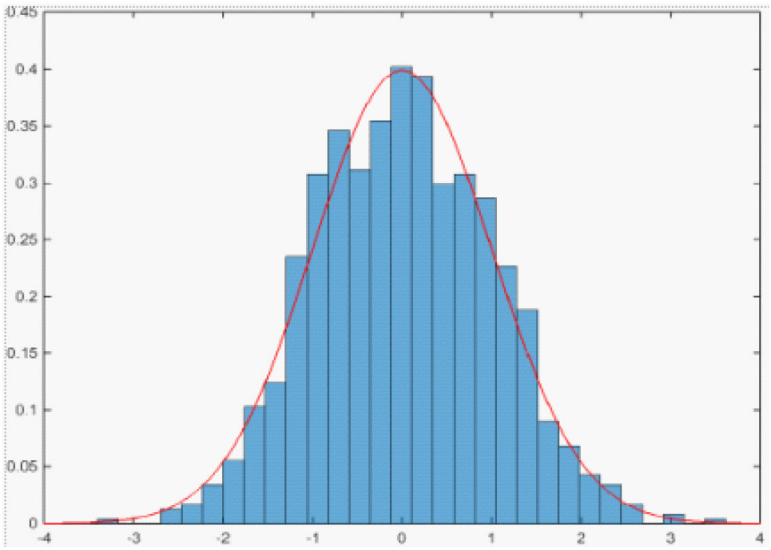
$$y_i = \alpha d_i + \sum_{j=1}^{p} \beta_i x_{ij} + \varepsilon_i,$$

$$d_i = \sum_{j=1}^{p} \gamma_i x_{ij} + \nu_i, \qquad [14]$$

donde $p = 200$, $n = 100$, $\alpha = 0$, $|\beta_j|$ y $|\gamma_j| = O\left(\frac{1}{j^2}\right)$, $\varepsilon_i \sim N(0,1)$, $\nu_i \sim N(0,0,32)$, $\mathbb{E}(\varepsilon_i \nu_i) = 0$, y x $\sim N(0,\Omega)$ con $\Omega_{kj} = (0,5)^{|j-k|}$. Utilizando 1.000 simulaciones, el resultado de contrastar $H_0 : \alpha = 0$, tras regularizar los parámetros β_j y γ_j mediante Lasso y aplicar PSD con un nivel nominal del 5 %, es una tasa de rechazo del 5,3 %, muy cercana al nivel nominal del contraste. Por contra, cuando se usa PS para regularizar únicamente la forma reducida de y_i, el rechazo efectivo es del 47 % (figura 2), corroborando de esta forma los resultados obtenidos para $p = 1$.

Figura 2.

(Post Selección Doble (PSD): Lasso al 5%)



Fuente: Elaboraciòn propia.

6. REGULARIZACIÓN EN MODELOS NO LINEALES

Si bien los resultados anteriores se cumplen para modelos lineales, su generalización a modelos no lineales no resulta excesivamente complicada. La lógica subyacente es similar a la

anterior. No obstante, ahora aparecen algunos nuevos resultados que permiten relajar alguno de los supuestos previos, como son las restricciones inherentes al supuesto de parsimonia en el número de parámetros verdaderos (la dependencia entre p y n). Como ejemplo ilustrativo en esta sección, usaremos un modelo lineal parcial (Robinson, 1988). En este modelo existe un parámetro de interés, α_0, asociado a un tratamiento o al cambio en una determinada política, d, esta vez en presencia de un gran número de controles, x, cuya forma funcional sea desconocida (en vez de ser lineal) y potencialmente complicada, tanto en la primera como en la segunda ecuación. En concreto, el DGP ilustrativo en este caso es el siguiente

$$\begin{aligned} y &= \alpha_0 d + g_0(\mathrm{x}) + \varepsilon, \\ d &= m_0(\mathrm{x}) + v, \end{aligned} \qquad [15]$$

donde x es un vector (n x p) de controles (con $p \simeq n$ o $p \gg n$) cuya inclusión es necesaria para que se cumpla $\mathbb{E}(\varepsilon|d,\mathrm{x}) = \mathbb{E}(v|\mathrm{x}) = 0$. El subindice "0" en los parámetros anteriores indica su verdadero valor y, en consonancia con los estudios observacionales, se supone que $m_0(\mathrm{x}) \neq 0$. De forma similar, se denota como $l_0(x)$ a la verdadera función de x que predice y en la forma reducida.

Un procedimiento habitual para estimar α_0 consiste en utilizar un método de regularización de ML (Lasso o cualquier otro entre los mencionados) de forma iterativa. Por ejemplo, usando un estimador incial de $\alpha_0^{(0)}$, se puede computar el residuo $\left(y - \hat{\alpha}_0^{(0)} d\right)$ y utilizar técnicas de ML en la regresión de dicho residuo sobre x para estimar $g_0^{(1)}(\mathrm{x})$ de forma no paramétrica A continuación se computa un nuevo residuo $\left(y - g_0^{(1)}(\mathrm{x})\right)$ que se regresa sobre d, obteniendo otro estimador $\hat{\alpha}_0^{(1)}$, y así sucesivamente hasta que el procedimiento iterativo converja. Este método resulta similar a la aplicación de PS en el modelo lineal que, como vimos, funciona mal. De nuevo, la intuición es que ML produce excelentes predicciones pero aumenta el SVO que es lo que importa a la hora de estimar la primera ecuación en [15].

La alternativa consiste en aplicar PSD de la siguiente manera. Primero, se utiliza ML para predecir d e y dado x, y de esta manera estimar las esperanzas condicionales $E(y|\mathrm{x}) = l_o(\mathrm{x})$ y $E(d|\mathrm{x}) = m_o(\mathrm{x})$. A continuación, se obtienen los residuos $\hat{\varepsilon} = y - \widehat{E(y/x)}$ y $\hat{v} = d - \widehat{E(d/x)}$. Finalmente, se regresa $\hat{\varepsilon}$ sobre $\hat{v}$ por MCO para obtener $\hat{\theta}$. Al igual que en el modelo lineal, el procedimiento PSD funciona correctamente, siendo parecida la intuición de por qué lo hace. No obstante aparecen nuevos resultados procedentes de la no linealidad que tienen interés. Para entender los argumentos en este caso, conviene analizar las condiciones de momentos que subyacen a los dos procedimientos anteriores (PS y PSD), las cuales vienen dadas por:

$$\mathbb{E}\left[\left(y - \alpha_0 d - g_0(\mathrm{x})\right)d\right] = 0 \text{(PS)},$$
$$\mathbb{E}\left[\left(y - E(y|\mathrm{x}) - \left(d - \left(E(d|z)\alpha_0\right)\right)\left(d - E(d|\mathrm{x})\right)\right] = 0 \text{(PSD)}.$$

En el caso de aplicar PS, la primera condición de momentos implica que:

$$\hat{\alpha}_{PS} = \left(\sum_{i=1}^{n} \frac{d_i^2}{n}\right)^{-1} \sum_{i=1}^{n} \frac{d_i\left(y_i - \hat{g}_0(x)\right)}{\sqrt{n}}$$

es decir:

$$\sqrt{n}(\hat{\alpha}_{PS}-\alpha_0)=\left(\sum_{i=1}^{n}\frac{d_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{d_i\varepsilon_i}{\sqrt{n}}+\left(\sum_{i=1}^{n}\frac{d_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{d_i\left(g_0(x)-\widehat{g_0}(x)\right)}{\sqrt{n}}:=(A)+(B). \quad [16]$$

Una vez más, mientras que el término (A) converge asintóticamente a una distribución $N(0, E(d^2)^{-1})$, el término (B) difiere de 0. De hecho, (B) diverge ya que la tasa de convergencia de los estimadores no paramétricos resulta ser más lenta que la de los paramétricos: $n^{-\varphi}$ con $0{,}25 < \varphi < 0{,}5$, en vez de $n^{-0{,}5}$. Por tanto, en el límite, dicho término equivale a:

$$(B)\approx\sqrt{n}n^{-\varphi}\to\infty,$$

con lo que, al igual que en la regresión lineal, la aplicación de PS a este modelo pueda acarrear un SVO muy elevado. Por contra, la condición de momentos para PSD implica que:

$$\hat{\alpha}_{PDS}=\left(\sum_{i=1}^{n}\frac{\hat{v}_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{\hat{\varepsilon}_i\hat{v}_i}{\sqrt{n}},$$

tal que $\hat{\varepsilon}=y-\hat{l}_0(\mathrm{x})$ y $\hat{v}=d-\hat{m}_0(\mathrm{x})$ Por consiguiente, dado que $\hat{\varepsilon}=\varepsilon-(\varepsilon-\hat{\varepsilon})$ y $\hat{v}=v-(v-\hat{v})$ se obtiene:

$$\begin{aligned}&\sqrt{n}(\hat{\alpha}_{PDS}-\alpha_0)=\\&\left(\sum_{i=1}^{n}\frac{v_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{\varepsilon_i v_i}{\sqrt{n}}+\left(\sum_{i=1}^{n}\frac{v_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{\left(l_0(\mathbf{x})-\hat{l}_0(x)\right)\left(m_0(\mathbf{x})-\hat{m}_0(\mathbf{x})\right)}{\sqrt{n}}+r(\mathbf{x}) \quad [17]\\&:=(A)+(B)+(C)\end{aligned}$$

donde de nuevo el término (A) es asintóticamente normal. En cuanto a (B), a diferencia de lo que ocurría al emplear PS, ahora sí que converge a 0 pues cada uno de los errores de predicción de d e y son de orden $n^{-\varphi_k}$ (k=l,m); por tanto:

$$(B)\approx\sqrt{n}n^{-(\varphi_m+\varphi_l)}\to 0.$$

Finalmente, para conseguir un buen comportamiento del estimador PSD queda por demostrar que el término (C) es asintóticamente despreciable, esto es, $o(1)$. Nótese que (C) captura un término residual, $r(\mathrm{x})$, que depende de los productos cruzados de ε y v con $(\varepsilon-\hat{\varepsilon})$ y $(v-\hat{v})$. Belloni *et al.* (2014b) demuestan que este resultado se puede alcanzar mediante la partición de la muestra total de n observaciones en al menos dos submuestras independientes donde, en una de ellas se usa ML para estimar $l_0(\mathrm{x})$ y $m_0(\mathrm{x})$ mientras que en la otra se regresa el residuo de predicción $\hat{\varepsilon}$ sobre $\hat{v}$, ambos obtenidos a partir de la primera submuestra. De hecho, dicho procedimiento permite relajar parcialmente el supuesto de *Parsimonia* necesario en la aplicación de PSD a modelos lineares con un gran número de controles.

Todos los resultados anteriores se pueden enmarcar de forma general en términos de la denominada *Condición de Ortogonalidad* de Neyman (1979) que pasamos a discutir brevemente. Por una parte, sea W la matriz de datos, θ_0 un conjunto de dimensión reducida con los

parámetros de interés, y $\eta_0 = (l_0, m_0)$ el vector de las verdaderas funciones predictivas de un conjunto p-dimensional de controles x sobre d e y (también sobre z si se requiriera un instrumento). Por otra parte, sea $\Psi\ (W, \theta_0, \eta_0)$ la función que captura las condiciones de momentos que permiten identificar θ_0 como la solución de la minimización de $\mathbb{E}\ (\Psi\ (W, \theta_0, \eta_0) = 0$.

La cuestión abordada por Neyman es cómo modificar el estimador de θ_0 de forma que su distribución asintótica no se vea afectada por pequeños cambios en η_0, dado que sus componentes son desconocidos y habrán de reemplazarse en la anterior condición de momentos por estimaciones no paramétricas de los mismos. Esta última propiedad se traduce intuitivamente en la siguiente condición de ortogonalidad:

$$\partial_\eta\, \mathbb{E}(\Psi(W, \theta_0, \eta)\, \downarrow_{\eta=\eta_0}=0, \qquad [18]$$

donde ∂ es la derivada parcial convencional cuando la función Ψ sea diferenciable (en términos más generales, se tomaría la derivada direccional o de Fréchet). En el caso del modelo lineal parcial, se puede comprobar fácilmente que la condición de momentos verifica la siguiente igualdad:

$$\mathbb{E}(\Psi(W, \theta_0, \eta_0) = \mathbb{E}[(y\text{-}l(x)\text{-}\theta_0(d\text{-}m(\mathrm{x}))(d\text{-}m(\mathrm{x})) = 0,$$

en la que, diferenciando respecto a los dos componentes de η y sustituyendo en $\eta = \eta_0$, se cumple [18]. Nótese que dicha condición de ortogonalidad es similar a la usada en el método $C(\alpha)$ de Neyman (1979) que permite realizar inferencia sobre la estimación de θ_0 con independencia de η_0, siempre y cuando Ψ se interprete como una funcion de verosimilitud.

Finalmente, con algunos supuestos adicionales, los argumentos esgrimidos anteriormente pueden generalizarse a modelos completamente no lineales en las variables pero aditivos en las perturbaciones, del tipo:

$$y = h_0\ (d,\mathrm{x}) + \varepsilon,$$

$$d = m_0\ (\mathrm{x}) + v, \qquad [19]$$

con $\mathbb{E}(\varepsilon \mid d, \mathrm{x}) = \mathbb{E}(v \mid \mathrm{x}) = 0$, donde $h_0(d, \mathrm{x})$ es una función desconocida de la variable de interés d y de los controles x, que puede estimarse con métodos semi o no paramétricos (véase Chernovzhukv *et al.,* 2020). Este modelo engloba a los discutidos previamente (inchuyendo a aquellos estimados por IV en los que basta añadir otra ecuación relacionando d con el instrumento z, como en [9]). Un caso particular de este tipo de modelos puede utilizarse para estimar el efecto medio de un tratamiento binario, $d_i = \{0, 1\}$, sobre la variable y. Puesto que d no es aditivamente separable de x, este modelo permite la existencia de efectos heterogéneos en el tratamiento. En este caso, la interpretación natural del parámetro θ_0 se corresponde con el efecto promedio del tratamiento (ATE), definido por:

$$\theta_0 = \mathbb{E}\ [h_0(1,\mathrm{x}) - h_o(0,\mathrm{x})],$$

el cual puede expresarse alternativamente (utilizando la distribución conjunta de d y x) como:

$$\theta_0 = \frac{d\,y}{\Pr(d=1|\,\mathrm{x})} - \frac{(1-d)y}{1-\Pr(d=1|\,\mathrm{x})},$$

donde $Pr\,(d{=}1 \mid \mathrm{x})$ es la probabilidad de recibir el tratamiento por parte de un individuo con características x, lo que en la literatura se conoce como *propensity score*. Combinando ambas definiciones de θ_0, se puede utilizar la siguiente condición de momentos:

$$\Psi(W,\theta,\eta) = h(1,x) - h(0,\mathrm{x}) + \frac{d\,(y - h(1,\mathrm{x}))}{m(\mathrm{x})} - \frac{(1-d)(y - h(0,\mathrm{x}))}{1-m(\mathrm{x})} - \theta,$$

que verifica la condición de ortogonalidad de Neyman con $\eta = (h(1,x),\ h(0,\mathrm{x}),\ m(\mathrm{x}))$. Chernozhukov *et al.* (2017) proponen un método de PSD consistente en utilizar una partición de la muestra de tamaño n en K submuestras, denominadas I_k (k = 1,2,...,K), con n/K observaciones cada una. Para cada I_k, se utilizan las observaciones de su conjunto complementario de submuestras restantes, denotado I_k^c para estimar los componentes del vector η por ML, mientras que I_k se usa para obtener $\hat{\theta}_0$ como la solución de la condición de momentos $n^{-1}\ \Sigma_{i\in I_k^c}\Psi\ (W,\theta,\ \eta(I_k^c))=0$. Finalmente, se construye el estimador promedio $\hat{\theta}_{0K} = K^{-1}\ \Sigma_{k=1}^{K}\theta_0\left(I_k, I_k^c\right)$. Bajo condiciones menos restrictivas de parsimonia e isometría, dicho estimador satisface el mismo resultado que se obtuvo para los modelos total o parcialmente lineales, esto es:

$$\sigma^{-1}\sqrt{n}\left(\hat{\theta}_{0K} - \theta_0\right) \rightsquigarrow N(0,1),$$

donde $\sigma^2 = \mathbb{E}[\Psi^2\ (\mathrm{w},\ \theta_0,\ \eta_0\ (\mathrm{x})]$.

Otro método que resulta más sencillo de implementar en el caso de tratamientos heterogéneos en contextos no experimentales es el denonominado *método de controles modificados (MCM)*. A través de una estrategia de identificación basada en la selección de observables, MCM parte de un caso particular de la función $h_0\ (d,\mathrm{x})$ en [19], consistente en el siguiente modelo con interacciones:

$$y_i = \beta'\,\mathrm{x}_i + d_i\,\delta'\,\mathrm{x}_i + \varepsilon_i, \qquad [20]$$

donde, para simplificar, se supone que el tratamiento (condicional en los observables x_i) se administra aleatoriamente a un 50 % de la muestra de individuos, de forma que $\mathbb{E}(\varepsilon_i \mid d_i, \mathrm{x}_i) = 0$. Por tanto el primer término de la derecha en [20] representa una aproximación lineal de la esperanza condicional de la variable y para los no participantes en el programa, $\mathbb{E}[y_{0i} \mid \mathrm{x}] = \beta'\,\mathrm{x}$, mientras que el segundo término proporciona otra aproximación lineal del ATE, $\mathbb{E}[y_{1i} - y_{0i} \mid \mathrm{x}] = \delta'\,\mathrm{x}$. Cuando p sea muy grande, el problema de regularización de los coeficientes se agudiza al duplicarse el conjunto de regresores. Tian *et al.* (2014) proponen abordar la regularización en dos etapas. Para ello, usan la transformación de la variable indicador d_i en la forma $T_i = 2d_i - 1$, de manera que el modelo de interacciones [20] pasa a ser:

$$y_i = \beta'\,\mathrm{x}_i + \frac{T_i}{2}\delta'\,x_i + \varepsilon_i.$$

Dado que $d_i \in \{0,1\}$, se cumple $T_i/2 \in \{-0{,}5,0{,}5\}$, por lo que $\mathbb{E}(T_i) = 0$. Esta modificación, que no altera el vector de coeficientes de interés δ, consigue que los dos conjuntos de regresores en [20] sean ortogonales, ya que $Cov\,(x_{ij}, T_i x_{ik}) = Cov\,(x_{ij}, x_{ik})\,\mathbb{E}(T_i) = 0$ para $j,k \in \{1,2,...,p\}$; nótese que la primera igualdad procede de la asignación aleatoria del tratamiento y la segunda de $\mathbb{E}(T_i) = 0$. Ello permite abordar la estimación de los coeficientes δ independientemente de los coeficientes β en un modelo más parsimonioso que [20], dado por :

$$y_i = \frac{T_i}{2}\delta' x_i + \varepsilon_i,$$

que constituye la regresión básica del MCM a la que se puede aplicar Lasso o cualquier otro procedimiento de ML para seleccionar aquellos controles que presenten efectos heterogéneos. Pese a haber utizado el supuesto de asignación aleatoria en aras a la simplificación, Chen *et al.* (2017) han demonstrado que la idea básica del MCM puede extenderse a estrategias de identificación no experimentales (en contextos observacionales) mediante su combinación con procedimientos de ponderación de probabilidad inversa (IPW).

7. APLICACIÓN EMPÍRICA

A continuación se proporciona una ilustración empírica de los procedimientos de regularización discutidos previamente mediante la aplicación de PSD en el modelo Angrist y Krueger (1991, AK en adelante) comentado en la sección segunda. Recordemos que estos autores encuentran relaciones estadísticamente significativas entre el trimestre del año en que uno nace, el nivel educativo y los ingresos para las cohortes de las décadas de 1920, 1930 y 1940 en EE. UU. Recordemos que los nacidos durante el primer trimestre del año obtienen menos educación y tienen menores ingresos que los nacidos durante los restantes trimestres del año ya que las regulaciones sobre asistencia escolar obligatoria en EE. UU. típicamente exigen que los estudiantes comiencen el primer grado en el otoño del año en que cumplen 6 años y que permanezcan en la escuela hasta que cumplan los 16 años. Por consiguiente, las personas nacidas a principos de año generalmente ingresan en el primer grado cuando tienen cerca de 7 años de edad y cumplen los 16 a mediados del décimo grado. Por contra, las personas nacidas en el tercer o cuarto trimestre generalmente comenzarán la escuela justo antes o justo después de cumplir 6 y terminarán el décimo grado antes de cumplir los 16. El modelo a estimar es el siguiente:

$$\begin{aligned} w_i &= \alpha s_i + \beta' \mathrm{x}_i + \varepsilon_i, \qquad & \mathbb{E}(\varepsilon_i \mid \mathrm{x}, \mathrm{z}) = 0, \\ s_i &= \gamma' \mathrm{z}_i + \phi' \mathrm{x}_i + v_i, \qquad & \mathbb{E}(v_i \mid \mathrm{x}, \mathrm{z}) = 0, \end{aligned} \qquad [21]$$

donde w_i es el logaritmo del salario del individuo i, s_i denota los trimestres de educación obligatoria, x_i es un vector de p controles, y z_i es un vector de m variables instrumentales ($m > 1$) que afectan a la educación pero no directamente al salario. Los datos proceden del censo de EE. UU. de 1980 y contienen observaciones para casi 330.000 hombres nacidos entre

1930 y 1939[6]. En concreto, z es un conjunto de 510 variables formado por: una constante, indicadores (*dummies*) de 9 años de nacimiento, 50 dummies del Estado donde nacieron, y 450 interacciones de los dos conjuntos anteriores de dummies. En concreto, AK usan los siguientes instrumentos: tres dummies del trimestre de nacimiento, sus dobles interacciones con el estado de procedencia y año de nacimiento y la triple interacción de todas ellas a la vez, es decir un total de 1.530 instrumentos potenciales. Se remite al lector interesado a AK (1991) para obtener el resto de detalles de su estimación. El coeficiente de interés es α, el cual recoge el impacto causal de la educación sobre los ingresos.

En la literatura se encuentan dos opciones básicas para estimar [21]: (i) usar como instrumentos solamente las tres dummies del trimestre de nacimiento o, alternativamente, (ii) utilizar 180 instrumentos resultantes de las tres dummies de trimestre de nacimiento y sus interacciones con las 9 de años de nacimiento y las 50 de estado de procedencia (excluyendo triples interacciones). Hansen, Hausman y Newey (2008) argumentan que el uso del conjunto de 180 instrumentos en estimaciones por MCB presentan un sesgo sustancial pero mayor precisión que cuando solo se usan tres instrumentos por el problema de "instrumentos débiles". Una posible solución a este problema es usar el estimador LIML de Fuller (1977), denominado FULL, consistente en corregir el estimador por MCB por sesgos de orden n^{-1}. Por ejemplo, suponiendo (para simplificar) que $\beta = \phi = 0$ en [21] , FULL proporcional el estimador $\hat{\alpha}_{FULL}$ dado por:

$$\hat{\alpha}_{FULL} = \arg\min_{\alpha} \frac{(\mathrm{x}-\alpha s)' Q_z (\mathrm{x}-\alpha s)}{(\mathrm{x}-\alpha s)'(\mathrm{x}-\alpha s)} \Rightarrow \hat{\alpha}_{FULL} = \left(s' Q_z s - \tilde{k} s' s\right)^{-1} \left(s' Q_z \mathrm{w} - \tilde{k} s' \mathrm{w}\right),$$

$$\tilde{k} = [\hat{k} - \left(1-\hat{k}\right)\frac{c}{n}][1-\left(1-\hat{k}\right)\frac{c}{n}], Q_z = z\left(z'z\right)^{-1} z',$$

tal que $c \geq 4$ es un parámetro a elegir y $\hat{k} = \hat{\varepsilon}' Q_z \hat{\varepsilon} / \hat{\varepsilon}' \hat{\varepsilon}$, donde $\hat{\varepsilon}$ son los residuos en el modelo [21].

El cuadro 1 presenta estimaciones de los rendimientos a la educación mediante MCB y FULL para diferentes conjuntos de instrumentos. Las tres primeras filas corresponden a las

Cuadro 1.

Estimación del rendimiento de la educación en AK (1991)

nº IVs	*MCB*	*FULL*
3	0.106 (0.019)	0.109 (0.021)
180	0.096 (0.010)	0.103 (0.014)
1.530	0.069 (0.005)	0.108 (0.042)
1(*)	0.087 (0.031)	---
12(**)	0.088 (0.014)	0.089 (0.014)

Nota: Estimadores MCB y FULL del parámetro α en el modelo de regresión [21]. Desviación típica entre paréntesis. (*) Lasso *plug-in,* (**) Lasso CV.
Fuente: Elaboración propia mediante los comandos *poivgress* y *lasso linear* de Stata.

[6] Los datos están disponibles en la web: https://economics.mit.edu/faculty/angrist/data1/data/angkru1991

agrupaciones naturales de los instrumentos comentados anteriormente de tamaño 3.180 y 1,5, respectivamente. Las dos últimas filas ofrecen los resultados basados en el uso de LASSO para seleccionar instrumentos con niveles de penalización dados por Lasso *plug-in* y CV en 10 subconjuntos de la muestra. De acuerdo con el primer valor de λ, Lasso únicamente selecciona la dummy de haber nacido en el cuarto trimestre, mientras que con CV elige 12 instrumentos entre los que se encuentran las dummies de haber nacido en tercer y cuarto trimestres. Todos los estimadores de α se obtienen utilizando PDS como técnica de regularización.

El primer resultado a destacar es que, con 180 o 1.530 instrumentos, existen algunas diferencias entre las estimaciones por MCB y FULL. Sin embargo, desaparecen al usar PSD con un número pequeño de instrumentos (1 o 10), lo cual indica que este procedimiento evita el sobreajuste en la primera etapa de IV. Además, aunque Lasso desconoce la relevancia de las dummies de trimestre de nacimiento entre los 180 o 1.530 instrumentos, siempre incluye alguna de ellas en el conjunto seleccionado (especialmente la variable dummy del cuarto trimestre, es decir la de los individuos más favorecidos por el tratamiento). Con este instrumento se estima el rendimiento anual de la escolarización en 0,087 con una desviación típica estimada de 0,031, mientras que con 180 y 1.530 se encuentra alrededor de 0,11 (MCB) y 0,10 (FULL). En general, estos resultados demuestran que la selección de instrumentos por PSDS es factible, produciendo estimaciones sensatas y comparables con las disponibles en esta literatura.

8. CONCLUSIONES

En este trabajo se ofrece una panorámica de los métodos de regularización disponibles en la literatura de ML que pueden ser útiles para abordar preguntas relevantes en economía laboral. Se destaca que los procedimientos de post selección doble (PSD) ofrecen resultados muy superiores a los de post selección (PS) sencilla. Estas ventajas se producen tanto en la especificación del conjunto de controles para estimar el efecto de un determinado tratamiento basado en variables observables cuando hay datos masivos en contextos no experimentales, como cuando se usa un gran número de variables instrumentales para tratamientos endógenos. Además, estos resultados son válidos para para modelos lineales y no lineales.

Referencias

Angrist, J. y Frandsen, B. (2020). Machine Labour. *NBER WP,* 26584.

Angrist, J. y Krueger, A. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics,* 196, pp. 979-1014.

Athey, S. e Imbens, G. (2019). Machine Learning Methods Economists Should Know About. *Annual Reviews of Economics,* 11, pp. 685-725.

Belloni, A., Chen, D., Chernozhukov, V. y Hansen, CH. B. (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica,* 80, pp. 2369–2429.

Belloni, A., Chen, D., Chernozhukov, V. y Hansen, CH. B. (2013). Inference for High-Dimensional Sparse Econometric Models. *Advances in Economics and Econometrics. 10th World Congress,* Vol. 3, edited by D. Acemoglu, M. Arellano y E. Dekel, pp. 245–295. Cambridge University Press.

Belloni, A., Chen, D., Chernozhukov, V. y Hansen, CH. B. (2014a).Inference on Structural and Treatment Effects with High-Dimensional Data. *Journal of Economic Perspectives,* 2014.

Belloni, A., Chen, D., Chernozhukov, V. y Hansen, CH. B. (2014b). Inference on Treatment Effects with High-Dimensional Controls, with Application to Abortion and Crime. *Review of Economic Studies,* 81, pp. 608-650.

Belloni, A., Chernozhukov, V. y Wang, L. (2011). Square-root Lasso: Pivotal Recovery of Space Signal via Conic Programming. *Biometrika,* 98, pp. 791-806.

Bickel, P. J., Ritov, Y. y Tsybakov, A. B. (2009). Simultaneous Analysis of Lasso and Dantzig Selector. *Annals of Statistics,* 37, pp. 1705 –1732.

Chen, S., Tian, L., Cai, T. y Yu, M. (2017). A General Statistical Framework for Subgroup Identification and Comparative Treatment Scoring. *Biometrics,* 73, pp. 1199-1209.

Chernozhukov, V., Chetverikov, D., Demiert, M., Dufflo, E., Hansen, CH. B. y Newey, W. (2017).Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review P&P,* 107, pp. 261-265.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K. y Robins, J. (2020). *Locally Robust Semiparametric Estimation.* Mimeo.

Chernozhukov, V., Hansen, CH. B. y Spindler, M. (2015). Post-Selection and Post-Regularization Inference: An Elementary, General Approach. *Annual Review of Economics,* 7, pp. 649-688.

Fuller, W. A. (1977). Some Properties of a Modification of the Limited Information Estimator. *Econometrica,* 45, pp. 939-954.

Hansen, CH. B., Hausman, J. y Newey, W. K. (2008). Estimation with Many Instrumental Variables. *Journal of Business & Economic Statistics,* 26, pp. 398–422.

Hastie, T., Tibshirani, R. y Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York, NY: Springer.

Heckman, J., Lochner, L. y Todd, P. (2006). Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond. En: Eric Hanushek y Finis Welch (eds.), *Handbook of Education Economics,* Vol. 1, chapter 7. Elsevier.

Mullainathan, S. y Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives,* 31, pp. 87-106.

Neyman, J. (1979). C(α) Tests and their Use. *Shankhya: The Indian Journal of Statistics,* 41, pp. 1-21.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56, pp. 931–954.

Tian, L., Alizadeh, A. A., Gentles, A. J. y Tibshirani, R. (2014). A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. *Journal of the American Statistical Association,* 109, pp. 1517-1532.

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

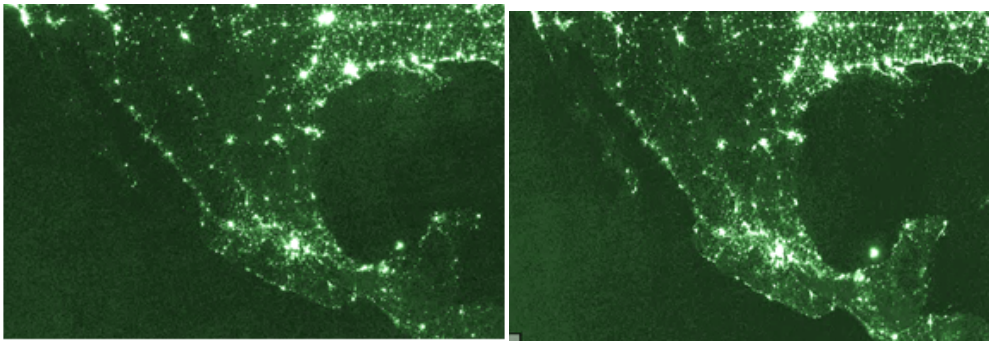
o entre distintas regiones dentro de un país. Tales comparaciones son una herramienta clave para el análisis del crecimiento económico del país. El uso de imágenes de satélites para medir el crecimiento del PIB, a través de la intensidad de las luces, es una herramienta que mejora dicha medición, sobre 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)



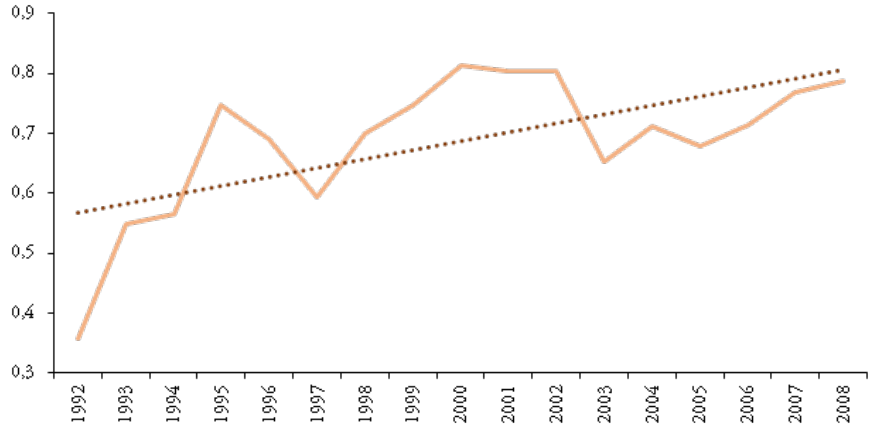
Fuente: Oficina Nacional de Administración Oceánica y Atmosférica de Estados Unidos. https://sos.noaa.gov/datasets/nighttime-lights-comparison-1992-2000-and-2008/

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,

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

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.

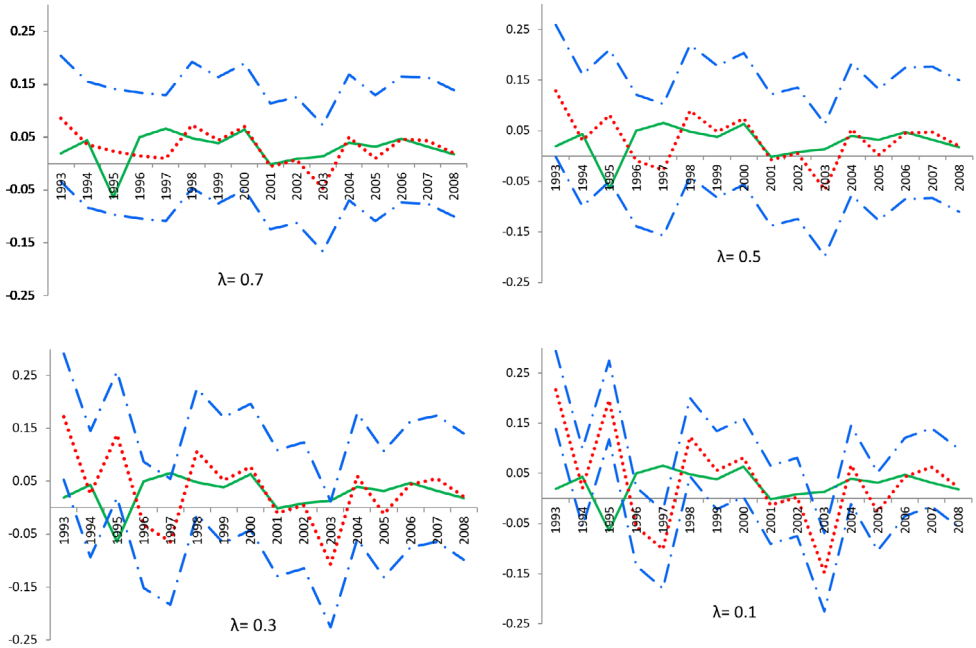
Resultados de la estimación para México con $\hat{\beta}$ = 0.808 y diferentes valores de λ

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)



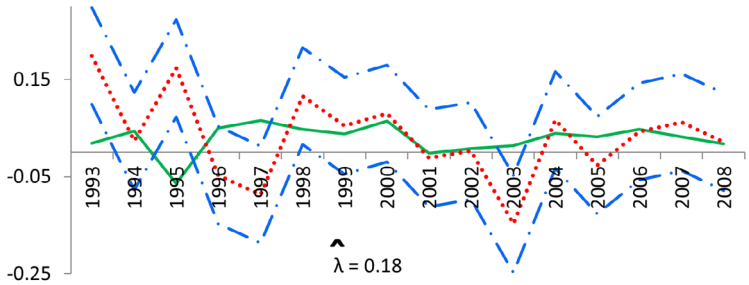
Fuente: Elaboración propia.

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)

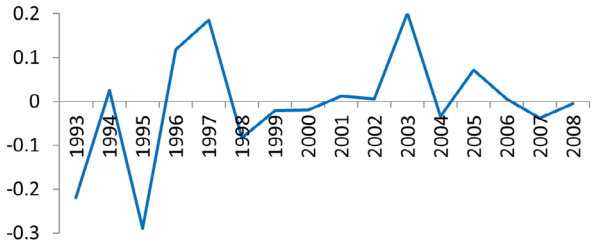


Fuente: Elaboración propia.

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

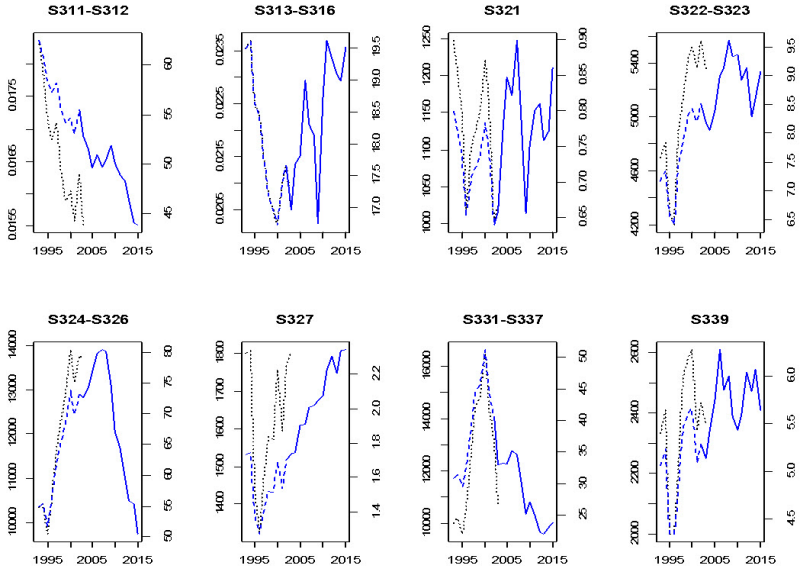
Los datos que surgieron como resultado de la aplicación de la fase de conversión están expresados en millardos de pesos a precios constantes de 2008 y estas son las unidades que se usan en las figuras y tablas de la parte numérica de este trabajo. Cabe recordar que la conversión surge de un procedimiento 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.

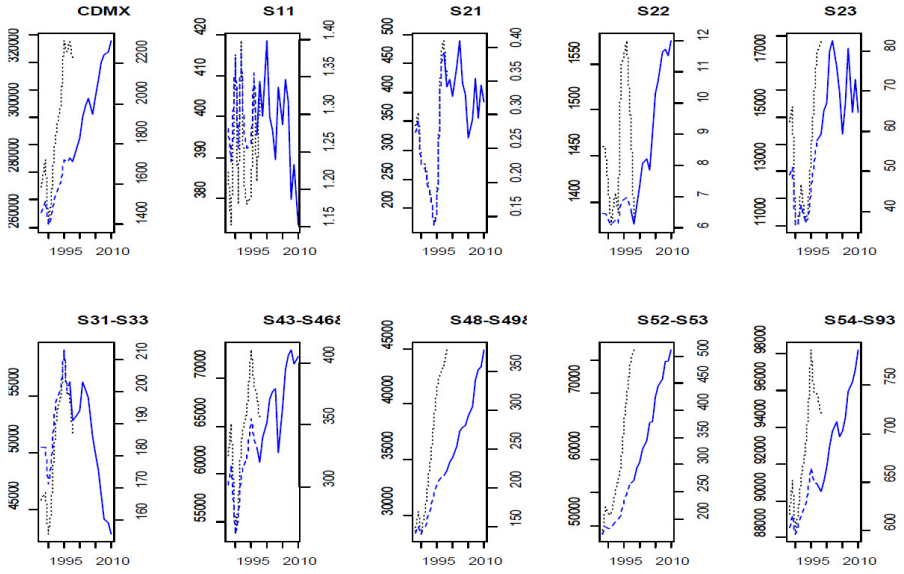
Conversión por subsectores de manufacturas. Serie observada (base 1993) en línea negra punteada, periodo de conversión en línea azul punteada y serie base 2008, en línea azul



Fuente: Elaboración propia.

Figura 7.

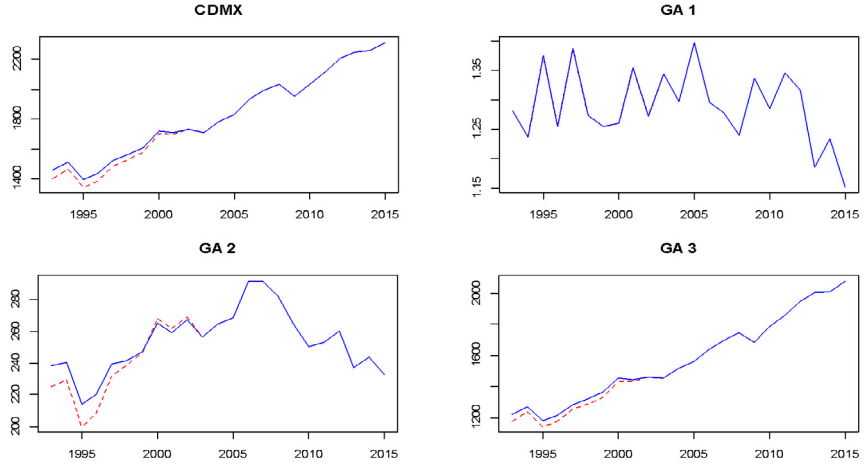
Conversión por sectores. Serie observada (base 1993) en línea negra punteada, periodo de conversión en línea azul punteada y serie observada base 2008, en línea azul continua



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.

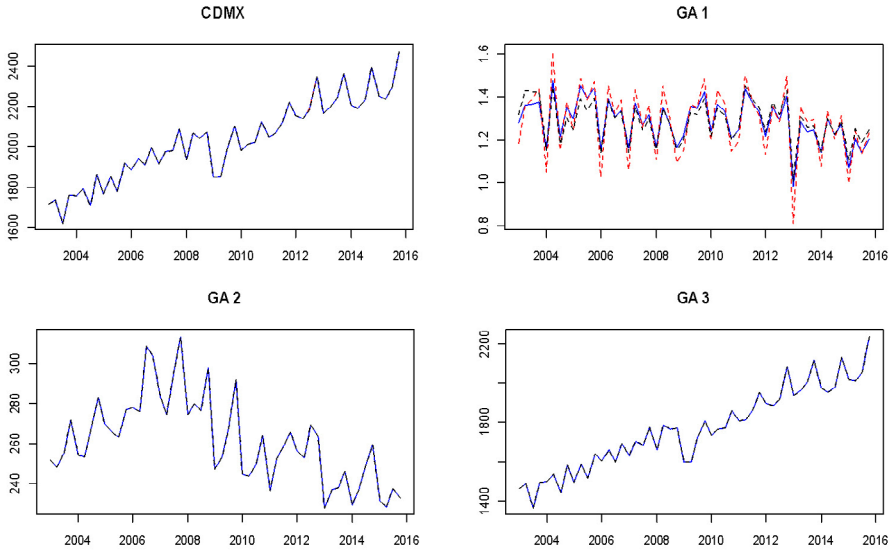
Desagregación temporal univariada de cada GA y del total

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

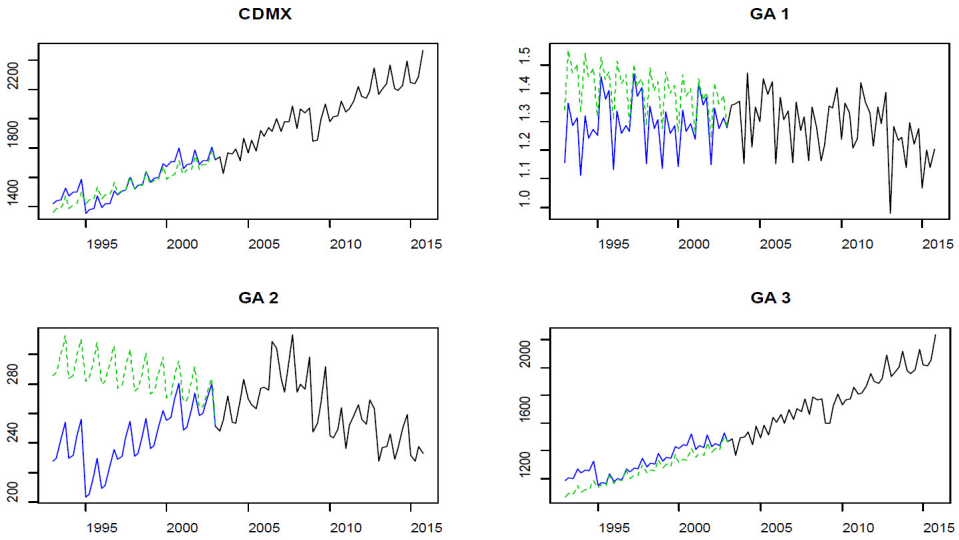
El modelo resumido en la tabla 4 cumple con los supuestos de estacionariedad y de no autocorrelación del error, por lo que se considera estadísticamente válido y se obtienen las matrices de varianzas para los errores de pronóstico un periodo hacia atrás, que se usan al generar las retropolaciones restringidas,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

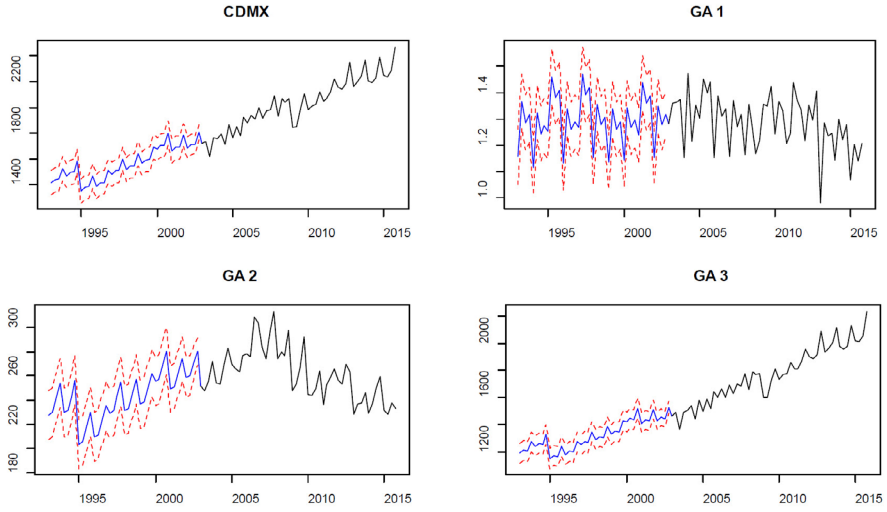
Reconciliación con las cifras trimestrales nacionales 1993-2002

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.

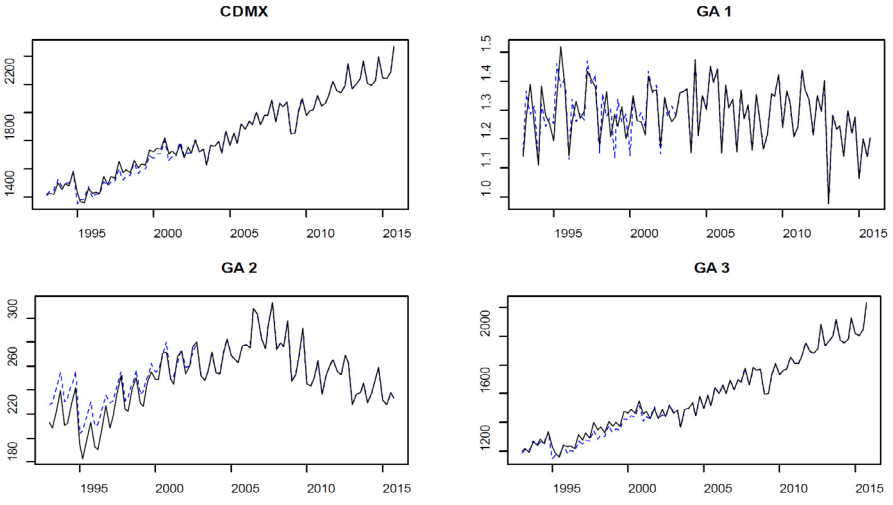
Retropolación restringida multivariada e intervalos de predicción: GA y total
(Millardos de pesos a precios constantes de 2008)



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).

CAPÍTULO V

Éxitos y retos de *big data* en análisis económico: un recorrido a través de ejemplos

Pilar Poncela*
Eva Senra

La ingente cantidad de información disponible presenta soluciones y retos a problemas existentes en análisis económico. Cada vez son más las aplicaciones de éxito, bien basadas en actualización de la metodología estadística disponible, bien en la utilización de nuevas bases de datos. El reto pendiente es pasar de aplicaciones puntuales de éxito del uso de big data a su utilización generalizada por parte de los responsables del análisis económico. Presentamos diversos ejemplos (integración financiera, *nowcasting* y generación de nuevos indicadores de innovación y movilidad), señalando algunas oportunidades que big data proporciona y apuntando algunos retos que quedan por resolver.

Palabras clave: análisis de conglomerados, calidad, indicadores, procesamiento natural del lenguaje.

* Se agradece el apoyo financiero del Ministerio de Ciencia, proyectos números MINECO/FEDER PID2019-107161GB-C32 y PID2019-108079GB-C22/AEI/10.13039/501100011033.

1. INTRODUCCIÓN

La revolución del *big data* nos proporciona una ingente cantidad de información generada por los propios individuos a partir de sus interacciones con las redes sociales (texto, fotos y vídeos), publicaciones en blogs, texto de búsquedas en internet o repositorios, datos empresariales como las transacciones comerciales (*e-comercio,* tarjetas bancarias, móviles), información generada por máquinas, a través del conocido como Internet de las Cosas (IOT) que nos permite disponer de datos de sensores (tráfico, tiempo, cámaras de seguridad), o dispositivos de rastreo (GPS, localización del móvil), entre otros. Esta revolución también ha dado lugar a la necesidad de adaptar las técnicas estadísticas para su aplicación a bases de datos masivos.

Las profesiones relacionadas con el análisis de datos se encuentran entre las más demandadas en estos momentos, entre otros motivos, debido al big data. Las grandes empresas ya incorporan estos perfiles en sus compañías, las convierten en *data-driven* y reconocen la necesidad de almacenar y preservar sus datos, organizarlos y definir el tipo de objetivos e indicadores que necesitan. Y las que no lo hacen tienen en mente la necesidad de gestionar sus datos y no perder oportunidades de negocio a corto, medio y largo plazo.

Si bien el problema de la toma de decisiones dentro del contexto de una empresa es complejo, la construcción y utilización de indicadores basados en big data que puedan ser utilizados en la estadística oficial y sean útiles en la toma de decisiones de política económica es todavía un reto pendiente.

Este capítulo tiene dos objetivos: primero, presentar varios ejemplos donde se han utilizado técnicas de big data en análisis económico y, segundo, reflexionar cuándo esto no constituye una moda pasajera, o una colección de casos de éxito, sino una revolución que ha venido para quedarse e incorporarse de manera sistemática en su aplicación en política económica.

El primer objetivo ilustra, por un lado, la necesidad de adaptación de las técnicas estadísticas en un contexto de mayor información mediante una aplicación a la integración financiera utilizando datos diarios de bolsa a través del análisis de conglomerados o *cluster*. Dada la gran abundancia de datos, es necesario restringir la comparación entre los distintos índices bursátiles solamente a las dinámicas relevantes, ilustrando cómo se puede modificar el análisis clásico de conglomerados de series temporales a grandes conjuntos de datos.

Por otro lado, se ilustra la construcción de indicadores a partir de bases de datos alternativas a través de tres ejemplos, que responden a necesidades diferentes. Así, el segundo ejemplo analiza si la incorporación de la información contenida en las noticias es útil para generar predicciones de corto plazo, o incluso del trimestre actual o pasado, antes de la publicación del dato oficial de los principales indicadores macroeconómicos (*nowcasting*). En tercer lugar, se ilustra un procedimiento de generación de nuevos indicadores de innovación basados en la minería de textos. Esto es de gran necesidad, puesto que la rápida innovación en algunos sectores impide disponer de información oficial sobre esta actividad. Finalmente, se presenta

un ejercicio experimental de integración de los datos de geolocalización de los teléfonos móviles, por parte del Instituto Nacional de Estadística (INE), para poder estimar en tiempo real la movilidad de forma alternativa a como se venía haciendo mediante respuestas a encuestas ligadas al Censo cada 10 años. A consecuencia de la pandemia, desde el sector privado, Google y Apple han hecho públicos indicadores de movilidad basados en la información de sus usuarios. El ejercicio emprendido por el INE ha permitido dar respuesta igualmente a la necesidad de proporcionar indicadores de movilidad en tiempo real desde el sector público.

Los ejemplos anteriores no pretenden ser una lista exahustiva del uso de datos masivos en análisis económico, sino que ilustran diversos aspectos de su utilización tales como la modificación de técnicas estadísticas existentes, la incorporación de nuevas bases de datos o la generación de nuevos indicadores ante la presencia de datos masivos en casos concretos, siendo todos ellos casos de éxito.

El segundo objetivo de este capítulo es analizar indicios de calidad necesarios para que los ejemplos anteriores puedan conducir a una práctica sistemática de la utilización de big data en análisis económico. Los distintos indicadores presentados en las aplicaciones se enfrentan a diferentes retos de cara a su validación. Aquellos dirigidos a monitorizar la actividad en tiempo real y nowcasting tienen una clara metodología de validación basada en la evaluación de los errores de predicción con los métodos habituales de exactitud predictiva, una vez publicado el indicador oficial que pretenden adelantar. Por el contrario, la validación no es inmediata en aquellas situaciones en las que no existe un indicador de referencia oficial al que asemejarlo, como es el caso de la generación de nuevos indicadores de innovación. Por último, este también sería el caso si el Censo de 2021 no incluye las preguntas referentes a movilidad que permitan comparar las estimaciones obtenidas con big data con las oficiales.

Este capítulo se organiza de la siguiente forma. Las secciones 2 a 5 presentan diversos ejemplos de aplicación de big data. Así, la sección 2 analiza la adaptación de técnicas tradicionales a grandes conjuntos de datos a través de un estudio de integración financiera entre países. La sección 3 muestra la utilidad de bases de datos basadas en noticias para la monitorización de la actividad en tiempo real y la predicción a corto plazo. La sección 4 presenta cómo la necesidad de generar nuevos indicadores de fenómenos ágiles y muy específicos, como los relacionados con la innovación, lleva a considerar la utilización de bases de datos alternativas. La sección 5 recoge los resultados de un estudio experimental llevado a cabo por el INE en relación con la movilidad a través del uso de datos de geolocalización. La sección 6 analiza las ventajas de los indicadores presentados e introduce algunos de los retos a los que nos enfrentamos para validar la utilización de los mismos de manera sistemática en análisis económico.

2. INTEGRACIÓN FINANCIERA: ADAPTACIÓN DE TÉCNICAS TRADICIONALES

A través del análisis de la integración financiera a nivel europeo, se ilustra la modificación de una técnica estadística tradicional, el análisis de conglomerados o cluster, para

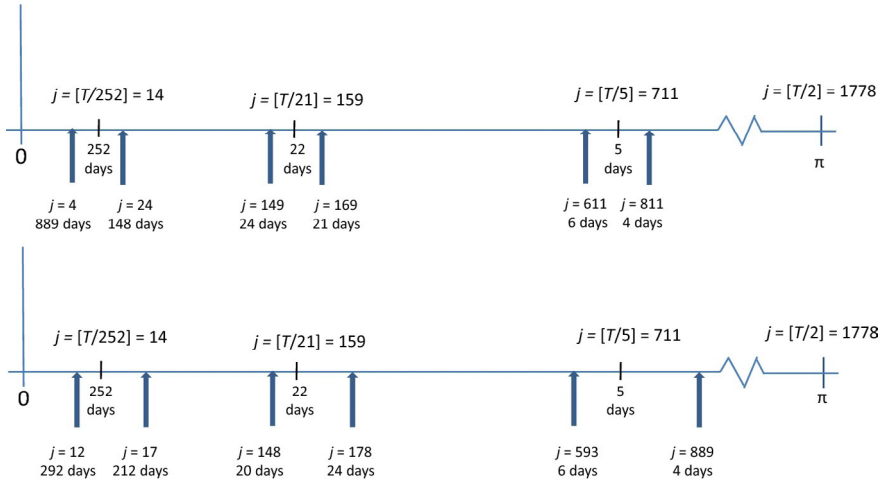
adaptarla al incremento del tamaño de la base de datos (Caiado, Crato y Poncela, 2020). El problema económico que justificó su adaptación es la necesidad de evaluar el grado de integración financiera en la Unión Europea (UE). La UE ha sido un gran catalizador para la liberación de los mercados en Europa. No obstante, la integración financiera no se ha conseguido aún y el sesgo nacional en las decisiones de inversión es notable. Tanto la Comisión Europea (CE) como el Banco Central Europeo (BCE) monitorizan el grado de integración financiera de manera continua (véase, por ejemplo, BCE, 2018, o Nardo *et al.,* 2017). Existen diversas formas de abordar dicha integración financiera, dependiendo de qué dimensión de la misma se desee analizar. Una de las formas es comprobar si se cumple la ley del precio único. Por ejemplo, si esta se da, el precio de un determinado activo debería ser el mismo en todos los mercados bursátiles. Para medir el grado de integración financiera a nivel europeo se analizan los grupos de índices financieros de acuerdo a su proximidad utilizando el análisis de conglomerados.

El análisis de conglomerados, aquí aplicado a series temporales, permite agrupar las mismas en base a su proximidad o semejanza. Esta se mide eligiendo una característica o propiedad P que defina el comportamiento de las series temporales y calculando la distancia entre las series atendiendo a esta propiedad. Piccolo (1990) propuso utilizar la distancia euclídea entre los coeficientes de la representación autorregresiva de las series. Desde una perspectiva no parámetrica, es decir, sin necesidad de estimar un modelo para las series, Galeano y Peña (2000) introdujeron medidas de distancia basadas en la función de autocorrelación de las mismas y, recientemente, Alonso y Peña (2019) han incluido la información sobre la dependencia lineal para formar los grupos. En la línea de las propuestas no paramétricas, Caiado *et al.* (2006) introdujeron los métodos en el dominio de la frecuencia y propusieron utilizar la distancia entre los periodogramas de las series. El periodograma mide la variabilidad (más precisamente el cuadrado de la amplitud de la onda) asociada a cada frecuencia. Es decir, descompone la varianza muestral asociada a una serie como la suma de las varianzas asociadas a cada frecuencia. Caiado *et al.* (2006) calculan la proximidad entre dos series calculando la distancia entre sus periodogramas. Dadas las observaciones de una serie temporal de longitud T, se define el periodograma para frecuencias angulares $w_j = (2\pi j)/T$, $j = 1, ..., [T/2]$, donde [z] denota la parte entera de z. Cuando el número de datos es muy elevado, para disminuir el número de cálculos que hay que realizar, Caiado *et al.* (2020) propusieron realizar el análisis de conglomerados calculando la distancia entre periodogramas no para todas las frecuencias sino sólo para aquellas asociadas a las principales fluctuaciones (aquellas donde la varianza es mayor). Esto puede ser de utilidad en macroeconomía para el análisis del ciclo de negocios o en finanzas. En este último caso es conocido que las series financieras diarias pueden presentar oscilaciones a las frecuencias diaria, semanal, mensual y anual. La propuesta de Corsi (2009) sobre modelos autorregresivos heterogéneos para volatilidad estocástica es un buen reflejo de este hecho, por lo que bastaría medir la distancia entre periodogramas sólo para ciertas frecuencias. Para ilustrar cómo se seleccionarían las frecuencias elegidas, supongamos, por ejemplo, que estamos analizando series diarias y estamos interesados solamente en el ciclo anual. Para series de tiempo de longitud T = 3556 (tamaño muestral de las series que posteriormente analizaremos en la aplicación sobre integración financiera), seleccionaremos ciclos de alrededor de 252 días laborables, es decir, fluctuaciones correspondientes a la frecuencia anual que, en este caso,

correspondería a la abscisa del periodograma j_s = T/252 = 14. Como es posible que exista cierta heterogeneidad en los ciclos anuales (por ejemplo, el número de días festivos no es el mismo en todos los países), nos gustaría seleccionar un intervalo alrededor de la frecuencia de interés, en este caso, I_{14}. Si este intervalo es simétrico alrededor de dicha frecuencia, por ejemplo, $I_{14} \pm 10 = [I_4; I_{24}]$ seleccionaríamos ciclos entre 148 y 889 días laborables. Aunque el intervalo es simétrico alrededor de la frecuencia de interés, es asimétrico en el número de días que consideramos alrededor de 252. Nuestra propuesta es usar intervalos que nos den ciclos de $\pm h$ días alrededor de un número de días dado, aunque resulten en intervalos asimétricos en la frecuencia. El mismo razonamiento se puede seguir para seleccionar el intervalo de frecuencias para captar el ciclo mensual y semanal. En la figura[1] 1 se han representado los ejes del periodograma. En el eje de abscisas, se señalan las principales frecuencias de variación de los índices considerados, así como intervalos simétricos en frecuencia (panel superior) o en el tiempo (panel inferior) alrededor de estas frecuencias. Únicamente compararíamos el periodograma para estos intervalos.

Figura 1.

Intervalos de frecuencias de muestreo en el periodograma. Panel superior: intervalos simétricos en frecuencia. Panel inferior: intervalos simétricos en el tiempo

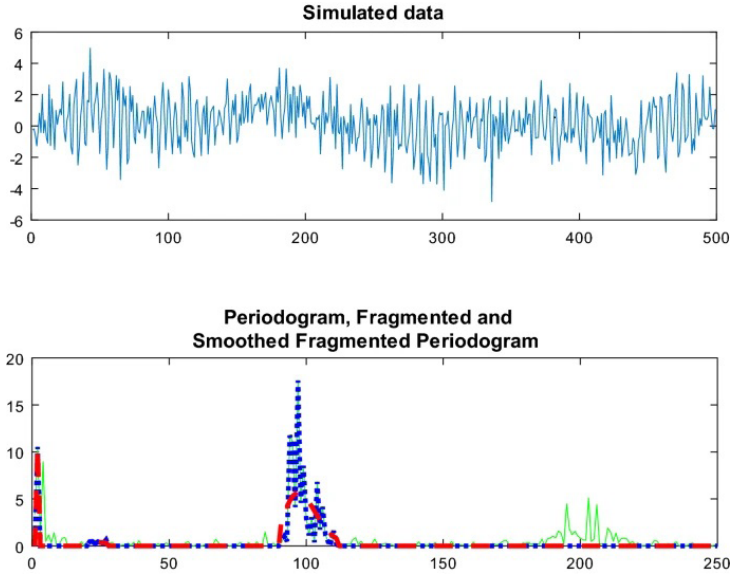


Para disminuir la varianza en la estimación del periodograma, Caiado *et al.* (2020) propusieron suavizarlo antes de proceder al cálculo de la distancia. La figura 2 muestra muestra en el panel superior un conjunto de datos simulados con el modelo de Corsi (2009) y en el panel inferior, el periodograma completo, el periodograma fragmentado para las frecuencias de interés y su suavizado. Mediante simulaciones, Caiado *et al.* (2020) comprobaron que usar el periodograma fragmentado suavizado aumenta la tasa de series correctamente clasificadas.

[1] Todos los gráficos de esta sección provienen del artículo de Caiado *et al.* (2020) y se reproducen aquí bajo licencia de Creative Commons.

Figura 2.

Panel superior: serie simulada con variación estacional semanal, mensual y anual. Panel inferior: periodograma (línea continua verde), periodograma fragmentado (línea punteada azul) y periodograma fragmentado suavizado (línea discontinua roja)



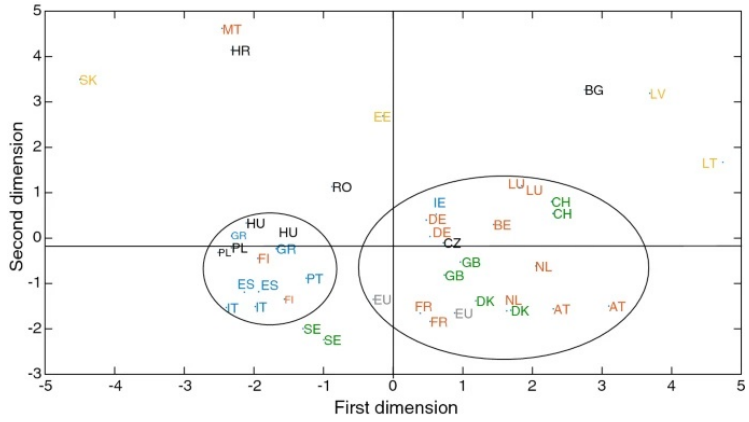
Caiado *et al.* (2020) aplican la propuesta anterior de clasificación para grandes conjuntos de datos a precios diarios de cierre de 44 mercados bursátiles[2] en Europa correspondientes al período comprendido entre el 2 de enero de 2003 y el 31 de diciembre de 2016. Para analizar el efecto que tuvo la crisis de deuda soberana en la integración financiera en Europa, se dividen las series en dos subperíodos, antes de la crisis (2 de enero de 2003 a 30 de junio de 2011) y después de la crisis de deuda soberana (1 de julio de 2011 a 31 de diciembre de 2016) y se analizan las series de retornos definidas como las tasas de variación relativas de los precios que se aproximan por las primeras diferencias del logaritmo. Se repite el siguiente ejercicio con cada submuestra para ver cómo cambian los resultados: se calcula el periodograma fragmentado suavizado de cada una de las 44 series y se calcula la matriz de distancias entre dichos periodogramas. Después de aplicar componentes principales a la matriz de distancias, se hace un gráfico de los 44 retornos financieros en función de las dos primeras componentes principales (escalado multidimensional). En la figura 3 se muestran los resultados correspondientes al período anterior a la crisis de deuda soberana y en la figura 4, después de la crisis. Se identifican las series con el acrónimo del país al que pertenece el índice bursátil[3]. Se recogen con distintos colores los clubs de países habitualmente analizados

[2] De algunos países, se consideran varios índices. Por ejemplo, para España, se dispone de las series correspondientes al IBEX 35 y al Índice General de la Bolsa de Madrid.

[3] Así, tanto el IBEX 35 como el Índice General de la Bolsa de Madrid aparecen como ES en las gráficas.

Figura 3.

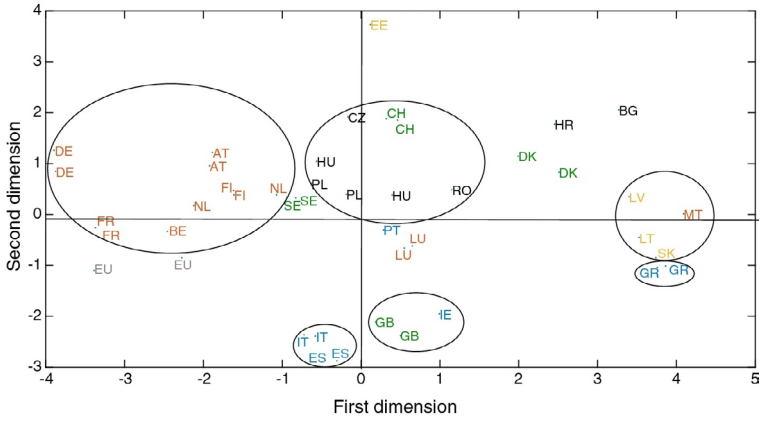
Escalado multidimensional de índices financieros en Europa antes de la crisis de deuda soberana. Muestra: 2/1/2003 a 30/6/2011



en integración financiera (Nardo *et al.,* 2017), rodeándose con un círculo los distintos grupos obtenidos mediante análisis de conglomerados jerárquico. En azul se muestran los países en dificultades o que han experimentado un deterioro significativo de la zona del euro (Irlanda, Grecia, España, Italia y Portugal); en rojo, los del núcleo de la zona del euro (Austria, Bélgica, Finlandia, Francia, Alemania, Luxemburgo y los Países Bajos); en verde, los del centro y oeste de Europa no pertenecientes a la zona del euro (Dinamarca, Gran Bretaña, Suecia y Suiza);

Figura 4.

Escalado multidimensional de índices financieros en Europa después de la crisis de deuda soberana. Muestra: 1/7/2011 a 31/12/2016



en naranja, los de la zona del este del euro (Estonia, Letonia, Lituania, Eslovaquia); en negro, los de la zona no euro oriental (Bulgaria, República Checa, Hungría, Polonia, Rumania y Croacia) y, finalmente, en gris, los índices europeos globales (eurostock50 y stxe600).

Desde el punto de vista de análisis económico, dichas figuras representan cómo la crisis de deuda soberana cambió el mapa de la integración financiera en Europa. Aunque hay indicios en la figura 3 de que existen dos grupos de países, los que no muestran dificultades financieras y los que sí lo hacen, en la figura 4 la separación entre países es mucho más clara. En el primer período aparecen en un mismo grupo los países sin dificultades, pertenezcan o no al área del euro, mientras que en la figura 4 los países del área euro sin dificultades forman un único grupo.

3. INDICADORES PARA LA MONITORIZACIÓN DE LA ACTIVIDAD ECONÓMICA EN TIEMPO REAL A TRAVÉS DE NOTICIAS

La generación de predicciones macroeconómicas, cada vez a horizonte temporal más corto, es clave en cuestiones de politica económica. Uno de los principales caballos de batalla de los modelos tradicionales de predicción macroeconómica es el retraso en la publicación de los datos oficiales. Esto ha dado lugar al desarrollo de modelos que incorporen información a distinta frecuencia mezclando, por ejemplo, datos mensuales y trimestrales para la generación de predicciones del producto interior bruto, y que tengan en cuenta los distintos calendarios de publicación de los datos incorporando la información tan pronto como está disponible, en vez de esperar a tener paneles balanceados de datos. Véase, por ejemplo, Camacho, Pérez-Quirós y Poncela (2013) y Banbura *et al.* (2013) para una revisión de los métodos de predicción de corto plazo y nowcasting. El uso de variables provenientes de encuestas (denominadas *soft data* en el argot), disponibles con anterioridad a las variables cuantitativas de actividad económica real (o *hard data*), sirve para paliar en parte la falta de estos últimos antes de su publicación en los modelos de predicción. Así, por ejemplo, Camacho y Pérez-Quirós (2010) utilizan un modelo factorial para sintetizar la información de diversos indicadores económicos a fin de generar predicciones de PIB de la eurozona que incorporan datos trimestrales (distintas estimaciones del PIB y desempleo) y mensuales, tanto *hard* (producción industrial,...) como *soft*. La versión para la economía española, Spain -STING, se recoge en Camacho y Pérez-Quirós (2011). Véase, por ejemplo, Stock y Watson (2016) o Bok *et al.* (2018) para una revisión de la literatura sobre el modelo factorial para la predicción macroeconómica y nowcasting o Poncela, Ruiz y Miranda (2021) quienes señalan la predicción en tiempo real y monitorización de la actividad económica como una de las principales aplicaciones del filtro de Kalman en análisis factorial dinámico revisando la literatura empírica sobre el tema.

No obstante, recientemente, han aparecido nuevas fuentes de datos alternativos, masivos, derivados del uso de tarjetas de crédito, noticias en prensa, en redes sociales, búsquedas por internet (por ejemplo, *Google trends*) o reseñas en Twitter (Loureiro y Alló, 2020) que están disponibles casi en tiempo real. Estas bases alternativas de datos masivos no son generadas por las agencias oficiales de estadística ni están diseñadas para reflejar el comportamiento de un

determinado agregado macroeconómico. Sin embargo, son de muy alta frecuencia, producidas en tiempo real, por lo que pueden ayudar a paliar las carencias de los modelos de predicción macroeconómica que se usan hoy en día cuando todavía no se dispone del dato oficial.

Una de las mencionadas bases de datos masivos alternativas lo constituyen las noticias en prensa. Estas han sido utilizadas de distintas maneras en análisis económico. Por una parte, se cree que reflejan la incertidumbre de la situación económica. La literatura actual argumenta que la incertidumbre afecta la actividad económica, véase, por ejemplo, Baker, Bloom y Davis (2016) quienes construyen un indicador de incertidumbre política y económica contando los artículos que contienen un determinado término. De manera análoga, el Banco de España ha construido indicadores de incertidumbre para la economía española y las principales economías latinoamericanas. (Véase, Ghirelli, Pérez y Urtasun, 2019; Ghirelli, Pérez y Urtasun, 2020, respectivamente). Otro uso de las noticias es explotar su contenido predictivo sobre el estado de la economía. Diversos bancos centrales, instituciones e investigadores han comprobado el poder predictivo de las noticias para los principales agregados macroeconómicos, en especial PIB, inflación y paro. Véase, para el Reino Unido, Rambaccussing y Kwiatkowski (2020), y Kalamara *et al.* (2020). Estos últimos encuentran que la información contenida en las noticias de los tres principales periódicos (*Daily Mirror, Daily Mail* y *The Guardian*) contiene poder predictivo sobre un simple modelo autorregresivo en predicciones con horizonte temporal de hasta nueve meses. La mejora predictiva de estos modelos disminuye considerablemente al incluir en los mismos factores extraídos de series de actividad económica, aunque siguen siendo útiles, sobre todo, en la vecindad de los puntos de giro, es decir, cuando cambia la fase del ciclo de negocios en las series de actividad real (no tanto para la inflación que, como apuntan Stock y Watson [2007], es difícil de predecir). El Banco de España encuentra resultados similares para predecir el PIB de España (véase Aguilar *et al.,* 2020) y Thorsrud (2020) encuentra el mismo tipo de resultados para el PIB trimestral de Noruega incorporando series temporales diarias extraídas de noticias en un modelo factorial. Para EE. UU., Barbaglia, Consoli y Manzan (2020) encuentran resultados análogos para la predicción de los principales agregados macroeconómicos utilizando indicadores de sentimiento construidos a partir de noticias de los seis principales periódicos de EE. UU. usando un total de 6,6 millones de artículos y 4, 2×10^9 palabras.

La utilización de las bases de datos basadas en noticias para la predicción en tiempo real y la monitorización de la actividad económica se lleva a cabo en tres pasos. El primer paso es construir la base de datos de noticias. El segundo paso consiste en pasar de estas bases de datos de noticias a series temporales que puedan ser utilizadas en los modelos de predicción. Para ello se pueden utilizar diversos algoritmos tales como contar las veces que aparece un determinado término, por ejemplo, en la prensa diaria o aquellos basados en diccionarios donde se da un valor entre -1 y +1 a un término específico dependiendo de si su connotación es negativa o positiva para lo que se analiza el entorno en el que aparece dicho término[4]. Por último, se utilizan modelos de aprendizaje automático (Kalamara *et al.,* 2020) o modelos econométricos más o menos sencillos para generar predicciones a muy corto plazo y

[4] En la siguiente sección referida a EURITO, se explican muy brevemente, algunas particularidades de los algoritmos de búsquedas de términos, por ejemplo, en relación a la presencia de sinónimos.

monitorizar el estado de la economía. Dentro de estos últimos, Barbaglia, Consoli y Manzan (2020) utilizan simples regresiones predictivas donde los regresores son los retardos de la variable a predecir que pueden ser aumentados con otros indicadores de actividad económica, mientras que Thosrud (2020) utiliza un modelo factorial dinámico.

En resumen, las noticias están disponibles a muy alta frecuencia, mucho antes de que dispongamos de los datos de estadísticas oficiales y son útiles para la predicción en tiempo real de la actividad económica y la monitorización de la economía. Para usarlas en modelos de previsión macroeconómica hay que pasar de los datos de texto a series temporales que podamos introducir en los modelos de predicción. Aunque una vez que se publican las estadísticas oficiales mensuales (datos construidos para reflejar el comportamiento de una determinada variable económica), los indicadores basados en noticias pueden perder su capacidad predictiva para el PIB trimestral, siguen siendo útiles en la vecindad de los puntos de cambio (Kalamara *et al.,* 2020), sobre todo en las recesiones (Barbaglia, Consoli y Manzan, 2020) y se han mostrado de gran utilidad en la recesión debida a la COVID-19 donde los indicadores basados en encuestas (los primeros disponibles informativos sobre el estado de la economía) no han reflejado a tiempo ni correctamente el *shock* económico (Aguilar *et al.* 2020).

4. EURITO: GENERACIÓN DE INDICADORES EN ÁREAS VÍRGENES

EURITO Research and Innovation Indicators (EURITO, 2018) es un proyecto financiado por la Comisión Europea, realizado por un consorcio de cuatro organizaciones: Nesta (Reino Unido), Fraunhofer (Alemania), Danmark Tekniske Universitet (DTU, Dinamarca) y la Fundación Cotec para la Innovación (España). El objetivo de EURITO es elaborar indicadores de desarrollo de la investigación e innovación a partir de la huella digital de sus actividades, en un área en la que los indicadores disponibles (European Innovation Scoreboard) están principalmente basados en encuestas o resultados como patentes y publicaciones que generalmente ofrecen una visión demasiado agregada y retardada de su desarrollo. EURITO es el acrónimo formado a partir del título del proyecto "EU Relevant, Inclusive, Timely, Trusted and Open Research Innovation Indicators" que resume los objetivos del mismo: indicadores EURopeos de innovación relevantes que puedan proporcionar la información necesaria para el conocimiento y desarrollo de políticas de investigación y desarrollo; Inclusivos permitiendo que la cobertura de los indicadores se extienda más allá de la investigación en disciplinas de las STEM o de empresas altamente tecnológicas, incluyendo la innovación en servicios o sectores menos tecnológicos, así como la consideración de redes informales; confiables en Tiempo real, garantizando la representatividad, calidad, oportunidad y pronta publicación, indicando sus limitaciones y validados por los investigadores y la industria; Abiertos (Open), de forma que tanto las fuentes de datos originales como la metodología y códigos de programación empleados permitan reproducir y multiplicar las oportunidades para su mejora, extensión y aplicación.

A modo de ejemplo, describimos la elaboración de indicadores sobre tecnologías emergentes desarrollado en el primer estudio piloto del proyecto (Nesta, 2019). La medición

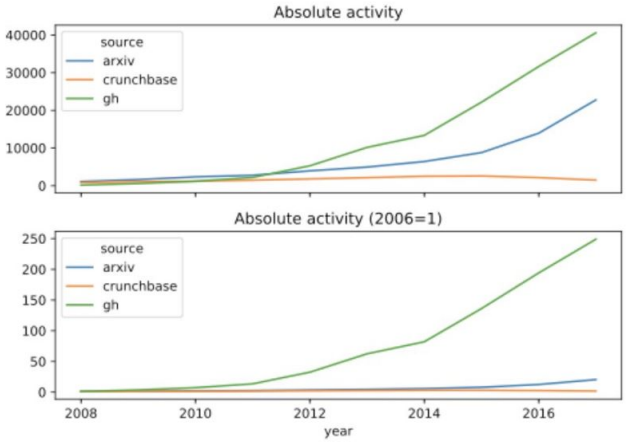
de la aparición y desarrollo de una tecnología emergente es relevante en sí misma, así como su impacto en la aplicación a la creación de nuevos productos, servicios o actividades de investigación. Prácticamente por definición una tecnología emergente, es susceptible de caer fuera de taxonomías preestablecidas utilizadas en el análisis industrial, científico y tecnológico lo que dificulta y retrasa su medición mediante indicadores tradicionales, Bakhshi y Mateos-García (2016). Sin embargo, las descripciones en formato texto incluidas en los documentos en repositorios de investigación científica, información sobre financiación, patentes, páginas webs de las compañías, plataformas de colaboración, entre otros, pueden considerarse como nuevas bases de datos no estructuradas que permitan de una forma ágil medir el desarrollo de una tecnología emergente.

El estudio piloto se centra en el desarrollo de la *inteligencia artificial,* entendida como el conjunto de tecnologías que usan datos y métodos de aprendizaje automático o *machine-learning* (Mateos-García, 2018). Los indicadores que se persiguen son el nivel de investigación en inteligencia artificial en la Unión Europea y su evolución en el tiempo, comparado con sus principales competidores en agregado y entre los Estados miembros, entre otros. A tal fin, se puede encontrar información en distintas bases de datos que recogen repositorios de artículos de investigación (arXiv, Microsoft Academic Graph), software (GitHub), financiación (CORDIS), patentes (PATSTAT), nuevas *startups* y compañías tecnológicas (CRUNCHBASE), redes de colaboración (Meetup) o habilidades demandadas y ofertadas (webs universidades y anuncios de empleo). Las bases de datos son heterogéneas y muestran importantes diferencias y problemas para la construcción de un indicador de actividad basado en el contenido de los textos. Una primera dificultad se encuentra, por ejemplo, en la distinta longitud de las descripciones entre las fuentes de datos lo que puede llevar a que las medidas de relevancia no sean comparables. También es preciso tener en cuenta la necesidad de disponer de los metadatos, y conocer indicadores que identifiquen cada observación como el sexo o la localización geográfica. En EURITO esta información estaba disponible en algunas bases de datos, pero donde no era el caso, se ha aproximado mediante algoritmos que identifican, por ejemplo, la localización geográfica a partir del nombre de la institución del investigador.

Como una primera aproximación para desarrollar el indicador de actividad, el estudio piloto selecciona tres bases de datos de acceso abierto con cobertura europea: Crunchbase, arXiv y GitHub. El objetivo es identificar las entradas relacionadas con el concepto de inteligencia artificial mediante la aplicación de técnicas de procesamiento natural del lenguaje. El procedimiento se inicia con una palabra clave (en nuestro caso *Artificial Intelligence*) y la búsqueda de sinónimos en un espacio multidimensional mediante similaridad semántica (algoritmo word2vec, Mikolov, Yih and Zweig, 2013). Esta búsqueda selecciona aquellas palabras que aparecen en un contexto similar a la palabra clave original. En el desarrollo del indicador se corrige la posibilidad de que algunos sinónimos sean demasiado genéricos y produzcan un elevado número de documentos irrelevantes. Para ello se eliminan de la lista de palabras claves aquellas con bajo *TF-IDF (Term-Frequency Inverse-Document Frequency)* que normaliza el número de veces que un término aparece en un documento por el número de veces que aparece en un *corpus*.

Figura 5.

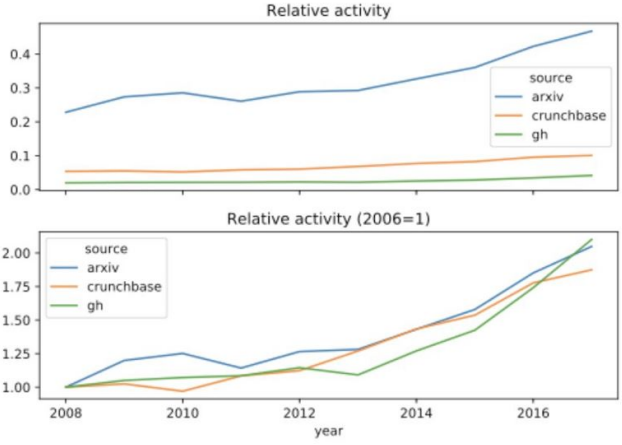
Número de documentos relacionados con actividad artificial en números absolutos e índice (2006=1)



Las figuras 5 y 6 muestran la evolución del nivel de actividad absoluta y relativa en inteligencia artificial en las tres bases de datos seleccionadas: Crunchbase, arXiv y GitHub (denotada como gh en el gráfico)[5]. Mientras que el nivel de actividad absoluto cuenta

Figura 6.

Proporción de documentos, relacionados con actividad artificial en números absolutos e índice (2006=1)



[5] Estos gráficos se toman del entregable relacionado con el primer estudio piloto (Nesta, 2019) y se reproducen con consentimiento de los autores.

directamente el número de entradas relacionadas con la inteligencia artificial, el nivel en términos relativos normaliza esta medida comparando con el total de entradas en cada base de datos.

Tal y como se observa, el desarrollo de la inteligencia artificial ha crecido tanto en números absolutos como en términos relativos en las tres fuentes de datos consideradas, siendo la más representativa arXiv, donde el crecimiento acumulado desde 2008 es del 250 % llegando a suponer el 40 % de las publicaciones totales en esta base de datos en el año 2017.

5. MOVILIDAD A PARTIR DE LA GEOLOCALIZACIÓN DE TELÉFONOS MÓVILES: NUEVOS INDICADORES PARA VIEJAS PREGUNTAS

Desde un punto de vista de política económica y social, el conocimiento de las cifras de población es necesario en muchos aspectos relacionados como por ejemplo la asignación de recursos públicos. El análisis de la movilidad cotidiana (por motivos laborales o educativos) y de la movilidad estacional (relacionado con el turismo nacional o internacional) resulta igualmente necesario para reconocer en cada momento las necesidades reales de los distintos territorios. El Censo de Población y Viviendas es una operación estadística elaborada por el Instituto Nacional de Estadística que se realiza cada diez años y permite conocer las características de las personas, hogares, edificios y viviendas. El Censo de 2011 fue novedoso pues incluyó, por primera vez, registros administrativos y encuestas dirigidas al 10 % de la población. Nuevamente, el INE se enfrenta a un reto metodológico, elaborando el Censo de Población y Viviendas de 2021 basado totalmente en el uso de registros. Dentro de los trabajos preparatorios del Censo 2021, el INE, bajo la categoría de *Estadística Experimental,* ha realizado un estudio de movilidad (EM1) a partir de registros georeferenciados de telefonía móvil, como fuente de datos alternativa a los cuestionarios censales (INE, 2020a).

Los resultados obtenidos provienen del análisis de la posición de más del 80 % de los teléfonos móviles en toda España, con la colaboración de los tres principales operadores de telefonía móvil (Orange, Telefónica y Vodafone). En relación con el ámbito de la investigación, se considera a la población residente en España (se excluyen los teléfonos de numeración extranjera) para el total del territorio español dividido en 3.214 *áreas de movilidad INE* (agrupación de municipios con menos de 5.000 empadronados hasta que superan esta cantidad, municipios entre 5.000 y 50.000 empadronados, o divisiones de aquellos municipios con más de 50.000 empadronados en barrios -*SCD, sub-city districts*).

A partir de los datos de posición de los teléfonos durante una semana laboral de referencia (los días 18 a 21 de noviembre de 2019), el INE proporciona la matriz de movilidad cotidiana que permite conocer la población residente en el área (según Padrón a 1 de enero de 2019), la población residente que se mantiene en su área, la población que llega al área, la población detectada durante el día en el área, la variación de población, los destinos a los que se desplazan sus residentes y el origen de los que llegan al área. De cara a la movilidad, se define el área de residencia del teléfono móvil como aquella donde el teléfono se encuentra con mayor frecuencia durante el período previo al considerado (entre dos y tres meses según

el operador). Asimismo, el área de destino se define como la más frecuente fuera del área de residencia en la que se encuentra el terminal al menos durante cuatro horas al día en la franja horaria de 10:00 a 18:00 y al menos dos días de los cuatro observados de la semana laboral normal de referencia. Esta información permite conocer casi en tiempo real, los movimientos de la población a un elevado nivel de desagregación geográfica, así como los trayectos que realiza.

La primera columna de la tabla 1 recoge las estimaciones de movilidad diurna obtenidas en el área de *Madrid (SCD Sol)*. SCD Sol es un área de Madrid central que cuenta con 7.309 residentes, de los cuales 2.026, el 27,7 %, se mantienen en el área y, al mismo tiempo, fue área de destino de 14.790 personas no residentes, más del doble de la población que procedían de más de 190 orígenes identificables. Estas estimaciones señalan esta área como fuerte receptora de población diurna de manera cotidiana.

Tabla 1.

Movilidad en Madrid - SCD Sol

	Semana referencia	*Fechas estacionales*			
Variable	*18-21 noviembre*	*20 julio*	*15 agosto*	*24 noviembre*	*25 diciembre*
Población residente	7.309	7.309	7.309	7.309	7.309
que se mantiene en el área	2.026	3.100	2.665	3.652	2.466
Población total					
detectada durante el día	16.816				
que pernocta en el área		16.839	15.710	31.390	30.609
Variación de población	12.653	9.530	8.401	24.081	23.300

Adicionalmente, el INE proporciona información sobre la población que pernocta fuera de su área de residencia, lo que permite complementar la información sobre los movimientos diurnos por motivos laborales o educativos con la estimación de la variación debida a causas estacionales. Se proporciona información sobre dos fechas de agosto (20 de julio y 15 de agosto), un domingo normal (24 de noviembre, continuidad de la semana de referencia considerada en la movilidad diurna) y el día de Navidad (25 de diciembre). Se consideran en esta ocasión todos los teléfonos presentes en territorio español en esas fechas y se determina el área de pernoctación a partir del lugar más frecuente en el que se encuentra el aparato desde las 22:00 horas del día anterior hasta las 6:00 horas en esa fecha.

Las columnas 2 a 5 de la tabla 1 recogen las estimaciones de movilidad estacional en las fechas señaladas. La población residente no cambia, sigue siendo la correspondiente a fecha 1 de enero de 2019 del Padrón. Estas estimaciones permiten señalar *Madrid Sol* como un destino fuertemente receptor de turistas nacionales tanto en las fechas consideradas de verano como de invierno. Se identifica el mayor atractivo del área de *Madrid Sol* en las fechas de invierno frente al verano, puesto que mientras que en las fechas del 20 de julio y 15 de

agosto las pernoctaciones son de algo menos del doble de la población residente, el 24 de noviembre[6] y el 25 de diciembre prácticamente las multiplican por 4.

5.1. La incidencia de la COVID-19 en la movilidad

La metodología empleada en la generación de la matriz de movilidad cotidiana ha permitido al INE, en colaboración con los tres principales operadores de telefonía móvil, estimar el movimiento de la población desde la irrupción de la pandemia de la COVID-19 y la declaración del estado de alarma desde el 16 de marzo hasta la actualidad, INE (2020b, 2020c). En una primera fase, durante el estado de alarma, el INE utilizó las mismas áreas de movilidad cotidiana y redefinió el área de residencia como aquella donde el teléfono ha pasado la mayor parte del tiempo en el horario de 0:00 a 6:00 horas y el área de destino como la más frecuente en el período de 10:00 a 16:00 horas con un mínimo de dos horas, respectivamente. Los resultados obtenidos, elevados al total de la población, han permitido conocer la variación de la movilidad frente a un día de una semana de referencia *normal* coincidente con la utilizada en el estudio de la movilidad cotidiana. Desde el 24 de junio al 30 de diciembre, se ha venido recogiendo la información de movilidad los miércoles y los domingos, recuperando las definiciones de área de residencia y destino de la matriz de movilidad cotidiana.

La figura 7 recoge la evolución del porcentaje de población que sale del área de residencia en el total nacional y en el área de *Madrid Sol* a partir de estos datos. El primer punto señala el porcentaje de movilidad en la semana de referencia de 2019[7] que era, tanto en el total nacional como en *Madrid Sol* cercana al 30 %, cifra que no se ha recuperado hasta la fecha, siendo el máximo a nivel nacional del 22 % en los días laborables de junio y julio y del 23 % el domingo 29 de noviembre en *Madrid Sol.* Destaca en la figura 7 la dualidad entre los días laborables y los fines de semana a nivel nacional durante el estado de alarma, que también se mantiene en la segunda mitad de 2020 situándose la movilidad los miércoles (línea punteada azul superior) alrededor del 20 % frente a los domingos (línea punteada azul inferior) que se reduce entre el 10 y el 15 %.

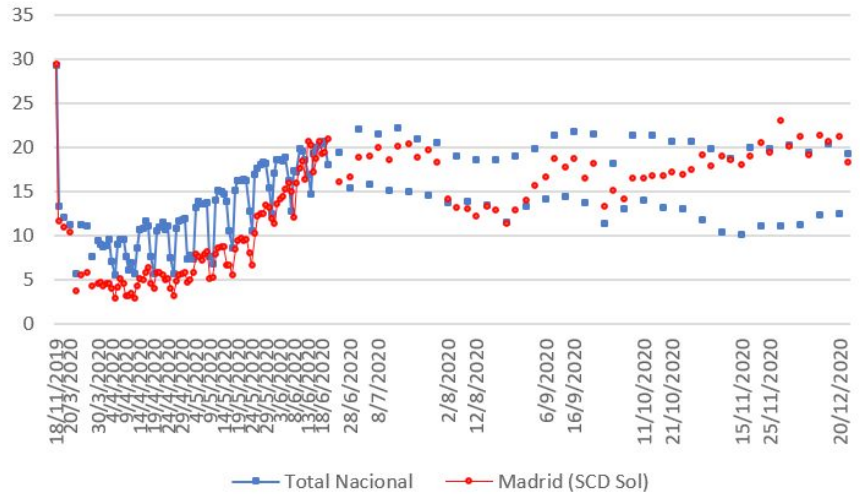
La figura 7 también permite analizar el distinto comportamiento en *Madrid Sol.* Por una parte, durante el primer período del confinamiento, donde la movilidad fue mucho menor que en el total nacional, manteniendo cifras alrededor del 3 % durante el mes de abril y no alcanzando el porcentaje de movilidad nacional hasta el mes de junio. Con posterioridad la diferencia se manifiesta principalmente en la inexistencia de comportamiento dual entre día laborable y domingo.

[6] Las cifras no son comparables con las de movilidad cotidiana de la primera columna, puesto que antes se recogía movilidad diurna y teléfonos españoles y ahora son pernoctaciones y todos los teléfonos presentes.

[7] En la figura, se presenta con una línea continua la información diaria durante el estado de alarma hasta el 23 de junio y, mediante puntos espaciados la información correspondiente a los miércoles y domingos desde el 24 de junio hasta el final de la muestra.

Figura 7.

Movimiento de personas por áreas de movilidad
(Porcentaje)



5.2. Otros indicadores de movilidad: Google y Apple

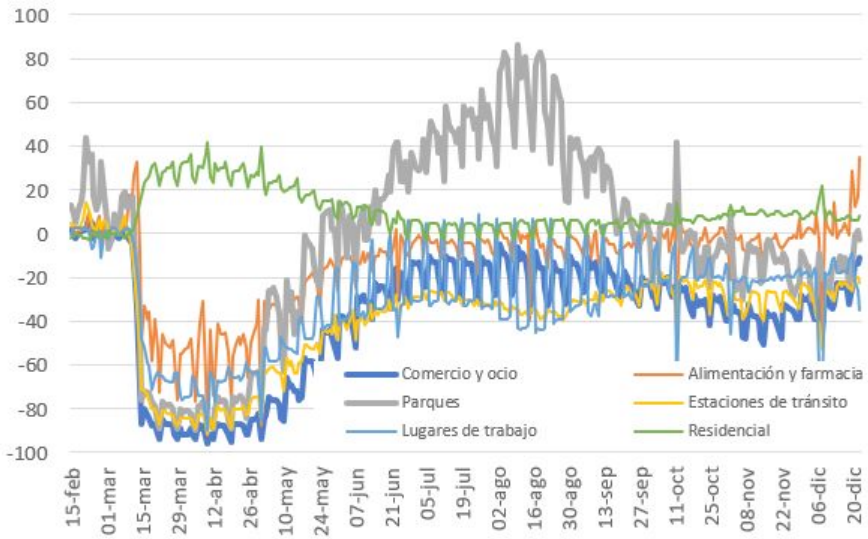
Alternativamente, Google (2020) y Apple (2020) están utilizando la información que proporcionan sus usuarios para generar indicadores de movilidad. Estas fuentes de datos ya están siendo utilizadas en el análisis económico, véase por ejemplo, Woloszko (2020) que monitoriza la actividad económica utilizando los indicadores de movilidad de Google para Kartal, Depren y Depren (2020) que analizan la reacción de los principales índices bursátiles de los países de Asia del este ante la COVID-19 utilizando los datos de Apple, entre otros.

Google proporciona estadísticas a partir de los datos de los usuarios que han habilitado el historial de ubicaciones de su cuenta. En concreto, informa sobre las tendencias de movimiento a lo largo del tiempo ordenadas por zonas geográficas y clasificadas en diversas categorías de lugares atendiendo a su actividad: *Comercio y Ocio, Alimentación y farmacias, Parques, Estaciones de tránsito, Lugares de trabajo y Residencial.* Los datos recogen el cambio en el número de visitas a dichos lugares, clasificados por su actividad, y su duración en comparación al valor medio de cada día de la semana durante un período de cinco semanas desde el 3 de enero hasta el 6 de febrero de 2020. Los datos están disponibles diariamente desde el 15 de febrero de 2020.

La figura 8 presenta los datos proporcionados por Google por actividad en tipo de destino.

Figura 8.

Tendencias de movilidad de Google
(Total España)



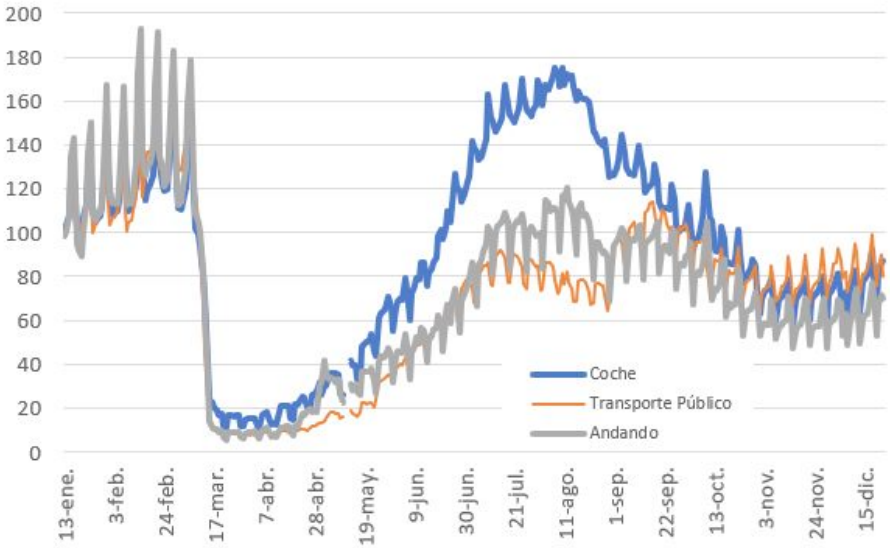
La figura 8 refleja claramente la falta de movilidad del confinamiento, siendo la categoría *Residencial* la única con valores positivos, en relación con el período de referencia pre-COVID-19, y la posterior relajación de las medidas que mantienen esta categoría en niveles positivos pero de menor magnitud. Igualmente, la figura 8 nos recoge la consecuente reducción en las visitas al resto de lugares durante el confinamiento, siendo el que menor descenso sufre la categoría de *Alimentación y Farmacia,* que aún así ve reducidas las visitas hasta valores alrededor del 60 % con respecto al período de referencia. Con posterioridad al confinamiento, la única categoría que alcanza niveles superiores al período de referencia han sido las correspondiente a *Parques* durante el período estival.

Apple también proporciona informes diarios de tendencias de movilidad a partir de las solicitudes de indicaciones en *Mapas de Apple.* En concreto, publica información diaria sobre el porcentaje del número de consultas realizadas en comparación con las realizadas el 13 de enero. Se dispone de datos sobre solicitudes de traslados en coche, transporte público o a pie, para el total nacional, por comunidades autónomas y las cuatro ciudades más grandes de España: Madrid, Barcelona, Valencia y Sevilla. La figura 9 muestra la evolución de la movilidad según la información proporcionada por Apple.

Los datos recogidos en la figura 9, al igual que las cifras de INE y las de Google, indican que existe una fuerte estacionalidad semanal a nivel nacional y un fuerte impacto de la pandemia en la movilidad. En el caso de Apple la información proporcionada muestra una

Figura 9.

Tendencias de movilidad de Apple
(Total España)



mayor recuperación de la movilidad mediante el uso del automóvil, siendo la única categoría que llega a alcanzar cifras superiores al 13 de enero de 2020 (pre-COVID), durante los meses de verano.

6. ALGUNOS RETOS Y VENTAJAS A PARTIR DE LOS EJEMPLOS

Los ejemplos anteriores muestran cómo las técnicas estadísticas y las nuevas bases de datos resultan prometedoras en el análisis económico y, en especial, en la generación de indicadores. Hemos visto que mediante técnicas de big data se pueden abordar problemas económicos relevantes, bien actualizando técnicas estadísticas existentes (integración financiera), generando nuevos indicadores en materias donde no existía información oficial (EURITO), adelantando indicadores oficiales (noticias) o proporcionando información en tiempo real de variables que antes se estimaban mediante encuestas cada diez años (movilidad).

No obstante, para su generalización y utilización en el análisis económico los indicadores deben validar su capacidad para representar fielmente la realidad que persiguen identificar. Un problema que es diferente dependiendo de cada una de las aplicaciones. En el caso de los indicadores basados en noticias y, en general, todos aquellos dirigidos a la predicción económica o el nowcasting, la validación viene de la mano de su utilidad para adelantar la información oficial que será publicada posteriormente. EURITO y los indicadores de

movilidad proporcionan nuevas variables en materias donde no existe información oficial publicada al respecto con la que comparar. La validez o credibilidad de estos datos deben fundamentarse en su adecuación a estándares de calidad, como los recogidos en el Código de Buenas Prácticas del Sistema Estadístico Europeo (ESS, 2017). Algunos de estos estándares son: Relevancia; Exactitud y Fiabilidad; Prontitud y puntualidad en la publicación de los datos; Coherencia y Comparabilidad, Accesibilidad y Claridad.

La ausencia de sesgos y la fiabilidad de los datos de partida y de las estadísticas generadas es uno de los retos pendientes. Realmente, con la irrupción del big data, ha cambiado el paradigma clásico de muestreo. La oportunidad de obtener información mediante bases de datos alternativas, no diseñadas para el propósito del estudio, puede apartar la posibilidad de mantener un esquema de inferencia tradicional. Por una parte, los tamaños de muestra tan grandes nos llevan a pensar que estamos prácticamente considerando la población y que los errores de muestreo son negligibles. No obstante, se pueden presentar dos problemas. El primero es que, al no disponer de un diseño muestral, estemos incurriendo en errores de cobertura y exista un importante sesgo por selección de muestra en los posibles análisis derivados. El segundo es que los errores de medida de los datos no sean completamente aleatorios y se propaguen con el tamaño de muestra en vez de compensarse. Además, es preciso que las modificaciones de las técnicas estadísticas que llevan a cabo los procedimientos de aprendizaje automático para adaptarse a las nuevas necesidades del big data, garanticen que cubren los términos que se refieren al problema de interés. Así, por ejemplo, en la aplicación sobre integración financiera, si dejamos fuera alguna frecuencia de fluctuación importante, podemos omitir información relevante para la comparación entre índices. O en el caso de EURITO, si el procedimiento de similaridad semántica utilizado no es capaz de recoger el espectro semántico relevante para el problema en cuestión.

El reto del análisis de la coherencia de los indicadores obtenidos depende del ejemplo analizado. Los indicadores construidos para monitorizar la actividad en tiempo real tienen una clara metodología de validación de su coherencia basada en la evaluación de los errores de predicción, con los métodos habituales de exactitud predictiva, una vez publicado el indicador oficial que pretenden adelantar. La evaluación de la coherencia no es inmediata en aquellas situaciones donde no existe un indicador de referencia oficial. En el caso de EURITO, el alto nivel de desagregación sectorial y geográfico y las características de los fenómenos de innovación estudiados, no permiten una validación cuantitativa clara más allá de la comparación con grandes agregados de I+D+i y el análisis de su coherencia geográfica y temporal. En el caso de las variables de movilidad, la publicación de distintos indicadores a partir de fuentes de información diferentes, aunque a distinto nivel de desagregación sectorial y geográfico, quizás pueda servir para estudiar la coherencia entre las series a partir de un estudio de características comunes, entre aquellas que se aproximen al mismo fenómeno.

Otro tipo de cuestiones de carácter institucional no tratadas en este capítulo vienen recogidas en Salgado (2017) y Salgado y Oancea (2020). Entre estas se encuentran el acceso al dato (privacidad, continuidad y coste de las fuentes de información), los recursos tecnológicos necesarios (humanos y de capital), la independencia profesional en la generación de los datos, la coordinación y la cooperación público-privada, entre otras.

Este recorrido a través de los ejemplos no pretende, ni mucho menos, ser exhaustivo sobre los retos a los que hay que enfrentarse, sino que simplemente intenta señalar alguno de los más relevantes que trascienden tras el simple análisis de estos cuatro ejemplos.

Referencias

Aguilar, P., Ghirelli, C., Pacce, M. y Urtasun, A. (2020). Can news help to measure economic sentiment? An application in Covid-19 times. *Documento de Trabajo,* No. 2027, Banco de España.

Alonso, A. M. y Peña, D. (2019). Clustering time series by linear dependency. *Statistics and Computing,* 29, pp. 655–676.

Apple (2020). Informes de tendencias de movilidad. https://covid19.apple.com/mobility

Baker, S. R., Bloom, N. y Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics,* 131(4), pp. 1593-1636.

Bakhshi, H. y Mateos-García, J. (2016). New data for innovation policy. *Working Paper.* London: Nesta.

Banbura, M., Giannone, D. M., Modugno, M. y Reichlin, L. (2013). Now-casting and the real-time data flow. En: *Handbook of Economic Forecasting,* 2, pp. 195-237.

Barbaglia, L., Consoli, S. y Manzan, S. (2020). *Forecasting with Economic News.* Manuscrito disponible en SSRN: https://ssrn.com/abstract=3698121

BCE (2018). *Financial Integration in Europe.* op.europa.eu/en/publication-detail/-/publication/c7c3826a-526a-11e8-be1d-01aa75ed71a1/language-en

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M. y Tambalotti, A. (2018). Macroeconomic now-casting and forecasting with big data. *Annual Review of Economics,* 10, pp. 615-643.

Caiado, J., Crato, N. y Peña, D. (2006). A periodogram-based metric for time series classification. *Comput Stat Data Anal,* 50, pp. 2668–2684.

Caiado, J., Crato, N. y Poncela, P. (2020). A fragmented-periodogram approach for clustering big data time series. *Advances in Data Analysis and Classification,* 14, pp. 117-146.

Camacho, M. y Perez-Quirós, G. (2010). Introducing the EURO-STING: Short Term INdicator of Euro Area Growth. *Journal of Applied Econometrics,* 25, pp. 663-694.

Camacho, M. y Perez-Quirós, G. (2011). Spain-Sting: Spain Short-Term Indicator of Growth. *The Manchester School,* 79, pp. 594–616.

Camacho, M., Perez-Quirós, G. y Poncela, P. (2013). Short-term forecasting for empirical economists. A survey of the recently proposed algorithms. *Foundations and Trends in Econometrics,* 6, pp. 101-161.

Corsi, F. (2009). Heterogeneous autoregressive model of realized volatility (HAR-RV). *J Financ Econom,* 7, pp. 174–196.

EURITO (2020). *EU Relevant, Inclusive, Timely, Trusted, and Open Research Innovation Indicators.* http://www.eurito.eu/

Galeano, P. y Peña, D. (2000). Multivariate analysis in vector time series. *Resenhas,* 4, pp.383–404

Ghirelli, C., Pérez, J. J. y Urtasun, A. (2019). A new economic policy uncertainty index for Spain. *Economics Letters,* 182, pp. 64-67.

Ghirelli, C., Pérez, J. J. y Urtasun, A. (2020). Economic Policy Uncertainty in Latin America. *Documento de Trabajo,* No. 2024, Banco de España.

Google (2020). *Informes de movilidad local sobre COVID-19.* https://www.google.com/covid19/mobility/

INE (2020a). *Estudio EM-1 de movilidad a partir de la telefonía móvil.* Diciembre 2020. http://www.ine.es/experimental/movilidad/experimental_em.htm

INE (2020b). *Análisis de la movilidad de la población durante el estado de alarma por COVID-19 a partir de la población de los teléfonos móviles.* Junio 2020. https://www.ine.es/covid/exp_movilidad_covid_proyecto.pdf

INE (2020c). *EM-3 - Estudio de movilidad a partir de la telefoníma móvil durante el período julio-diciembre 2020 (EM-3).* Noviembre 2020. https://www.ine.es/experimental/movilidad/exp_em3_proyecto.pdf

Kalamara, E., Turrell, A., Kapetanios, G., Kapadia, S. y Redl, C. (2020). Making text count: economic forecasting using newspaper text. *Bank of England Staff Working Paper,* No. 865.

Kartal, M. T., Depren, S. K. y Depren, Ö. (2020). How Main Stock Exchange Indices React to Covid-19 Pandemic: Daily Evidence from East Asian Countries. *Global Economic Review.* DOI: 10.1080/1226508X.2020.1869055.

Loureiro, M. y Alló, M. (2020). Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the U.K. and Spain. *Energy Policy,* 143, 111490.

Mateos-García, J. (2018). *The Complex Economics of Artificial Intelligence.* Disponible en SSRN 3294552.

Mikolov, T., Yih, W. y Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. En: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751). Atlanta, Georgia: Association for Computational Linguistics. Retrieved from: http://www.aclweb.org/anthology/N13-1090

Nardo, M., Ndacyayisenga, N., Papanagiotou, E., Rossi, E. y Ossola, E. (2017). Measures and drivers of Financial Integration in Europe. European Commission, Joint Research Centre. *Report EUR 28469 EN.* doi:10.2760/92134.

Nesta (2019). *Pilot 1: Emerging Technology Ecosystems.* http://www.eurito.eu/pilots-and-indicators/

Piccolo, D. (1990). A distance measure for classifying ARIMA models. *J Time Ser Anal,* 11, pp. 152–164.

Poncela, P., Ruiz, E. y Miranda, K. (2020). Factor extraction using Kalman filter and smoothing: this is not just another survey. *International Journal of Forecasting* (en prensa).

Rambaccussing, D. y Kwiatkowski, A. (2020). Forecasting with news sentiment: Evidence with UK newspapers. *International Journal of Forecasting,* 36, pp. 1501-1516.

Salgado, D. (2017). Big Data en la Estadística Pública: Retos ante los primeros pasos. *Revista de Economía Industrial.* 3 trimestre, pp. 121-129.

Salgado, D. y Oancea, B. (2020). On new data sources for the production of official statistics. *Working Paper,* 01/2020. Instituto Nacional de Estadística (INE).

Stock, J. H. y Watson, M. W. (2007). Why Has U.S. Inflation Become Harder to Forecast? Journal of Money, Credit and Banking, 39, Issue s1.

Stock, J. H. y Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. *Handbook of Macroeconomics,* 2, pp. 415-525.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of business cycles. *Journal of Business and Economic Statistics,* 38(2), pp. 393-409.

Woloszko, N. (2020). Tracking activity in real time with Google Trends. *OECD Economics Department Working Papers,* N. 1634. https://doi.org/10.1787/18151973

CAPÍTULO VI

Análisis de factores comunes estacionales en datos masivos

Fabio H. Nieto*
Daniel Peña
Stevenson Bolívar

Las variables que se estudian en economía y finanzas tienen con frecuencia un comportamiento estacional. Al analizar este tipo de variables para hacer inferencia sobre su dinámica no conviene desestacionalizarlas, ya que esto supone aplicar filtros de corrección estándar de la estacionalidad que pueden no ser adecuados para las variables estudiadas. En la actualidad podemos analizar conjuntos muy grandes de variables, dada la presencia cada vez más frecuente de datos masivos, y los métodos tradicionales para tratar la estacionalidad son difíciles de aplicar en estos casos. Una manera efectiva de hacerlo es mediante el análisis factorial, que resume todas las relaciones de dependencia comunes, incluyendo la estacional, en un conjunto pequeño de factores, permitiendo además una dinámica específica de cada serie. En este capítulo presentamos la metodología para construir un modelo factorial estacional e ilustramos su aplicación en un conjunto de variables macroeconómicas estacionales que corresponden al dinero en efectivo en circulación en algunos países latinoamericanos.

Palabras clave: datos masivos, estacionalidad, modelo factorial dinámico, series temporales multivariadas; procesos integrados.

* Los autores del presente trabajo agradecen a Luis Fernando Melo, investigador en econometría del Banco de La República, el Banco Central de Colombia, por sus sugerencias sobre el tipo de variables a utilizar en este estudio, y a Fabio D. Nieto-Mosquera, economista jefe del Banco Agrario de Colombia, por su ayuda en la interpretación económica de los resultados que se presentan en este documento. Daniel Peña agradece el apoyo parcial de la Agencia Nacional de Evaluación de la Calidad y Acreditación con referencia PID2019-109196GB-I00.

1. INTRODUCCIÓN

El análisis de factores comunes (dinámicos) en series temporales multivariadas ha recibido mucha atención en los últimos años. Este método puede considerarse como una forma de reducir la dimensión de un vector de variables y, cuando el número de series analizado es grande, es una herramienta muy utilizada para el manejo de *big data*. La idea central detrás del modelo factorial es explicar los comovimientos entre las variables observadas por un número muy pequeño de factores subyacentes latentes. Este ha sido un objetivo importante de los macroeconomistas en décadas recientes, como lo podemos consultar en los trabajos de Stock y Watson (2016, 2017) y Diebold (2003), entre otros.

Para series temporales multivariantes se han considerado tres tipos de modelos factoriales: con factores comunes dinámicos exactos (FCDE), factores comunes dinámicos aproximados (FCDA) y factores comunes dinámicos generalizados (FCDG). Entre muchos otros, los trabajos de Peña y Box (1987), Peña y Poncela (2006) y Lam y Yao (2012) están en la órbita de los FCDE; Bai (2004), Bai y Ng (2002, 2004) y Ahn y Horenstein (2013) han trabajado en el campo de los FCDA y Forni *et al.* (2000) han sido pioneros de los FCDG. En este último caso, el análisis se ha realizado bajo la premisa de que el número de variables es muy grande. En este libro y en Peña, Poncela y Ruiz (2021) pueden encontrase varios trabajos panorámicos sobre las aplicaciones de los modelos factoriales a la predicción y al análisis económico.

Los métodos citados han supuesto, explícita o tácitamente, que las variables analizadas no son estacionales. Sin embargo, y de manera reciente, el problema de analizar factores comunes en presencia de estacionalidad ha empezado a ser considerado, ya que la desestacionalización de series temporales puede inducir comportamientos ficticios en las variables bajo estudio, como se puede consultar en la literatura (ver, por ejemplo, Bell y Hillmer, 1984). Es frecuente que las variables macroeconómicas tengan estacionalidad, y un análisis de factores comunes dinámicos deberá tener en cuenta explícitamente esta característica. Modelos factoriales con estacionalidad han sido propuestos, entre otros, por Melo *et al.* (2001), Busetti (2006), Alonso *et al.* (2011), García-Martos, Rodríguez y Sánchez (2011), Camacho, Lovcha y Perez-Quiroz (2015) y Nieto, Peña y Saboyá (2016). El objetivo de este capítulo es explicar este último enfoque y destacar sus potenciales ventajas en el análisis de datos masivos de series económicas.

El resto de este capítulo está organizado como sigue: en la sección segunda presentaremos una descripción del concepto de estacionalidad y su modelización en series económicas, tanto en series univariantes como multivariantes. La sección tercera introduce el modelo factorial dinámico estacional propuesto por Nieto, Peña y Saboyá (2016) y sus propiedades principales. La sección cuarta aplica el análisis factorial estacional a datos del dinero en efectivo en circulación en 15 países latinoamericanos. Las conclusiones de este trabajo se presentan en la sección quinta.

2. ESTACIONALIDAD EN SERIES TEMPORALES

2.1. Estacionalidad univariante

Históricamente, el concepto de estacionalidad surge al tener en cuenta el efecto de las estaciones climatológicas sobre la variabilidad de una serie temporal. Jevons (1862), citado por Klein (1997), definió este efecto al analizar series temporales comerciales (precios, ventas, inventarios, etc.), comprobando que las observaciones mensuales o trimestrales tienden a tener un valor medio diferente en distintos meses o trimestres. Se induce entonces un comportamiento periódico a través del tiempo, con periodo $s = 12$ si los datos son mensuales, o $s = 4$ si son trimestrales. La práctica inicial fue desestacionalizar este tipo de variables, esto es, estimar el ciclo estacional y restarlo a la serie, pero para ello hay que suponer un cierto modelo estacional que puede no ser adecuado para la serie, lo que puede introducir ciclos o comportamientos ficticios en la variable desestacionalizada; veáse por ejemplo Bell y Hillmer (1984).

La estacionalidad de una serie temporal se ha modelado con varios tipos de modelos estadísticos. Inicialmente el ciclo estacional, generalmente ligado a la temperatura, se representaba por medio de una función determinista periódica, como por ejemplo, las funciones seno y coseno. Es claro que esta formulación de la estacionalidad es muy restrictiva: una serie mensual de ventas, puede tener estacionalidad por unas ventas promedio más altas en diciembre, efecto que se representará mal con una función sinusoidal. Una alternativa más flexible es estimar un modelo con s variables ficticias, una para cada periodo (y omitir la constante en el modelo), o introducir $s - 1$ variables ficticias más la constante. Por ejemplo, con datos mensuales los coeficientes de las 12 variables ficticias estiman la media de cada mes, mientras que con 11 y la constante los coeficientes son el incremento (o decremento si es negativo) de cada mes respecto a la media del que tomamos como referencia, cuya media se estima con la constante. En lugar de suponer una estacionalidad determinista, un modelo más general es permitir que los efectos evolucionen con el tiempo. Por ejemplo, en los modelos estructurales o de componentes no observables (Harvey, 1989), la serie se representa como:

$$y_t = T_t + S_{t,s} + I_t, \qquad [1]$$

donde T_t representa la tendencia de largo plazo de la variable, $S_{t,s}$ es la estacionalidad que tiene una dinámica con periodo s y seguirá un proceso definido, en general no determinista, e I_t denota la componente irregular. El modelo que se establezca para la evolución de $S_{t,s}$ determina el comportamiento estacional.

Algunos autores consideran que la estacionalidad aleatoria puede tener efectos permanentes o transitorios. En el primer caso la estacionalidad permanece en la serie, por ejemplo, si $S_t^{(i)} = S_{t-s}^{(i)} + \eta_t$ para $i = 1, ..., s$, donde η_t es un proceso estacionario, la estacionalidad es permanente y la serie no estacionaria. Sin embargo, si $S_t^{(i)} = \phi S_{t-s}^{(i)} + \eta_t$, con $\phi < 1$, la estacionalidad prevista desaparece a largo plazo y el proceso es estacionario. Por ejemplo,

Hylleberg *et al.* (1990) definen la estacionalidad como picos en el espectro de la serie en las frecuencias estacionales, que pueden estar causados por efectos permanentes o transitorios. Otros autores, como Box y Jenkins (1970), suponen que la estacionalidad es una característica permanente en la serie. Entonces, para conseguir una serie estacionaria es necesario aplicar diferencias estacionales, $\nabla_s = (1\text{-}B^s)$ donde B es el operador de retardo y $B^s y_t = y_{t\text{-}s}$. Esto es debido a que con estacionalidad estable los valores medios de la serie en distintos puntos del ciclo estacional serán diferentes.

Una forma general de modelar la estacionalidad es redefinir los valores temporales como $t = i + js$, donde $i = 1, ..., s$ son los instantes del ciclo estacional y $j = 0, ..., L\text{-}1$ representa los ciclos completos de periodo s, y supondremos para simplificar que tenemos $N = Ls$ datos. Entonces, $y_t - y_{i+js}$. Para cada valor de i podemos construir una serie temporal que relaciona los valores de los mismos periodos a lo largo del ciclo estacional, por ejemplo los meses dentro del año. Supongamos que estas series son estacionarias. Llamando c_i a la media de cada una de las s series, su estructura será del tipo:

$$y_{i+js} = c_i + \Phi_i^{-1}(B^s)\Theta_i(B^s)u_{i+js} \qquad [2]$$

donde en esta ecuación i es fijo, representa el periodo, $u_{i+js} = u_t$ es un proceso estacionario de media cero. Los ruidos de estas s ecuaciones pueden unirse en un proceso único de ruido que estará incorrelado para los retardos múltiplos de s pero puede tener dependencia a retardos 1,2,.... Esta serie de ruido común, u_t tal que $u_{i+js} = u_t$, seguirá un cierto proceso ARMA del tipo:

$$u_t = \phi^{-1}(B)\theta(B)a_t, \qquad [3]$$

donde a_t es ruido blanco. Las s+1 ecuaciones [2] y [3] representan la modelización del proceso estacional estacionario. En el caso particular de que los modelos [2] sean todos idénticos y con los mismos parámetros, es decir todos los periodos estacionales tienen la misma media y la misma dinámica, sustituyendo [3] en [2] tenemos un modelo para la serie completa y_i dado por:

$$\Phi(B^s)\phi(B)(y_t - c) = \theta(B)\Theta(B^s)a_t$$

Este es un modelo ARMA donde los operadores AR y MA factorizan en dos bloques de retardos B y B^s, pero las medias de las observaciones, tanto marginales como condicionadas al periodo, son las mismas. Por ejemplo, si la serie es mensual este modelo implica que todos los meses tienen la misma media y estructura de dependencia de los meses pasados. En este sentido son intercambiables. La estacionalidad es solo una dependencia transitoria que hace que si un mes el valor es especialmente alto o bajo es previsible que este efecto se extienda a corto plazo a los años futuros.

Supongamos ahora que las series y_{i+js} son no estacionarias. Por ejemplo siguen un paseo aleatorio $y_{i+js} = y_{i+(j\text{-}1)s} + u_{i+js}$. Entonces, suponiendo que $\nabla_s y_{i+js}$ es estacionario para $i = 1, ..., s$, si las s series siguen el mismo modelo estacionario utilizando el mismo razonamiento anterior podemos escribir el modelo común para todas las series:

$$\Phi(B^s)\phi(B)\nabla_s y_t = \theta(B)\Theta(B^s)a_t. \qquad [4]$$

Observemos que la diferencia estacional es imprescindible para permitir que la media de distintos periodos estacionales no sea la misma y exista un efecto estacional permanente. El modelo [4] puede también incluir diferencias regulares $\nabla = (1\text{-}B)$.

La recogida de datos masivos con alta frecuencia hace que muchas series actuales tengan varios tipos de estacionalidad. Por ejemplo, si medimos el consumo de electricidad cada hora tendremos una estacionalidad horaria dentro de cada día, con $s_1 = 24$ datos, diaria dentro de cada semana, con $s_2 = 7$ días x 24 horas=168 datos, y mensual dentro del año, con $s_3 = 30$ días x 24 horas=888 datos. Este modelo, suponiendo que cada tipo de dependencia es el mismo en todos los periodos de cada ciclo estacional, puede modelarse como:

$$\Phi_1(B^{s_1})\Phi_2(B^{s_2})\Phi_3(B^{s_3})\phi(B)\nabla_{s_1}\nabla_{s_2}\nabla_{s_3}\nabla y_t = \Theta_1(B^{s_1})\Theta_2(B^{s_2})\Theta_3(B^{s_3})\theta(B)a_t.$$

Otra alternativa es utilizar un modelo estructural generalizando [1], por ejemplo para tres estacionalidades, del tipo:

$$y_t = T_t + \sum_{j=1}^{3} S_{t,s_j} + I_t$$

La forma más efectiva de modelar muchas estacionalidades es un tema actual de investigación. Veáse por ejemplo De Livera, Hyndman y Snyder (2011).

2.2. Estacionalidad multivariante

En contraste con los muchos trabajos sobre la estacionalidad en series univariantes, esta característica ha sido poco estudiada con series multivariantes (véase por ejemplo Reinsel, 1997; Lütkepohl, 2013 o Tsay, 2014). Dado un conjunto de m series con estacionalidad del mismo periodo, s, el correspondiente vector de series y_t de dimensión m con la notación de la sección anterior, puede escribirse de forma estructural, suponiendo el caso mas interesante de no estacionariedad, como:

$$\mathbf{y}_{i+js} = \mathbf{T}_{i+js} + \mathbf{S}_i^{(j)} + \mathbf{n}_{i+js}, \qquad [5]$$

o en forma reducida VARMA como:

$$\Phi(B^s)\phi(B)\nabla\nabla_s \mathrm{y}_t = \theta\,(B)\Theta\,(B^s)\mathrm{a}_t \qquad [6]$$

que generaliza [4] con operadores matriciales. Esta representación tiene sin embargo dos problemas. El primero es que ahora el orden de los operadores es importante porque las matrices $\Phi(B^s)$ y $\phi(B)$ no conmutan en general con lo que $\Phi(B^s)\phi(B) \neq \phi(B)\Phi(B^s)$. En segundo lugar, es frecuente que no todas las m series necesiten diferencias regulares o estacionales. Es bien conocido, véase por ejemplo Hylleberg (1990) y Busetti (2006), que un vector de series no estacionarias está cointegrado si existe al menos una combinación lineal $\alpha'\mathrm{y}_t$ estacionaria.

En estos casos no conviene diferenciar el vector de series porque introducimos términos de media móvil no invertible. Lo mismo ocurre con la diferenciación estacional, ya que las series pueden compartir estacionalidades comunes.

Para ilustrar este último aspecto consideremos la formulación [5]. Si el vector de series está cointegrado y $\alpha' y_{i+js}$ es estacionario esto implica que $\alpha' T_{i+js}$ y $\alpha' S_t^{(j)}$ deben serlo también. Por ejemplo, supongamos que T_{i+js} es una tendencia determinista lineal común, $T_{i+js} = a1 + b1t$, donde $1' = (1, ...,1)$, y $S_i^{(j)} = 0$. Entonces, los m-1 vectores α_j que verifican $\alpha_j' T_{i+js} = 0$ definen relaciones de cointegración. Lo mismo ocurre si $T_{i+js} = pr_{i+js}$, donde r_t es un paseo aleatorio univariante. Los m-1 vectores ortogonales al vector p definen relaciones de cointegración. Análogamente, si los procesos $S_i^{(j)}$ no tienen media cero, y las series tienen estacionalidad permanente, pero existe un vector $\delta' S_i^{(j)} = 0$, que conduce a la serie univariante $\delta' y_{i+js}$ no estacional diremos que existe cointegración estacional de periodo s. Por ejemplo, supongamos que existen solo dos tipos de estacionalidad en las series de manera que para las $m/2$ primeras (suponemos m par) hay una estacionalidad determinista con coeficientes $c_i^{(1)}$ y para las $m/2$ restantes coeficientes $c_i^{(2)}$. Definiendo los vectores $u_1 = (1, ...,1, 0...,0)$ y $u_2 = (0...,0, 1, ...,1,)$ el modelo [5] puede escribirse como:

$$y_{i+js} = T_{i+js} + u_1 c_i^{(1)} + u_2 c_i^{(2)} + n_{i+js}$$

y los m-2 vectores δ_j que verifican $\delta_j' u_1 = \delta_j' u_2 = 0$ conducen a series univariantes no estacionales. Tenemos m-2 relaciones de cointegracion estacional o dos factores comunes estacionales. Lo mismo ocurre con estacionalidades no deterministas. Por ejemplo, si $S_i^{(j)} = b c_i^{(j)} + v_i^{(j)}$, donde $( )$ es un componente estacional estocástico común y $v_i^{(j)}$ un vector de ruido blanco, las combinaciones ortogonales a b definen m-1 relaciones de cointegración estacional.

Una forma alternativa de modelar la cointegración es con modelos factoriales (veáse Escribano y Peña, 1994; Peña y Poncela, 2006; y Peña y Tsay, 2021). Con muchas series es más interesante estudiar los factores comunes, que suelen ser pocos, que las relaciones de cointegración, que pueden ser muchas. Esta es una ventaja clara para modelar la estacionalidad multivariante de muchas series con un modelo factorial estacional, como el que presentamos a continuación.

3. UN MODELO FACTORIAL DINÁMICO ESTACIONAL

Muchas series temporales relacionadas suelen compartir varios grupos de estacionalidad, o factores estacionales comunes. Siguiendo el enfoque de Peña y Poncela (2006) de analizar factores de tendencia común, Alonso *et al.* (2011), García-Martos, Rodríguez y Sánchez (2011) y Nieto, Peña y Saboyá (2016) han estudiado modelos factoriales estacionales. El enfoque de estos últimos autores parte del modelo factorial de Peña y Box (1987):

$$y_t = Pf_t + e_t, \qquad [7]$$

donde y_t es el vector de variables observables de dimensión m, f_t un vector de factores no observables (o latentes) de dimensión r, con $r < m$, P es una matriz de dimensión $m \times r$,

llamada la matriz de pesos del vector de factores f_t y $\{e_t\}$ es una sucesión de vectores idéntica e independientemente distribuidos, con media el vector 0, matriz de varianzas y covarianzas Σ_e y distribución multinormal cada uno. Conceptualmente, el modelo indica que cada variable depende de una combinación lineal de los factores f_{1t}, ..., f_{rt} y de un ruido intrínsico a esa variable. Los pesos o ponderaciones de los factores en una variable, son las entradas de la fila de P correspondiente a ella. Así, por ejemplo, $y_{1t} = p_{11}f_{1t} + \cdots + p_{1r}f_{rt} + e_{1t}$.

Con el fin de extraer los ciclos estacionales comunes del conjunto de variables observables, y además otras características comunes a las variables, como tendencias de largo plazo y movimientos de corto plazo, se supone $f_t = \left(f'_{1t}, f'_{2t}, f'_{3t}\right)'$, donde f_{1t} es un vector de dimensión r_1 y representa las tendencias comunes (movimientos de largo plazo), f_{2t} es de dimensión r_2 y especifica la estacionalidad común (ciclos) de las variables observadas y f_{3t} de dimensión r_3 representa los factores comunes estacionarios (movimientos comunes de corto plazo). Nótese que $r_1+r_2+r_3 = r$ y con $P = [P_1, P_2, P_3]$, donde P_i es una submatriz de dimensión $m \times r_i$, $i = 1,2,3$, se obtiene que,

$$y_t = P_1 f_{1t} + P_2 f_{2t} + P_3 f_{3t} + e_t,$$

la cual es una expresión análoga al modelo de componentes no observables multivariado (Harvey, 1989).

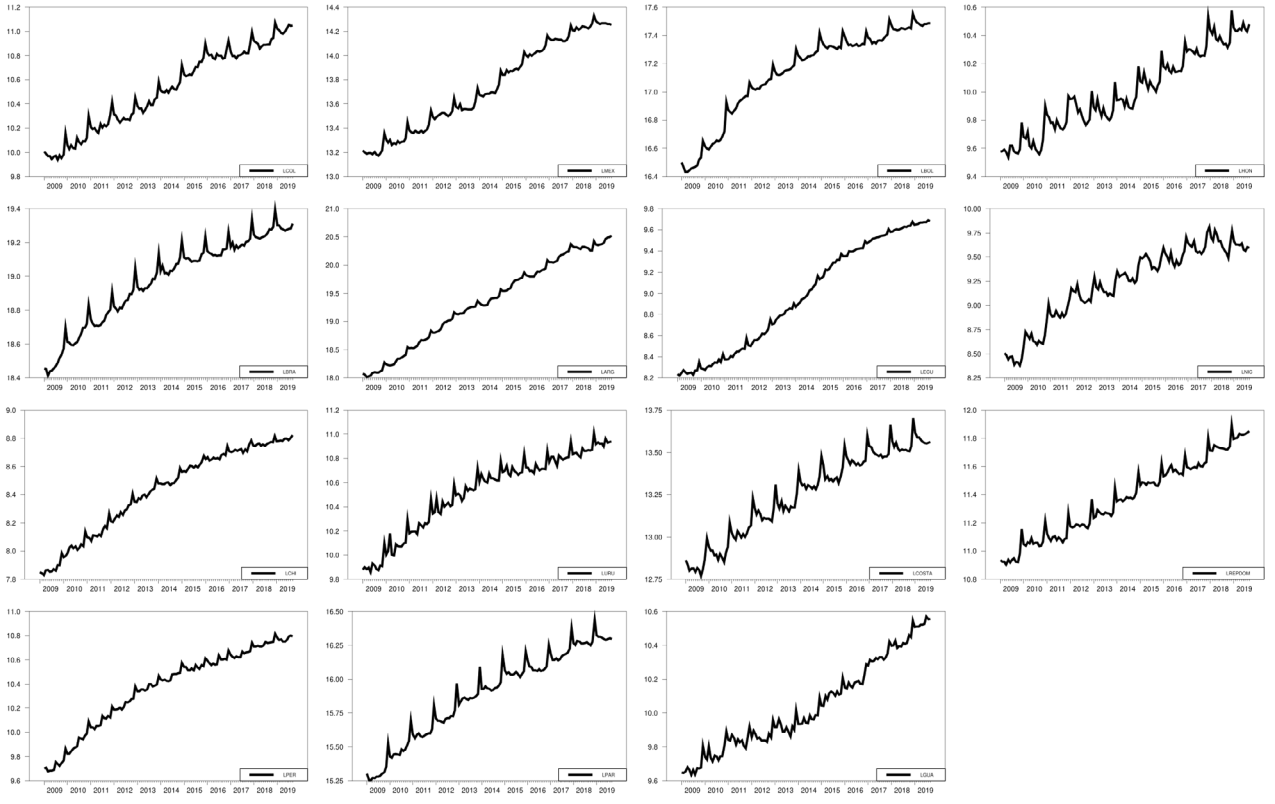
3.1. Un ejemplo de estacionalidad común

Antes de establecer las propiedades principales del modelo factorial estacional de Nieto, Peña y Saboyá (2016), consideramos el siguiente ejemplo empírico, el cual nos permite motivar la aplicación de esa metodología. Las variables de interés corresponden al dinero en efectivo en circulación, de 15 países de América Latina. Estos son: Colombia (1), Brasil (2), Chile (3), Perú (4), México (5), Argentina (6), Uruguay (7), Paraguay (8), Bolivia (9), Ecuador (10), Costa Rica (11), Guatemala (12), Honduras (13), Nicaragua (14) y República Dominicana (15). La fuente de los datos es Bloomberg y el período muestral considerado es enero, 2009-septiembre, 2019, para un total de 129 observaciones por país. En la figura 1 presentamos las gráficas de las series temporales en logaritmos, donde comenzando por Colombia, observamos de arriba hacia abajo columna por columna, en el orden establecido antes para los países. Como podemos observar, hay una tendencia creciente de tipo lineal en algunos países y de tipo curvilíneo en otros, notando que la tendencia de Ecuador difiere un poco de las demás (cambio de concavidad). También, es muy clara la presencia de estacionalidad en cada una de ellas, aunque la estacionalidad de las variables de Argentina y Ecuador no es tan marcada como en los otros países.

Es bien sabido que la inflación induce la tendencia creciente de este agregado monetario, mientras que la estacionalidad tiende a ser muy típica en épocas del año donde la demanda de liquidez de los agentes de la economía, tiende a elevarse por motivo de transacciones de navidad, verano, vacaciones colectivas, etcétera.

Figura 1.

Series temporales en logaritmos del dinero en circulación, en algunos paises latinoamericanos



Fuente: Elaboracion propia.

A simple vista, se intuye la presencia de, al menos, una tendencia común y un ciclo estacional común, los cuales inducen de alguna manera la tendencia y la estacionalidad de cada serie temporal. La pregunta es: ¿cuántas tendencias comunes y cuántos ciclos estacionales comunes están detrás del origen de los datos? ¿Hay movimientos de corto plazo comunes? ¿Es el número total de factores comunes mucho menor que el número de variables observadas?

Con el propósito de responder las preguntas anteriores, vamos a describir las propiedades del modelo factorial estacional de Nieto, Peña y Saboyá (2016). La idea básica es extraer la información presente en los datos, acerca del número y el tipo de los factores comunes, a partir de la estructura de asociación estadística conjunta de las variables observables. Es bien conocido que la estructura de autocorrelación conjunta de los factores comunes determina aquella de las variables observables (ver, por ejemplo, Peña y Box, 1987).

3.2. Propiedades del modelo factorial estacional

Sea $N = Ls$ el tamaño de muestra, definimos las matrices de autocovarianza muestral generalizadas (CVMG) $C(k, N)$ como:

$$C(k,N) = \frac{s^{2d}}{N^{2d}} \sum_{t=k+1}^{N} (\mathbf{y}_{t-k} - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})' \, , k = 0,1,2,..., \qquad [8]$$

donde $\bar{\mathbf{y}} = \frac{1}{T}\Sigma_{t=1}^{N}\mathbf{y}_t$ y d es el orden de integración de las tendencias estocásticas comunes. Nótese que $L = N/s$ es el número de ciclos estacionales en la muestra que supondremos un número entero y N múltiplo de s. Esta sucesión de matrices cuantifica la estructura de autocovarianzas conjunta de las variables. Vamos a ver que los valores y vectores propios de estas matrices son informativos sobre la presencia de factores comunes, regulares o estacionales en las series. Cuando N es muy grande y se cumplen ciertas condiciones sobre el modelo factorial estacional se comprueba que los m valores propios de las matrices $C(k,N)$, ordenados de mayor a menor, poseen la siguiente propiedad: (i) si el retardo k es estacional, es decir tiene la forma js, para algún número natural j, entonces los últimos $m - (r_1 + r_2)$ valores propios son muy cercanos a cero. (ii) Si k no es estacional los últimos $m - r_1$ valores propios son también muy cercanos a cero.

Para examinar esta característica del modelo en la práctica, estimamos las matrices $C(k,N)$ con la serie temporal multivariante, hasta un retardo K previamente establecido, y computamos los valores propios de cada una de ellas. Obtenemos las m sucesiones de valores propios, indexadas por el retardo $k = 0, 1, ..., K$, y observamos el comportamiento de las sucesiones tanto en los retardos estacionales como en los no estacionales. Determinamos entonces un valor para r_1 a partir de los retardos no estacionales y un valor para r_2 a partir de los retardos estacionales, siguiendo las propiedades anotadas antes.

Con base en las propiedades y características del modelo [1], citadas anteriormente, la propuesta metodológica para especificar el número y tipo de factores comunes, en el vector de

variables observables, sigue los siguientes pasos: 1) graficar las sucesiones de valores propios de las matrices CVMG para identificar el número de tendencias comunes r_1 y el número r_2 de ciclos estacionales comunes; 2) por medio de algún contraste o prueba estadística, determinar el número total r de factores comunes. Algunas alternativas son la prueba de Peña y Poncela (2006), extendida por Nieto, Peña y Saboyá (2016) y Bolívar, Nieto y Peña (2020), y los contrastes de Ahn y Horenstein (2013), Lam y Yao (2012) y Caro y Peña (2020); 3) tomar $r_3 = r - (r_1 + r_2)$.

Una vez identificados el número y tipo de factores comunes, esto es, r_1, r_2 y r_3, se debe especificar un modelo estadístico ARIMA/SARIMA para los factores comunes. Al hacerlo, los parámetros poblacionales fijos del modelo factorial estacional pueden ser estimados, usando el método de máxima verosimilitud y los factores comunes son estimados usando el denominado suavizador de punto fijo (Harvey, 1989). La idea es transformar el modelo factorial en un modelo de estados (Harvey, 1989; Gómez y Maravall, 1994). En muchos paquetes estadísticos se encuentran códigos para la estimación de un modelo de estados. Sin embargo, cuando la dimensión del vector de variables observables es muy grande (big data), pueden ocurrir problemas computacionales complejos. En ese caso, recomendamos utilizar el método de reducción de la dimensión de un modelo de estados de Jungbaker y Koopman (2015), veáse también Poncela, Ruiz y Miranda (2020).

3.3. Un ejemplo simulado

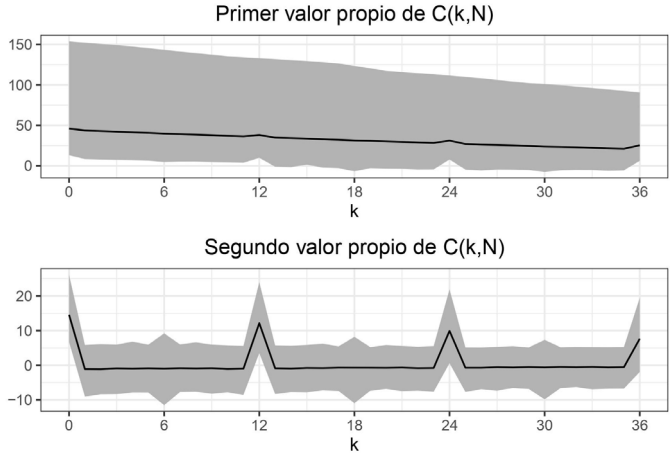
Con el fin de ilustrar el procedimiento propuesto antes, para especificar el número total de factores comunes, el número de tendencias comunes y el número de ciclos estacionales comunes, utilizaremos un modelo factorial estacional con $m = 20$, $r_1 = 1$, $r_2 = 1$ y $r = r_1 + r_2 = 2$. Las entradas de la matriz P son generadas aleatoriamente, a partir de una distribución uniforme sobre el intervalo [-1,1], fijando $p_{1,2} = 0$. Además, se garantiza que P sea ortonormal. La matriz de varianzas y covarianzas de la sucesión de vectores de ruido es $\Sigma_e = I_{20}$, la matriz identidad de orden 20. Para el primer factor, esto es, $f_{11,t}$, consideramos el modelo ARIMA(0,1,1) dado por $f_{11,t} = f_{11,t-1} + a_{11,t} + 0{,}4a_{11,t-1}$ y para el segundo, es decir, $f_{21,t}$, el modelo estacional puro SARIMA $(0,1,1)_{12}$ dado por $f_{21,t} = f_{21,t-12} + a_{21,t} + 0{,}6a_{21,t-12}$. Acá, $\{a_{11,t}\}$ y $\{a_{21,t}\}$ son, cada una, sucesiones de variables aleatorias idéntica e independientemente distribuidas, con media cero, varianza 1 y distribución normal. Este tipo de modelos se encuentra con mucha frecuencia en el análisis de series temporales económicas, ya que determinan el llamado proceso de suavizamiento exponencial simple, un método muy usado para obtener, de manera recurrente, pronósticos de una variable de interés (ver, por ejemplo, Peña, 2005).

El diseño del experimento de simulación es el siguiente: simulamos 1.000 series temporales multivariantes de longitud 300 cada una, luego, para cada serie, caculamos los valores propios de las matrices CVMG para retardos $k = 0,1, ..., 36$. Obtenemos así 1.000 primeros valores propios, 1.000 segundos valores propios, etc. Calculamos el promedio aritmético y los cuantiles 0.025 y 0.975 de cada muestra de valores propios, de cada retardo k, y construimos las 20 sucesiones de los valores promedio junto con intervalos del 95 % de confianza. En la

figura 2 observamos las sucesiones de los dos primeros valores propios con sus bandas de confianza resaltadas en color gris (que resultan al unir los cuantiles, tanto inferiores como superiores, de cada retardo). Observamos que la banda de confianza del primer valor propio (parte superior), prácticamente no contiene al cero, con lo que podemos validar que el primer valor propio es positivo en todo retardo *k*. En la parte inferior de la figura 2, observamos que la banda de confianza no contiene al cero en los retardos estacionales (0,12,24,36) mientras que sí lo incluye en los retardos no estacionales. En la figura 3, presentamos la gráfica del tercer valor propio y observamos que, excepto por el retardo $k = 0$, la banda de confianza incluye a cero en todos los demás retardos. Este comportamiento se repite en el resto de sucesiones de valores propios medios. De esta manera, concluimos que en los retardos estacionales hay

Figura 2.

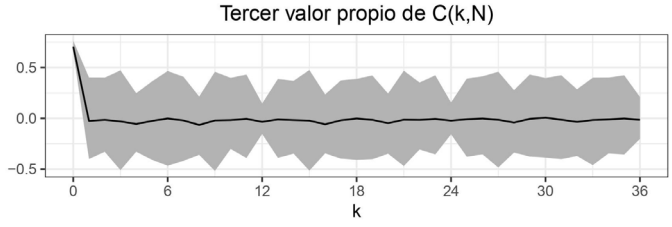
Sucesiones del primer valor propio promedio (arriba) y segundo valor propio promedio (abajo) de las matrices SGCV del ejemplo simulado, con bandas del 95 % de confianza



Fuente: Elaboración propia.

Figura 3.

Sucesión del tercer valor propio medio de las matrices SGCV del ejemplo simulado



Fuente: Elaboración propia.

dos valores propios positivos y en los no estacionales un valor propio positivo, como lo indica el modelo simulado.

4. ANÁLISIS DEL EFECTIVO EN CIRCULACIÓN EN LATINOAMÉRICA

En esta sección aplicamos la metodología del análisis factorial en presencia de estacionalidad, en el grupo de variables determinadas por el efectivo en circulación de 15 paises latinoamericanos, el cual fue descrito en la sección anterior. Como es sabido en el contexto económico, el comportamiento de este tipo de variables le permite a la autoridad monetaria, evaluar la liquidez agregada de la economía.

Un análisis exploratorio preliminar de los datos, en donde se aplican pruebas de raíz unitaria de Dickey-Fuller alrededor de tendencia lineal, sugiere la presencia de una tendencia estocástica en la mayoría de las series temporales. Adicionalmente, los correlogramas de la primera diferencia finita de cada serie temporal, exhiben autocorrelaciones significativas al 5 % en los retardos 12, 24 y 36, con un lento decrecimiento. Este hecho indica la necesidad de una diferencia estacional en cada serie temporal. En consecuencia, es plausible considerar que exista, al menos, una tendencia estocástica común y un ciclo estacional común.

Las etapas del análisis son las siguientes:

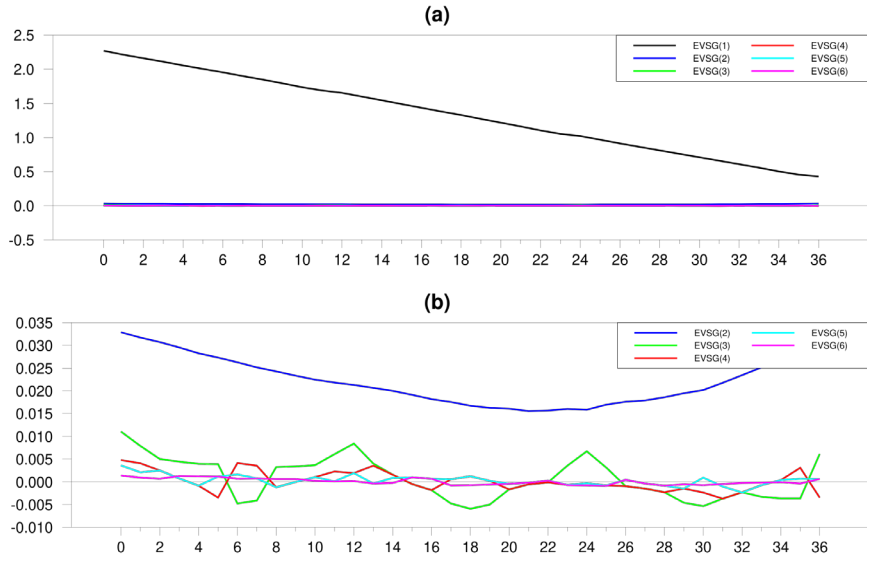
1. Identificación del número y del tipo de factores

Se computan las 15 sucesiones de valores propios de las matrices de covarianza muestral generalizada, para retardos $k = 0, 1, ..., 36$. Las seis primeras sucesiones se presentan en la figura 4(a) y se observa que la primera sucesión es muy preponderante, en el sentido que sus valores son mucho mayores que aquellos de las restantes sucesiones, en todos los retardos considerados. Esta sucesión señala la presencia de un factor no estacionario y no estacional muy fuerte. En la figura 4(b) presentamos las sucesiones del 2º al 6º valor propio. Podemos observar que la segunda sucesión indica también la presencia de otro factor no estacionario y no estacional, menos fuerte que el anterior. La tercera sucesión señala un factor estacional, pues sus valores en los retardos estacionales son relativamente grandes (son los terceros en magnitud) y en los no estacionales son relativamente pequeños, en comparación con las dos primeras sucesiones. El resto de sucesiones señalan valores muy pequeños en todos los retardos. En consecuencia, podemos especificar que $r_1 = 2$ y $r_2 = 1$.

Con el propósito de identificar el número total de factores r, intentamos utilizar la prueba de Peña y Poncela (2006). Sin embargo, como lo destacan Nieto, Peña y Saboyá (2016), dado que $N/m = 129/15 \approx 8{,}6 < 20$, no tenemos soporte metodológico para realizarla. De hecho, un intento de su aplicación nos arroja el valor $r = 11$, al nivel 5 %, lo cual, desde el punto de vista de reducción de dimensión, no es un valor razonable. En su lugar, podemos considerar métodos alternativos, como los de Ahn y Horenstein (2013), Lam y Yao (2012) y Caro y Peña (2020). Al hacerlo así, se identifica un solo factor común. Por la concepción y filosofía de esos métodos, podríamos decir que ellos solo identifican el factor no estacionario y no estacional fuerte, el que induce la tendencia creciente de las variables.

Figura 4.

(a) Primeras seis sucesiones de valores propios de las matrices de CVMG
(b) Cinco últimas sucesiones (EVSG(i): i-ésimo valor propio)



Fuente: Elaboración propia.

Con el fin de explorar por la presencia de factores comunes adicionales, una práctica usual es corregir los datos originales por estimaciones iniciales de los factores detectados. Para obtener estimaciones iniciales de los factores, podemos proceder de la siguiente manera: con la filosofía del análisis de componentes principales y soportado también por los resultados de Nieto, Peña y Saboyá (2016), utilizamos los tres primeros vectores propios de $C(0,N)$, esto es, los asociados a los tres primeros valores propios, los cuales denotamos $\hat{P}^0_{\cdot 1}$, $\hat{P}^0_{\cdot 2}$ y $\hat{P}^0_{\cdot 3}$, respectivamente. Luego, tomamos $f^0_{11,t} = (\hat{P}^0_{\cdot 1})'\mathbf{y}_t, f^0_{12,t} = (\hat{P}^0_{\cdot 2})'\mathbf{y}_t$ y $f^0_{21,t} = (\hat{P}^0_{\cdot 3})'\mathbf{y}_t$, como las estimaciones iniciales de los factores. Sea $\hat{\mathbf{f}}^{(0)}_t = (f^0_{11,t}, f^0_{12,t}, f^0_{21,t})'$. En la figura 5 podemos observar dichas estimaciones iniciales. Es importante señalar en este punto que los dos primeros vectores propios de las matrices $C(0, N)$, $C(1, N)$ y $C(12, N)$ son muy parecidos, luego el primer par de vectores propios de $C(1, N)$ y $C(12, N)$ pueden ser utilizados también para computar los dos primeros factores comunes preliminares.

Ahora, la corrección por el primer factor común la realizamos a través de la expresión $\hat{\mathbf{y}}^{(1)}_t = \mathbf{y}_t - \hat{P}^0_{\cdot 1} f^0_{11,t}$ y aplicamos entonces los contrastes a la serie temporal $\hat{\mathbf{y}}^{(1)}_t$. Los resultados de este proceso a dos etapas se presentan en la tabla 1, donde las celdas de la tabla indican el número de factores identificado en cada etapa. Como podemos observar, dichos contrastes detectan globalmente un total de dos factores comunes, los cuales corresponderían a los dos factores no estacionarios y no estacionales identificados con los valores propios de las matrices CVMG, pero no detectan el factor común estacional.

La decisión final acerca del número r la tomaremos de la siguiente forma: sea $\hat{\mathbf{y}}_t = \hat{P}^{(0)}\hat{\mathbf{f}}_t^{(0)}$, donde $\hat{P}^{(0)}$ es la estimación inicial de la matriz P que se obtiene al utilizar como sus columnas, los vectores propios de $C(0, N)$ descritos antes. La serie temporal $\hat{\mathbf{y}}_t$ representa los datos ajustados por los tres factores identificados. Al repetir el paso 1 de la metodología propuesta en este capítulo, las sucesiones de valores propios de las matrices CVMG no señalan factores comunes adicionales. Combinando este resultado con aquellos presentados en la tabla 1, podemos concluir que una adecuada especificación del número total de factores comunes es $r = 3$.

Tabla 1.

AH.GR y AH.ER : Ahn y Horenstein (con las dos alternativas de razón de valores propios). LY(K): Lam y Yao con K retardos. CP(K): Caro y Peña con K retardos

Contraste	*Datos originales*	*Ajuste por el factor 1*
AH.GR	1	1
AH.ER	1	1
LY(9)	1	1
LY(12)	1	1
CP(9)	1	1
CP(12)	1	1

Fuente: Elaboración propia.

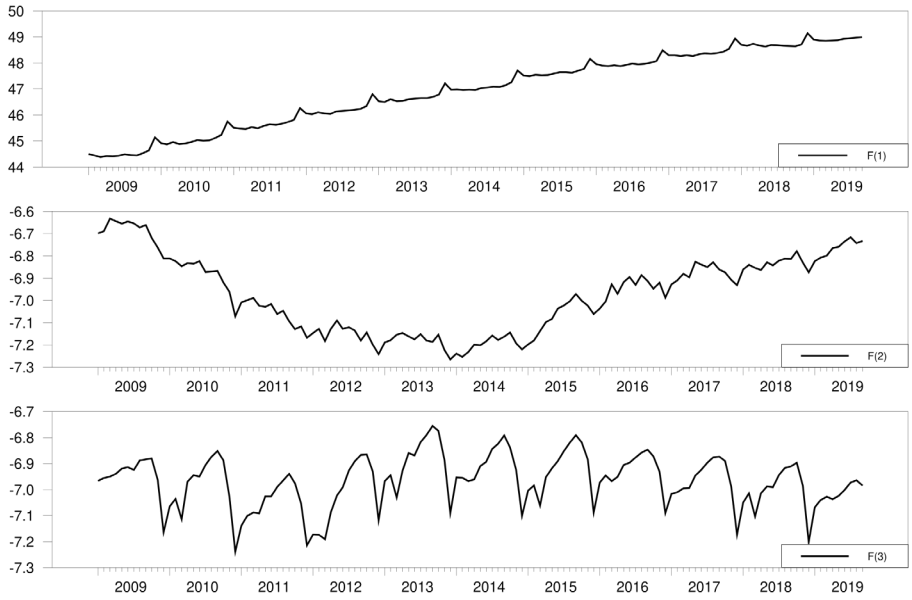
2. Identificación de los modelos para los factores comunes

Antes de realizar la estimación de los parámetros fijos del modelo, debemos identificar modelos ARIMA/SARIMA para los tres factores comunes. Con este fin, utilizamos las estimaciones iniciales de los factores como series temporales de los mismos e identificamos sus modelos vía, por ejemplo, el método clásico de Box y Jenkins (1976). Procedimientos automáticos de identificación también pueden ser usados para ese propósito como, por ejemplo, el paquete estadístico TSW (Caporello y Maravall, 2008).

Al hacerlo así, encontramos que los siguientes modelos son adecuados para especificar la dinámica de los factores $f_{11,t}$, $f_{12,t}$, y $f_{21,t}$: $(1-\phi B)\,(1-B)^2 f_{11,t} = a_{11,t}$, $(1-B)^2 f_{12,t} = a_{12,t}$ y $(1-B^{12})\, f_{21,t} = (1+\Theta B^{12})a_{21,t}$. Nótese que (i) el orden de diferenciación ordinaria especificado es $d = 2$, que concuerda con la observación empírica de una tendencia estocástica alrededor de tendencias de tipo lineal en las variables observables, (ii) el segundo factor común propuesto es una caminata aleatoria integrada, un modelo muy frecuente en análisis de series temporales económicas y (iii) el modelo estacional puro para el tercer factor representa un suavizamiento exponencial simple para las subseries mensuales.

Figura 5.

Factores comunes iniciales del efectivo en circulación



Fuente: Elaboración propia.

3. Estimación del modelo factorial

Para la estimación de los parámetros fijos del modelo factorial propuesto, utilizamos el método de máxima verosimilitud. Por razones de identificabilidad del modelo, restringimos a la matriz P de tal forma que P'P = I_3. En la tabla 2 presentamos las estimaciones de las componentes de P, redondeadas a dos cifras decimales, y de las varianzas de los ruidos intrínsicos ($\times 10^4$). Adicionalmente, ϕ = -0,57 y Θ = -0,06. En la figura 6 presentamos la gráfica de los pesos de cada factor en las diferentes variables, para una interpretación más cualitativa de los diferentes factores.

Tabla 2.

$\hat{P}_{\cdot i}$ es la estimación final de la columna *i* de *P*, i = 1, 2, 3; $\hat{\sigma}_i^2$ estimación de la varianza intrínsica del país *i*, 1 = 1, ...,15

País	$\hat{P}_{\cdot 1}$	$\hat{P}_{\cdot 2}$	$\hat{P}_{\cdot 3}$	$10^4\,\hat{\sigma}_i^2$
Colombia	0.09	0.05	-0.03	46.9
Brasil	0.07	-0.06	-0.01	49.6

Tabla 2. *(continuación)*

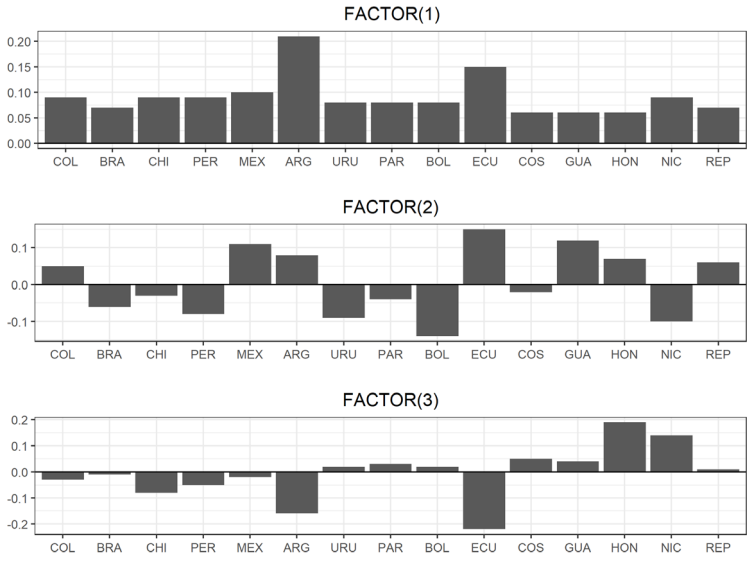
$\hat{P}_{.i}$ es la estimación final de la columna *i* de *P*, i = 1, 2, 3; $\hat{\sigma}_i^2$ estimación de la varianza intrínsica del país *i*, 1 = 1, ...,15

País	$\hat{P}_{.1}$	$\hat{P}_{.2}$	$\hat{P}_{.3}$	$10^4\,\hat{\sigma}_i^2$
Chile	0.09	-0.03	-0.08	14.6
Perú	0.09	-0.08	-0.05	16.3
México	0.10	0.11	-0.02	5.8
Argentina	0.21	0.08	-0.16	16.5
Uruguay	0.08	-0.09	0.02	68.5
Paraguay	0.08	-0.04	0.03	39.6
Bolivia	0.08	-0.14	0.02	14.0
Ecuador	0.15	0.15	-0.22	7.3
Costa Rica	0.06	-0.02	0.05	71.4
Guatemala	0.06	0.12	0.04	71.8
Honduras	0.06	0.07	0.19	9.0
Nicaragua	0.09	-0.10	0.14	133.3
R. Dominicana	0.07	0.06	0.01	58.3

Fuente: Elaboración propia.

Figura 6.

Pesos de cada factor en las 15 variables latinoamericanas



Fuente: Elaboración propia.

Como observamos en la tabla 2, 1ª columna de $\hat{P}$ (o en la parte superior de la figura 6), los pesos del primer factor son todos positivos, indicando que este factor es un promedio ponderado de todas las variables. Además, Argentina tiene un peso relativamente grande en el factor 1, seguido por Ecuador. Estas dos variables indicarían una alta influencia en esta tendencia común. Es conocido en el medio latinoamericano que la economía de Argentina ha padecido de una profunda estanflación en los últimos 10 años, con crecimientos anuales del PIB real de apenas 1,4 % (uno de los más bajos de Latinoamérica) y una alta inflación que se ha mantenido en promedio por encima del 21 %.

Sobre el segundo factor no estacionario y no estacional (segunda columna de $\hat{P}$), podríamos decir que es un contraste entre el grupo de Brasil, Chile, Perú, Uruguay, Paraguay, Bolivia y, curiosamente, Nicaragua y el resto de países. Bolivia recibe el mayor peso negativo y Ecuador, el mayor positivo.

Observando la tercera columna de $\hat{P}$ (o la parte inferior de la figura 6), podemos decir que el factor común estacional contrasta, esencialmente, al grupo de países centroamericanos (excepto México), con Chile, Argentina y Ecuador. Como podemos observar, Ecuador recibe el mayor peso negativo mientras que Honduras el mayor positivo.

De lo observado en la composición de los tres factores, podemos interpretar que Ecuador tiene un desempeño relativamente importante en este análisis factorial. Cabe notar que la economía de Ecuador está "dolarizada" desde hace 20 años, aproximadamente, y esto ha hecho que el Banco Central pierda la potestad de hacer emisión primaria y controlar la base monetaria de la economía.

Otro hecho interesante lo constituye la varianza del ruido intrínsico de Nicaragua, pues su valor es relativamente grande, en comparación con el resto de países. Así mismo, su peso en los tres factores es relativamente relevante. La economía de Nicaragua se ha contraído a tasas anuales del orden del -4 %, desde 2017 aproximadamente, el déficit fiscal se triplicó hasta niveles del 3 % del PIB, y la deuda pública se disparó hasta máximos históricos del 42 % del PIB. La independencia del Banco Central ha permitido actuar al emisor en función de estabilizar los medios de pago de la economía y el sistema financiero. No obstante, la crisis de los últimos años ha tenido efectos significativos sobre los ciclos naturales de varios de los principales agregados monetarios.

A la luz de los resultados anteriores, el efectivo en circulación en la muestra de 15 países latinoamericanos considerada, es explicado por tres variables comunes subyacentes (latentes). Una, explica la fuerte tendencia de tipo aproximadamente lineal de las series temporales. Otra, explica la presencia de una tendencia estocástica común de tipo caminata aleatoria integrada, que se superpone a la tendencia fuerte de las variables observadas. Y la otra, indica que las variables comparten una estacionalidad común. Al obtener esta reducción de dimensión, emergen características interesantes de las variables consideradas. Por ejemplo, y de alguna forma, el agregado monetario de Argentina, Ecuador y Nicaragua se separa del resto de variables consideradas. No obstante, el agregado monetario de Honduras y Bolivia también tiene cierto desempeño relevante en estas economías.

5. CONCLUSIONES

Con datos masivos, mucho más eficaz que investigar la cointegración estacional es construir un modelo que, además de los factores habituales integrados y estacionarios, pueda incluir factores comunes estacionales. El modelo propuesto por Nieto, Peña y Saboyá (2016) es un buen punto de partida con este objetivo.

Este modelo se ha aplicado para analizar el efectivo en circulación, como un agregado monetario, en una muestra de 15 países de América Latina. La metodología empleada permite ilustrar potencialmente, cómo podemos reducir un conjunto relativamente grande de variables observables, en un conjunto de variables latentes no observables de dimensión menor, en presencia de estacionalidad.

El método utilizado nos facilita un procedimiento numérico y gráfico para detectar los ciclos estacionales comunes, además de eventuales tendencias estocásticas comunes y comovimientos de corto plazo. Además, ha permitido separar los agregados monetarios de Argentina, Ecuador y Nicaragua, cuyas economías tienen características particulares en relación con el resto de economías consideradas.

Referencias

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

Alonso, A. M., Rodríguez, J., García-Martos, C. y Sánchez, M. J. (2011). Seasonal dynamic factor analysis and bootstrap inference: Application to electricity market forecasting. *Technometrics,* 53(2), pp. 137-151.

Bai, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics,* 122(1), pp. 137-183.

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

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

Bell, W. R. y Hillmer, S. C. (1984). Issues involved with the seasonal adjustment of economic time series. *Journal of Business and Economic Statistics,* 2(4), pp. 291-320.

Bolívar, S., Nieto, F. H. y Peña, D. (2021). On a new procedure for identifying a dynamic common factor model. *Colombian Journal of Statistics,* 44(1), pp. 1-21.

Box, G. E. P. y Jenkins, G. M. (1976). *Time Series Analysis Forecasting and Control.* California: Holden-Day.

Busetti, F. (2006). Tests of seasonal integration and cointegration in multivariate unobserved component models. *Journal of Applied Econometrics,* 21(4), pp. 419-438.

Camacho, M., Lovcha, Y. y Pérez-Quiroz, G. (2015). Can we use seasonally adjusted indicators in dynamic factor models? *Studies in Nonlinear Dynamics and Econometrics,* 19, pp. 377-391.

Caro, A. y Peña, D. (2020). A Test for the Number of Factors in Dynamic Factor Models. *Working Paper Universidad Carlos III de Madrid.*

Caporello, G. y Maravall, A. (2008). *Program TSW, Version Beta 1.0.4.* Banco de España.

De Livera, A. M., Hyndman, R. J. y Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association,* 106(496), pp. 1513-1527.

Diebold, F. X. (2003). Big data dynamic factor models for macroeconomic measurement and forecasting. En: M. Dewatripoint y S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications* (pp. 115-122). Eighth World Congress of the Econometric Society.

Escribano, A. y Peña, D. (1994). Cointegration and common factors. *Journal of Time Series Analysis,* 15(6), pp. 577-586.

Forni, M., Hallin, M., Lippi, M. y Reichlin, L. (2000). The generalized dynamic factor model: Identification and estimation. *The Review of Economic and Statistics,* 82(4), pp. 540-554.

Gómez, V. y Maravall, A. (1994). Estimation, prediction, and interpolation for nonstationary series with the Kalman Filter. *Journal of the American Statistical Association,* 89(426), pp. 611-624.

García-Martos, C., Rodríguez, J. y Sánchez, M. J. (2011). Forecasting electricity prices and their volatilities using unobserved components. *Energy Economics,* 33(6), pp. 1227-1239.

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

Hylleberg, S., Engle, R., Granger, C. W. y Yoo, B. (1990). Seasonal Integration and Cointegration. *Journal of Econometrics,* 44(1-2), pp. 215-238.

Jevons, W. S. (1862). On the Study of Periodic Commercial Fluctuations. En: H. S. Foxwell (ed.), *Investigations in Currency and Finance* (pp. 13-118). London: Macmillan.

Jungbacker, B. y Koopman, S. J. (2015). Likelihood-based dynamic factor analysis for measurement and forecasting estimation. *Econometrics Journal,* 18(2), pp. C1-C21.

Klein, J. L. (1997). *Statistical Visions in Time A History of Time Series Analysis 1662-1938.* Cambridge (United Kingdom): Cambridge University Press.

Lam, C. y Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *Annals of Statistics,* 40(2), pp. 694-726.

Lutkepohl, H. (2013). *Introduction to Multiple Time Series Analysis.* Springer Science & Business Media.

Melo, L. F., Nieto, F. H., Posada, C. E., Betancourt, Y. R. y Barón, J. D. (2001). Un índice coincidente para la actividad económica de Colombia. *ENSAYOS Sobre Política Económica,* 19(40), pp. 46-88.

Nieto, F. H., Peña, D. y Saboyá, D. (2016). Common Seasonality in Multivariate Time Series. *Statistica Sinica,* 26(4), pp. 1389-1410.

Peña, D. (2005). *Análisis de Series Temporales.* Madrid: Alianza Editorial.

Peña, D. y Box, G. E. P. (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association,* 82(399), pp. 836-843.

Peña, D. y Poncela, P. (2006). Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference,* 136(4), pp. 1237-1257.

Peña, D., Poncela, P. y Ruiz, E. (2021). *Nuevos métodos de predicción económica con datos masivos.* Madrid: Funcas.

Peña, D. y Tsay, R. S. (2021). *Statistical Learning for Big Dependent Data.* Wiley.

Poncela, P., Ruiz, E. y Miranda, K. (2020). Factor extraction using Kalman filter and smoothing: This is not just another survey. *Working paper 2020-05, Statistics and econometrics, Universidad Carlos III de Madrid.* Madrid.

Reinsel, G. C. (1997). *Elements of Multivariate Time Series Analysis.* Springer.

Stock, J. H. y Watson, M. W. (2016), Dynamic factor models: A brief retrospective. En: E. Hillebrand y S. J. Koopman (eds.), *Dynamic Factor Models* (Advances in Econometrics, Vol. 35). Emerald Group Publishing Limited.

Stock, J. H. y Watson, M. W. (2017). Twenty years of time series econometrics in ten pictures. *Journal of Economic Perspectives,* 31(2), pp. 59-86.

Tsay, R. S. (2014). *Multivariate Time Series Analysis.* New Jersey: John Wiley & Sons.

CAPÍTULO VII

Explorando pautas en series estacionales múltiples mediante técnicas multivariantes

Enrique Martín Quilis*

Se utilizan técnicas multivariantes (modelos fatoriales, análisis de conglomerados) para identificar pautas comunes y específicas en un vector de series temporales de elevada dimensión cuya sección cruzada es de naturaleza espacial. Estas técnicas se aplican en un contexto de series temporales de alta frecuencia caracterizadas por la presencia de diversos componentes (tendencia, ciclo, estacionalidad, efectos de calendario) que implica un notable aumento de la dimensión efectiva del conjunto de datos. La metodología propuesta es aplicada a una base de datos territorial de la economía española cuya cobertura es muy amplia, tanto temporal (1974-2019) como espacial (nivel provincial).

Palabras clave: estacionalidad, ciclos, extracción de señales, alieneamiento dinámico óptimo, análisis de conglomerados, análisis factorial.

* Agradezco la colaboración de Ana Abad, José Antonio Campo y Rafael Frutos en diversas etapas de este proyecto, así como las sugerencias de Daniel Peña, Pilar Poncela y Esther Ruiz. Las opiniones presentadas corresponden al autor, sin que coincidan de forma necesaria con las de la Agencia Tributaria.

1. INTRODUCCIÓN

El análisis de *big data* se caracteriza por "las tres V": volumen, variedad (o diversidad) y velocidad (Sathi, 2012; Kolanovic y Krishnamachari, 2017; Diebold, 2020). Las series temporales de datos económicos de alta frecuencia (por ejemplo, mensual) y muy alta frecuencia (por ejemplo, diaria) están formadas por una multiplicidad de componentes subyacentes muy distintos entre sí: atípicos, efectos de calendario (por ejemplo, fiestas móviles y laboralidad), tendencia, ciclo, estacionalidad e irregularidad. De esta manera y de forma casi automática, adquieren una dimensión muy superior a la de los datos observados, haciéndolo además de forma heterogénea. Aparecen así, rápidamente, las dos primeras V (volumen y variedad), afectando a la tercera V (velocidad), al complicarse el procesamiento de los datos y reducir por consiguiente su velocidad.

En este trabajo se van a utilizar dos técnicas centrales en el análisis de *big data,* análisis de conglomerados y modelos factoriales, en un contexto de series temporales múltiples cuya dimensión transversal tiene un significado geográfico. Este contexto dinámico va a requerir un uso bastante intensivo de técnicas de modelización de series temporales y de extracción de señales para realizar un adecuado procesamiento y descomposición de las series observadas. Asimismo, la elección de un concepto de distancia adaptado a la naturaleza dinámica de los datos es esencial.

La metodología econométrica consta de tres etapas, siguiendo el principio de "divide y conquista" para hacer frente a la maldición de las dimensiones que pesa sobre el análisis de series temporales múltiples. En la primera se realiza un análisis univariante de todas las series consideradas, con el fin de determinar cuál es la transformación de Box-Cox más apropiada, aislar los efectos deterministas (atípicos, fiestas móviles y ciclo semanal) de los estocásticos y realizar una descomposición de estos últimos en tendencia-ciclo, estacionalidad e irregularidad. A su vez, la aplicación de métodos de extracción de señales permite separar la tendencia del ciclo.

La segunda etapa es de carácter bivariante y trata de estimar una matriz de distancia mediante la aplicación del algoritmo de alineamiento óptimo (*DTW*, por *Dynamic Time Warping*) a todos los pares de series. Este algoritmo tiene en cuenta tanto la naturaleza dinámica de los objetos cuya semejanza se desea medir como la posibilidad de que existan desfases temporales entre ellos (por ejemplo, relaciones de adelanto o desfase). La utilización del enfoque DTW permite enlazar de manera consistente las etapas 1 (univariante, dinámica) con la 3 (multivariante, estática).

Finalmente, en la tercera etapa se utilizan métodos de formación de conglomerados (tanto jerárquicos como partitivos) y de análisis factorial para identificar pautas comunes en los componentes estacional y cíclico de las series consideradas. Estos métodos permitirán comprobar en qué medida ambas aglomeraciones son afines, tanto espacial como dinámicamente.

Desde un punto de vista económico, se examina la relación entre la estacionalidad (un componente estructural, no estacionario y candidato natural para incorporar información geográfica) y el ciclo (un componente transitorio aunque persistente, asintóticamente estacionario y muy condicionado por los factores macroeconómicos de corto y medio plazo). En particular, se trata de responder a la pregunta: ¿resuelven los agentes económicos sus programas de optimización tomando la estacionalidad como una restricción exógena o, por el contrario, determinan conjuntamente su comportamiento tanto estacional como no estacional? Esta segunda posibilidad extiende la crítica a la distinción entre tendencia y ciclo basada en los modelos estocásticos de crecimiento óptimo a la diferenciación entre los componentes estacional y no estacional (Prescott, 1986; Todd, 1990). Desde este punto de vista, los agentes económicos resuelven sus programas de optimización de una forma indiferenciada, de manera que los elementos estacional y no estacional de sus decisiones son aspectos distintos de un único proceso de decisión, siendo meramente facetas diferentes de una misma respuesta a impulsos comunes (Barsky y Miron, 1989; Beaulieu, MacKie-Mason y Miron, 1992; Cecchetti, Kashyap y Wilcox, 1997; Gerenew y Gourio, 2018).

Si la agrupación estacional coincide con la cíclica, se obtiene evidencia a favor de este punto de vista y, en particular, de la conveniencia de realizar un análisis del ciclo y de la coyuntura desagregado según las pautas estacionales. Si, por el contrario, la información estacional y la cíclica guardan poca relación, la balanza se inclina a favor de la exogeneidad del fenómeno estacional.

La estructura del trabajo es la siguiente. En la segunda sección se exponen los métodos de estimación de los componentes estacional y cíclico. A continuación, se describe el método de cálculo de la distancia. La cuarta sección presenta la metodología de formación de conglomerados utilizada y, en la quinta, se describen los datos empleados. Los resultados empíricos se ofrecen en la sexta sección. El trabajo termina con un apartado de conclusiones.

2. ESTACIONALIDAD Y CICLOS

El procedimiento utilizado para estimar los componentes estacional y cíclico de las series temporales analizadas consta de tres etapas: corrección de los efectos asociados a las observaciones atípicas y efectos de calendario, extracción basada en modelos ARIMA de las señales estacional y de tendencia-ciclo y, por último, estimación del ciclo por medio de un filtro de paso en banda aplicado a la serie de tendencia-ciclo obtenida en la etapa anterior. A continuación, se describe brevemente cada fase.

Se considera que la serie temporal observada puede ser expresada de acuerdo con la siguiente expresión:

$$z_t = o_t + n_t, \qquad [1]$$

siendo z_t, con $t = 1..n$, la serie observada, posiblemente transformada mediante la función de Box-Cox; o_t es el componente determinista, resultado de la combinación de efectos de

calendario (ciclo semanal, Pascua móvil) e intervenciones ligadas a valores atípicos y n_t es el componente estocástico.

Por su parte, o_t representa una combinación de modelos de intervención asociados a factores de tipo extraordinario que afectan a la serie de manera no recurrente junto con los efectos de calendario vinculados con el ciclo semanal y la Pascua móvil. La expresión completa es:

$$o_t = \Sigma_{h=1}^{k} V_h(B) I_t(t_h) + \Sigma_{j=1}^{7} \beta_j d_{j,t} + \gamma P_t(\tau), \qquad [2]$$

donde $I(t_h)$ es una variable binaria de tipo impulso, siendo t_h el periodo en el que tiene lugar el acontecimiento atípico. El filtro $V_h(B)$ recoge los efectos dinámicos asociados a la observación anómala. En este trabajo se consideran tres posibles tipos de atípicos: aditivos, transitorios y cambios de nivel. Asimismo, $d_{j,t}$ = *[(número de días de tipo j en el mes t) – (número de domingos en el mes t)],* con j = *lunes, ..., sábados* y $d_{7,t}$ es la diferencia entre la duración del mes t y la duración media de todos los meses. Finalmente, $P_t(\tau)$ representa la proporción que representa la semana de Pascua en el mes t, habiéndose considerado que su efecto se registra en los días anteriores al Domingo de Resurrección. En este trabajo se asume $\tau = 8$.

La especificación del componente estocástico sigue una representación autorregresiva, integrada y de medias móviles (ARIMA) de tipo multiplicativo (Box y Jenkins, 1976):

$$n_t = \frac{\theta_q(B)\Theta_Q(B^{12})}{\phi_p(B)\Phi_P(B^{12})(1-B)^d(1-B^{12})^D} a_t, \qquad [3]$$

donde $\phi(B)$ y $\theta_q(B)$ son, respectivamente, polinomios de orden p y q en el operador de desfases B y $\Phi_p(B^{12})$ y $\Theta_Q(B^{12})$ son polinomios de orden P y Q en B^{12}. Los filtros $(1\text{-}B)^d$ y $(1\text{-}B^{12})^D$ son operadores de diferenciación regular y estacional, controlados por los parámetros enteros d y D, respectivamente. Por último, a_t es una secuencia de ruido blanco con esperanza nula y desviación típica constante σ_a.

A su vez, el término estocástico n_t admite una descomposición, según la hipótesis de los componentes subyacentes, en tendencia-ciclo (p_t), estacionalidad (s_t) e irregularidad (i_t):

$$n_t = p_t + s_t + i_t. \qquad [4]$$

Una vez estimado el modelo ARIMA con análisis de intervención (AI) descrito en [1]-[3], es posible extraer tanto una señal estacional como una de tendencia-ciclo aplicando filtros de error cuadrático medio mínimo compatibles con dicho modelo ARIMA. De esta manera, se obtiene una estimación de los componentes subyacentes adaptada a las propiedades dinámicas de la serie y, merced al principio de descomposición canónica, libre de elementos irregulares de tipo ruido blanco. Una descripción muy completa de los métodos y modelos para el tratamiento de observaciones atípicas, efectos de calendario, modelización univariante y descomposición de series temporales se encuenta en Peña, Tiao y Tsay (2001).

La expresión general de este proceso de filtrado para la estacionalidad es:

$$\hat{s}_t = V_s(B,F)\hat{n}_t = k_s \Pi(B)\Pi(F)\Psi_s(B)\Psi_s(F)\hat{n}_t. \quad [5]$$

Para la tendencia-ciclo:

$$\hat{p}_t = V_p(B,F)\hat{n}_t = k_p \Pi(B)\Pi(F)\Psi_p(B)\Psi_p(F)\hat{n}_t, \quad [6]$$

donde k_s y k_p son parámetros que normalizan las funciones de ganancia de los filtros, $\Pi(B)$ es la expansión autorregresiva del modelo ARIMA de n_t, $\Psi_i(B)$ $(i = s, p)$, es la expresión de medias móviles del modelo teórico de los componentes y $\hat{n}_t$ es la estimación del componente estocástico obtenida al eliminar de la serie observada z_t sus elementos deterministas o_t (Maravall, 1987).

La señal de tendencia-ciclo así obtenida permite la estimación de un componente cíclico independiente al aplicar a aquella un filtro de paso en banda diseñado desde el dominio de la frecuencia. Dicho filtro es de tipo Butterworth, especificado para aproximar con una precisión dada a uno cíclico de tipo ideal (Pollock, 1999; Oppenheim y Schaffer, 1989; Bógalo y Quilis, 2003; Proakis y Manolakis, 2006).

De esta forma, el componente cíclico se obtiene según la siguiente expresión:

$$\hat{c}_t = H_c(B,F)\hat{p}_t = H_c(B,F)k_p \Pi(B)\Pi(F)\Psi_p(B)\Psi_p(F)\hat{n}_t, \quad [7]$$

donde $H_c(B,F)$ es el filtro cíclico de paso en banda antes mencionado y c_t es la señal cíclica.

Este método bietápico puede ser interpretado de forma bayesiana, ya que combina información *a priori* (un filtro fijo de tipo Butterworth) con la contenida en la tendencia estimada a partir de la muestra (por medio de un filtro adaptable de tipo Wiener-Kolmogorov). De esta manera, se obtiene la información *a posteriori*: una serie de fluctuaciones alrededor de la tendencia de periodicidad comprendida entre dos y ocho años.

Las principales ventajas de este enfoque son:

- La estimación de los componentes es compatible con las propiedades agregadas de las series, de forma que se evita la inducción de fenómenos espurios como, por ejemplo, la estimación de un componente estacional en una serie que carece de estacionalidad.
- El filtro usado en la estimación se adapta a las características de la serie observada, de forma que series con características estacionales distintas tendrán asimismo filtros distintos.
- El preprocesamiento de las series mediante el modelo ARIMA-AI permite estimar los componentes estocásticos sin la influencia distorsionadora asociada a las observaciones atípicas y a los efectos de calendario, lo que redunda en una mejor estimación de los mismos.

3. MEDIDA DE DISTANCIA

La formación de conglomerados se basa en dos elementos principales: una matriz de distancia (o similitud) entre los objetos que se desea agrupar y un procedimiento o algoritmo que, a partir de dicha distancia, determina la relación de pertenencia de los objetos respecto a los grupos. Cuando los objetos que se desea agrupar son series temporales, la definición de un concepto de distancia deviene especialmente difícil, debido a su multidimensionalidad implícita asociada a la existencia de componentes subyacentes (por ejemplo, tendencia, estacionalidad, etc.) y a la posibilidad de que existan relaciones dinámicas entre ellas (por ejemplo, relaciones de adelanto, coincidencia o retraso).

Por todo ello, se han presentado diversas medidas de distancia para series temporales. Estas medidas pueden basarse en un enfoque no paramétrico, bien en el dominio del tiempo o de la frecuencia (Caiado, Crato y Peña, 2006; Caiado, Crato y Poncela, 2020), o basado en modelos (Piccolo, 1990). La elección del enfoque es objeto de un amplio debate, muy condicionado por la naturaleza de las series temporales y el objetivo último de la agrupación (Galeano y Peña, 2000; Liao, 2005; Wang, Smith y Hyndman, 2006; Rani y Sikka, 2012).

En este trabajo se utiliza el algoritmo de alineamiento dinámico temporal DTW para calcular una matriz de distancia entre los elementos de un vector de series temporales (Sakoe y Chiba, 1978). DTW es un enfoque no paramétrico, basado directamente en la semejanza entre perfiles temporales (*shape-based*) y que se adapta muy bien a las características de los objetos que se desea clasificar: componentes subyacentes estimados mediante técnicas de extracción de señales.

El procedimiento DTW consta de dos etapas: cálculo de una matriz de distancia inicial y algoritmo de alineación óptima mediante programación dinámica. La matriz de distancia final entre todos los pares de series temporales se obtiene aplicando dicho algoritmo a las correspondientes matrices iniciales. A continuación, se describen ambas etapas.

3.1. Matriz inicial de distancia

Asumiendo que $z_{i,t}$ y $z_{j,t}$ son dos componentes de un vector de series temporales de dimensión k, $Z_t = (z_{1,t}...z_{k,t})'$, la matriz de distancia inicial toma como punto de partida la diferencia absoluta entre todas las observaciones efectuadas en t y en s de las dos series[1]:

$$C_{t,s} = |z_{i,t} - z_{j,s}| \qquad t, s = 1...n. \qquad [8]$$

La información contenida en la matriz anterior se va acumulando sobre todos los pares (t,s) mediante la siguiente recursión:

$$D_{t,s} = C_{t,s} + min\,[D_{t-1,s}, D_{t-1,s-1}, D_{t,s-1}] \qquad t, s = 2...n. \qquad [9]$$

[1] Con el fin de aligerar la notación, se omiten los índices *i* y *j*.

La condición inicial para la recursión [9] es $D_{1,1} = C_{1,1}$. Si el número de observaciones es elevado, la matriz D puede adquirir un gran tamaño, resultando su cálculo computacionalmente costoso. Por esta razón y con el fin de reducir la carga numérica, se suele acotar su cálculo a una ventana de proximidad entre las observaciones (p.e. una distancia entre t y s inferior al 10 % del tamaño muestral).

La matriz [9] representa, para cada par temporal (t,s), una medida de la similitud entre las series $z_{i,t}$ y $z_{j,t}$ acumulada hasta dicho par. Para obtener una medida sintética de distancia entre ambas series es necesario transformar dicha matriz en un escalar. Esta transformación se realiza mediante el algoritmo de alineación óptima que se describe a continuación.

3.2. Algoritmo de alineación óptima

El primer paso de este algoritmo consiste en determinar un emparejamiento entre las observaciones de $z_{i,t}$ y $z_{j,t}$ de manera que la similitud entre ellas, cuantificada mediante [9], sea máxima. Como dicha cuantificación se realiza de forma acumulativa desde $t = 1$ hasta $t = n$, la determinación de dicho emparejamiento se realiza en sentido inverso, $t = n$ hasta $t = 1$, siguiendo el enfoque de la programación dinámica.

De esta manera, tomando $D_{n,n}$ como condición inicial, se retrocede hasta la observación inicial y se emparejan las observaciones de $z_{i,t}$ y $z_{j,t}$ para las que D es mímina, utilizando el mismo entorno que se ha considerado en la recursión [9] mediante la que ha sido calculada. Así, avanzando en sentido temporal inverso y minimizando D en cada paso, se determina una secuencia para ambas variables, $i(t)$ y $j(t)$, de forma que su similitud es máxima. A continuación se detalla el procedimiento.

La condición inicial del algoritmo se sitúa al final de la muestra temporal: $t,s = n$; $i(t) = j(t) = n$ y $c = 1$. A partir de ahí, se aplica un bucle inverso desde $t = n$ hasta $t = 1$:

$$a,b = arg\ min_{\alpha,\beta}(D_{t-\alpha,s}, D_{t-\alpha,s-\beta}, D_{t,s-\beta}). \qquad [10]$$

En este paso se determinan los índices temporales de las dos series $z_{i,t}$ y $z_{j,t}$ de forma que se emparejan de forma óptima, minimizando la distancia en el mismo entorno sobre el que ha sido calculada según [9].

Una vez que se ha determinado el par (a,b), se actualiza el pivote temporal: $t = t\text{-}a$ y $s = s\text{-}b$ y se asigna este par temporal a la alineación óptima $i(t)$ y $j(t)$. Nótese que tanto a como b solo pueden adoptar los valores 0 o 1, en virtud de la recursión [9] usada para calcular D. Finalmente, se incrementa el contador de iteraciones, $c = c + 1$, y se repite [10] hasta alcanzar el par inicial $t = s = 1$.

Una vez que se ha determinado el emparejamiento óptimo, $i(t)$ y $j(t)$, la distancia final agregada entre las series $z_{i,t}$ y $z_{j,t}$ se calcula sumando todos los valores de la matriz D que han sido asociados mediante dicho emparejamiento:

$$DTW_{i,j} = \Sigma_{t=1}^{n} D_{i(t),j(t)}. \qquad [11]$$

Como ya se ha señalado, la matriz de distancia final [11] se calcula sobre los componentes subyacentes de un vector de series temporales[2]. Asimismo, estas matrices servirán de métrica para los algoritmos de formación de conglomerados que se describen en la siguiente sección.

4. FORMACIÓN DE CONGLOMERADOS

Una vez estimados los componentes estacionales de las series que se desea analizar, se procede a formar grupos de series con patrones estacionales similares. La formación de estas agrupaciones se realiza mediante el análisis de conglomerados (Everitt *et al.,* 2011; Kassambara, 2017).

El análisis de conglomerados es una técnica estadística no paramétrica, de tipo exploratorio, que agrupa los objetos atendiendo a su semejanza. Se puede considerar también como un método de aprendizaje no supervisado, mediante el que se busca la identificación de patrones a través de la detección de pautas o características similares, sobre las que se dispone de poca o ninguna información *a priori* (Sathi, 2012).

Estas dos características proporcionan al análisis de grupos una gran flexibilidad, ya que se adapta muy bien a situaciones que requieren la identificación de regularidades empíricas como paso previo a la elaboración de modelos estadísticos explícitos. Asimismo, el análisis de conglomerados resulta especialmente útil cuando se desea reducir drásticamente la dimensión de grandes masas de información, de forma que se selecciona un representante de cada grupo en lugar de la población completa. En este sentido, este tipo de análisis se asemeja a un muestreo endógeno en el que los propios objetos determinan su representante.

Existen diversos métodos de formación de conglomerados basados en criterios de optimización. En general, se trata de formar grupos con la máxima heterogeneidad entre ellos y la mínima dentro de cada uno de ellos. Naturalmente, el mayor problema de estos métodos radica en la explosión combinatoria a que da lugar una búsqueda exhaustiva. Con el fin de dotar al procedimiento de optimización de un contenido operativo en situaciones reales, se han propuesto diversos algoritmos de formación de un número dado de conglomerados. En este trabajo se utiliza el de las k-medias por su relativa eficiencia y por la solidez de su planteamiento teórico (Faber, 1994; Everitt *et al.,* 2011).

Se ha preferido un algoritmo partitivo en lugar de uno jeráquico por su mayor flexibilidad, escalibilidad y porque asegura la idoneidad de los respresentantes de cada grupo (centroides), propiedad muy importante si se desea utilizar el análisis de conglomerados para reducir la muestra a procesar. Como ya se ha señalado, la matriz de distancia sobre la que

[2] En particular, los componentes estacional y cíclico estimados mediante la metodología descrita en la sección segunda.

se aplica este algoritmo es la que ha sido calculada mediante el proceso DTW descrito en la sección anterior.

El algoritmo de las k-medias opera de la siguiente forma. En primer lugar, se realiza una agrupación aleatoria de los objetos en *G* conglomerados. En este trabajo los objetos que se desea agrupar son las series temporales asociadas a los componentes estacional y cíclico de un vector de series temporales. El número de conglomerados se asume como dado[3].

A continuación, se selecciona una serie temporal como representante (centroide) de cada grupo, buscando aquella que se encuentra, en promedio, más próxima a todas las demás que forman parte del mismo conglomerado. Una vez determinados los centroides, se revisa la asignación de series a grupos. De esta manera, se considera que el objeto *m* pertenece al grupo *h* si la distancia que lo separa del correspondiente centroide de los *G* grupos considerados es la menor posible:

$$m \in h \quad \Leftrightarrow h = arg\, min_g\, (DTW_{m,g}) \quad g = 1..G. \qquad [12]$$

Después de revisar la asignación de las series temporales a los grupos, se determinan los nuevos centroides (representantes) siguiendo el mismo criterio de distancia mínima antes expuesto, modificándose asimismo la asignación según el criterio definido en la expresión [12].

El proceso de cálculo de centroides y reasignación continúa hasta que se satisface algún criterio de convergencia (por ejemplo, que la variación en valor absoluto de todos los centroides sea inferior a un umbral predeterminado).

El algoritmo de las k-medias requiere que el número de conglomerados a formar sea conocido. En muchas aplicaciones, como la presente, esta información preliminar no está disponible y ha de ser obtenida mediante alguna investigación previa. En este trabajo se ha utilizado una aglomeración jerárquica mediante el método de Ward (1963) con el fin de disponer de una estimación del número de grupos. Se ha seleccionado este método porque incorpora explícitamente una función objetivo compatible con los criterios de optimización antes expuestos.

En resumen, el proceso de formación de conglomerados consta de dos etapas: 1) determinación preliminar del número de grupos *G* mediante el examen del dendrograma generado por el método (jerárquico) de Ward y, 2) aplicación del algoritmo de las k-medias, tomando *G* como número apropiado de conglomerados.

5. DATOS

Los datos utilizados en este trabajo son las pernoctaciones en establecimientos hoteleros procedentes de la *Encuesta de Ocupación Hotelera (EOH),* elaborada por el Instituto Nacional

[3] Más adelante se detalla cómo se determina *G*.

de Estadística (INE, 2019). Esta encuesta está dirigida a hoteles y acampamentos, sobre la base del marco que proporcionan los directorios de las consejerías de turismo de las comunidades autónomas, cuya actualización se realiza de forma continua. La muestra está diseñada mediante un muestreo estratificado, siendo los estratos la provincia y la categoría hotelera. La investigación es exhaustiva excepto en aquellos estratos con un número grande de establecimientos, para los cuales se selecciona una muestra.

La recogida de la información es mensual, contestando cada establecimiento durante un período de siete días seguidos elegidos de manera que entre todos los establecimientos de cada estrato se cubren todos los días del mes. Dentro del marco conceptual de la *EOH,* se entiende por pernoctación la ocupación por una persona de una plaza o una cama supletoria dentro de una jornada hotelera y en un mismo establecimiento. La ocupación por una persona en el mismo día de dos o más plazas en establecimientos distintos da lugar a más de una pernoctación.

El período muestral utilizado es 1976-2019. De esta manera, se dispone de un conjunto de series temporales de gran longitud (más de 40 años), de forma que es posible realizar un análisis muy completo de su comportamiento cíclico. Esta amplitud temporal tiene un coste, bajo la forma de falta de homogeneidad en la información de base, especialmente en el caso de la sustitución, a partir de 1999, de la *Encuesta sobre Movimientos de Viajeros en Establecimientos Hoteleros* por la actual *Encuesta de Ocupación Hotelera.* Aunque ambas comparten los mismos objetivos y conceptos, la nueva encuesta amplió el ámbito de la investigación y cambió el marco poblacional, introduciendo un cambio de nivel de signo positivo. Análogamente, en 1993 se dejó de encuestar a los establecimientos hoteleros de menor categoría, lo que produjo un efecto escalón de signo contrario.

Afortunadamente, estos cambios metodológicos de tipo permanente son identificados y corregidos mediante el análisis de intervención descrito en la sección 2, haciendo que variables exógenas de tipo cambio de nivel recojan las variaciones metodológicas cuando estas resultan significativas. Por lo que se refiere a las características de las series elegidas que más interesan en este trabajo, destacan la periodicidad mensual y su elaboración a nivel provincial. Ambas características permiten estimar la pauta estacional de cada una de las provincias y su posterior clasificación de acuerdo a la misma. Además, estas series son un buen indicador del ciclo de la actividad turística debido a su cobertura y a su objeto de investigación.

Los datos correspondientes al año 2020, severamente afectados por la crisis sanitaria debida a la COVID-19, no han sido considerados en el estudio porque, dada su naturaleza extraordinariamente atípica y su gran impacto (INE, 2020), requieren un estudio específico que está claramente fuera del objetivo de este trabajo.

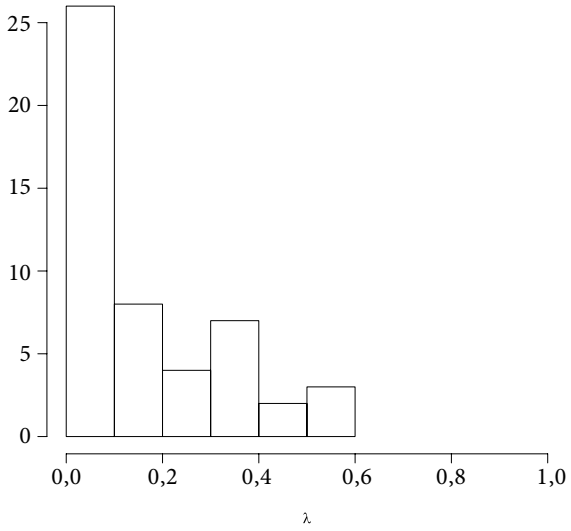
6. RESULTADOS EMPÍRICOS

A continuación se examinan los resultados obtenidos mediante la aplicación de los métodos expuestos en las secciones anteriores a los datos disponibles para este estudio.

La determinación de la transformación Box-Cox más apropiada se ha realizado utilizando el criterio de Guerrero (1993), implementado en el paquete R forecast (Hyndman y Khandakar, 2008). Como puede apreciarse en la figura 1, la distribución del parámetro óptimo λ de dicha transformación está marcadamente concentrado en torno a cero, por lo que se ha aplicado la transformación logarítmica a todas las series.

Figura 1.

Distribución del parámetro Box-Cox óptimo



Fuente: Elaboración propia.

La descomposición de las series log-transformadas, descontados los posibles efectos deterministas vinculados con observaciones anómalas y efectos de calendario, se ha realizado mediante el programa X13-ARIMA-SEATS[4] (U.S. Census Bureau, 2017; Sax y Eddelbuettel, 2018). En la figura 2 se presenta, a título de ejemplo, la descomposición completa de las pernoctaciones registradas en la provincia de Alicante, incluyendo la estimación de una tendencia secular mediante el filtro de Hodrick-Prescott y, como residuo, una señal cíclica.

En este trabajo la estimación de una señal cíclica independiente se obtiene directamente, mediante un filtro de paso en banda de tipo Butterworth (MathWorks, 2013). Este filtro está diseñado para extraer las fluctuaciones comprendidas entre dos y ocho años. La señal cíclica obtenida mediante este filtro muestra un perfil similar al obtenido con el de Hodrick-Prescott pero es mucho menos irregular[5]. Esta mayor suavidad y precisión hace más fiable el cálculo

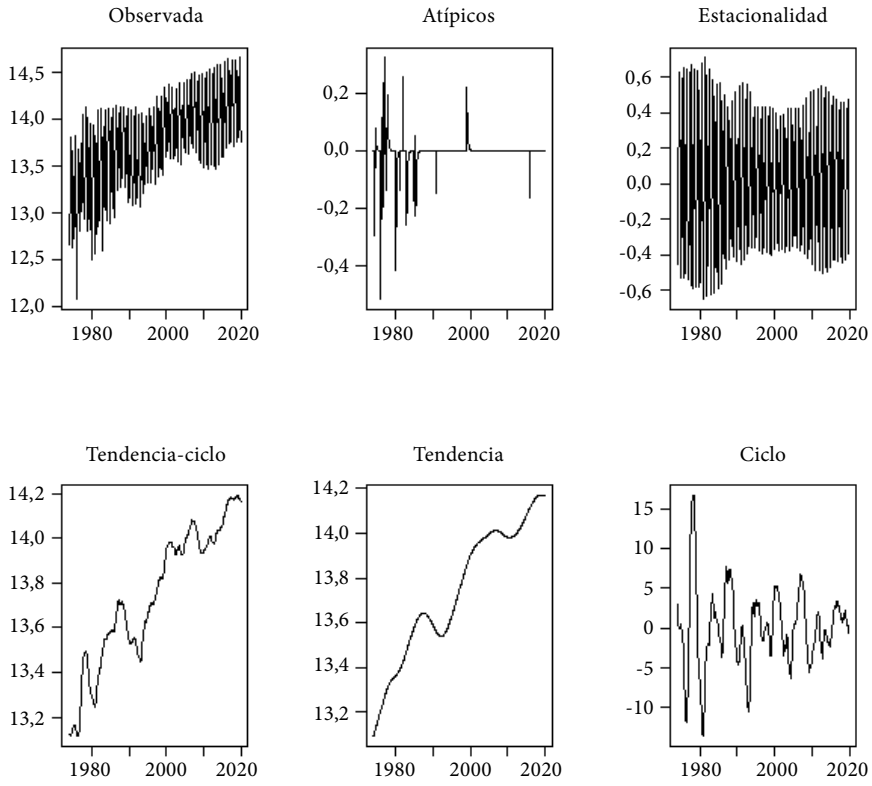
[4] Usando la opción de descomposición basada en modelos ARIMA.

[5] El filtro de Butterworth (paso en banda) elimina la irregularidad pero el de Hodrick-Prescott (paso bajo) la mantiene intacta.

de la distancia entre las series consideradas, especialmente dada la naturaleza no paramétrica y basada en perfiles del método DTW. En la figura 3 se comparan las señales cíclicas del ejemplo considerado (Alicante).

Figura 2.

Descomposición de la serie de pernoctaciones en Alicante

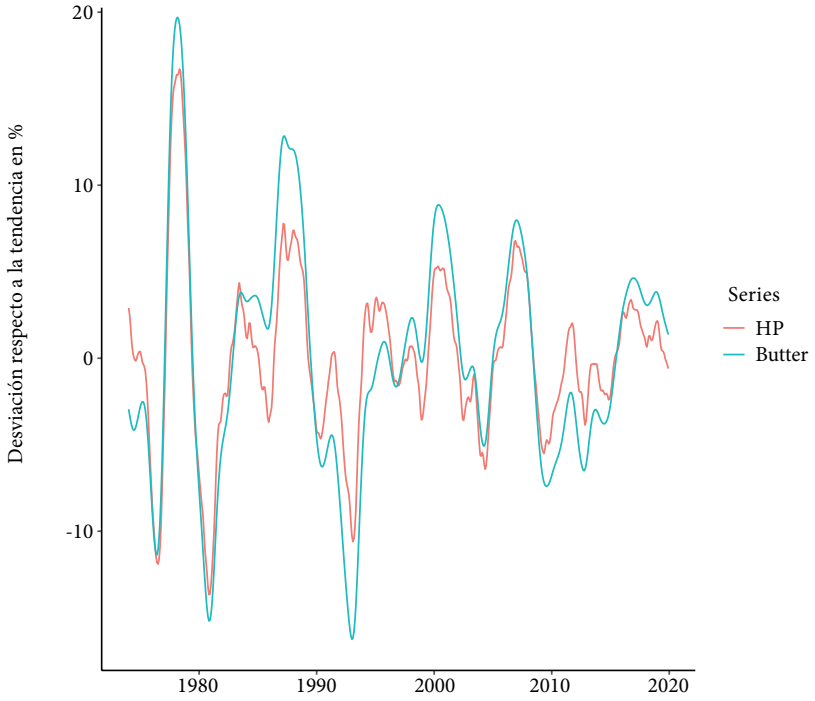


Fuente: Elaboración propia.

La estimación de la matriz de distancia mediante DTW y el análisis de conglomerados se ha realizado mediante diversos paquetes programados en el lenguaje R (R Core Team, 2019), destacando dtwclust (Sardá-Espinosa, 2019), cluster (Maechler *et al.,* 2017), factoextra (Kassambara, 2020) y dendextend (Galili, 2015). Finalmente, la estimación de los modelos factoriales estáticos se ha efectuado con la librería factorLib, programada en Matlab (Quilis, 2019).

Figura 3.

Señal cíclica (Alicante): estimaciones alternativas



Fuente: Elaboración propia.

6.1. Aglomeración estacional

La aglomeración generada por el método de Ward aplicado a la matriz de distancias DTW de los cincuenta factores estacionales puede ser examinada mediante el índice silueta (Rousseeuw, 1987). La figura 4 muestra los índices correspondientes para cada provincia, considerando un número de grupos comprendido entre dos y siete.

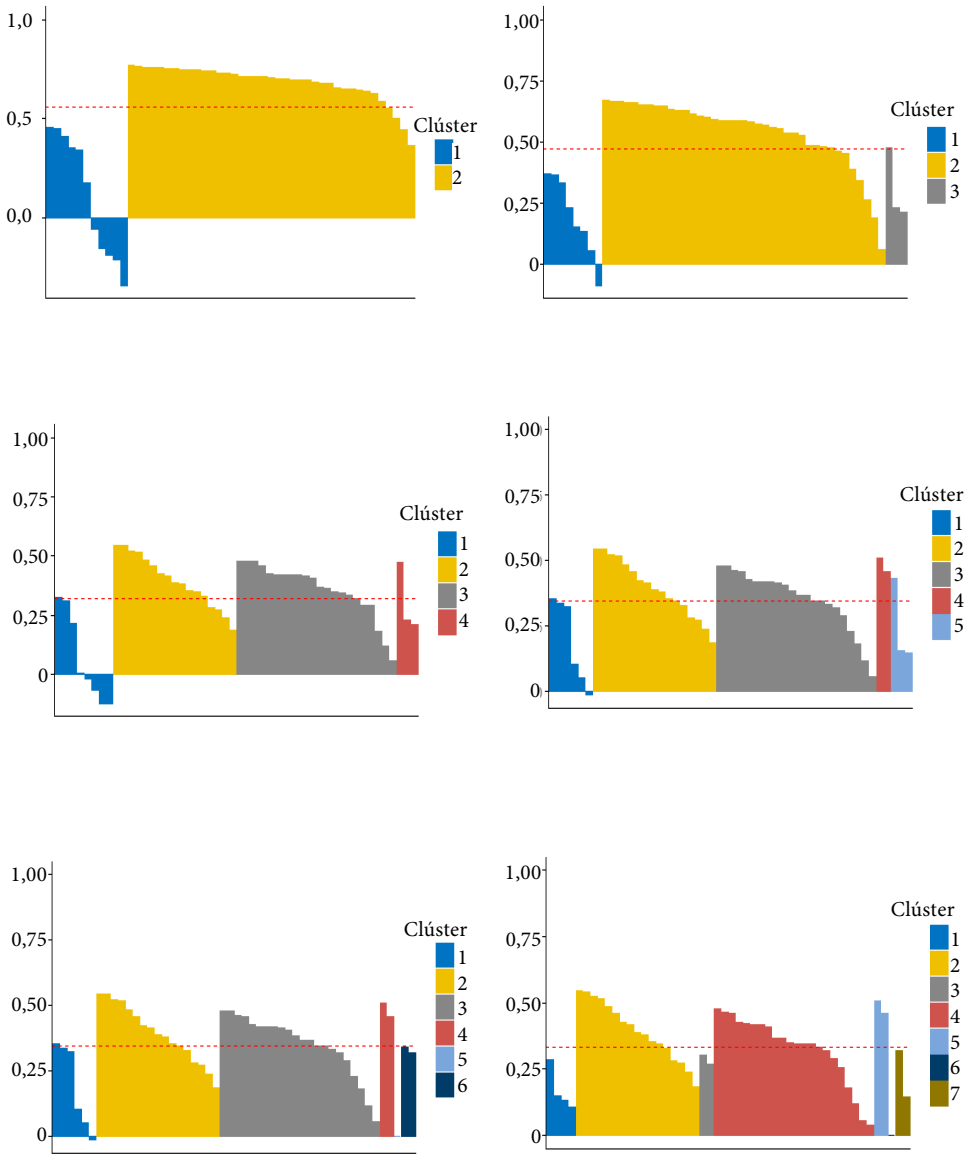
Buscando reducir al máximo posible el número de asignaciones inadecuadas[6], el número de grupos apropiado se sitúa entre cinco y siete. El valor medio del índice es muy similar en los tres casos, por lo que resulta difícil usarlo como criterio de selección.

Con el fin de refinar el análisis, se parte de seis grupos y se comparan las agrupaciones correspondientes con la distribución de las cargas en un modelo con dos factores estáticos. La figura 5 muestra dicha comparación.

[6] Representadas por valores negativos del índice silueta.

Figura 4.

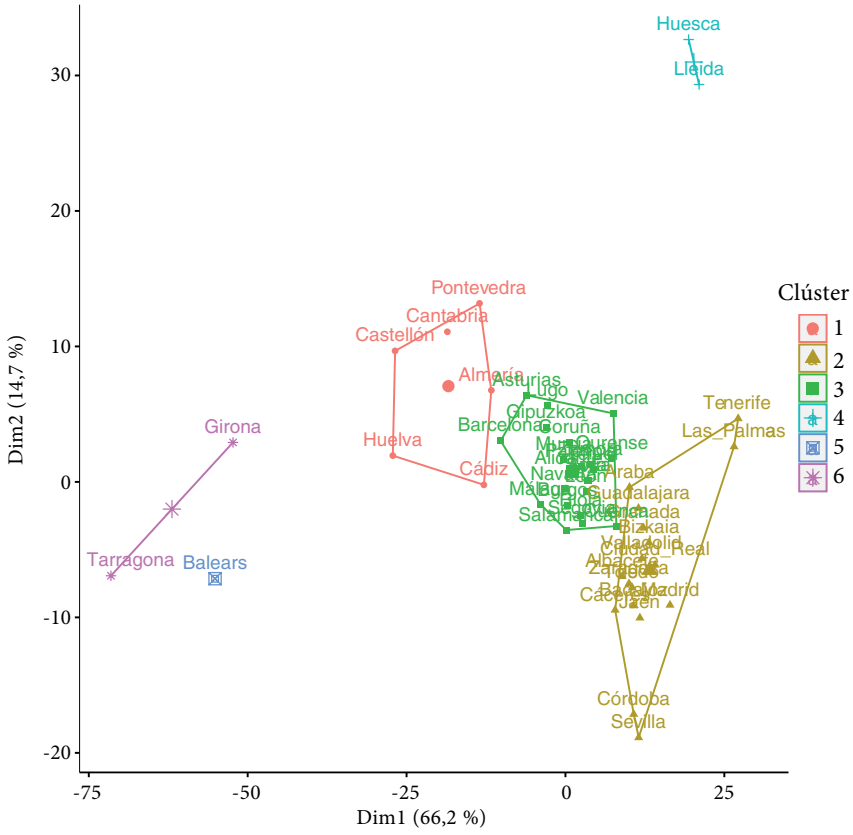
Estacionalidad: agrupación jerárquica (Ward). Índice silueta



Fuente: Elaboración propia.

Figura 5.

Estacionalidad: agrupación jerárquica (Ward) vs. cargas factoriales



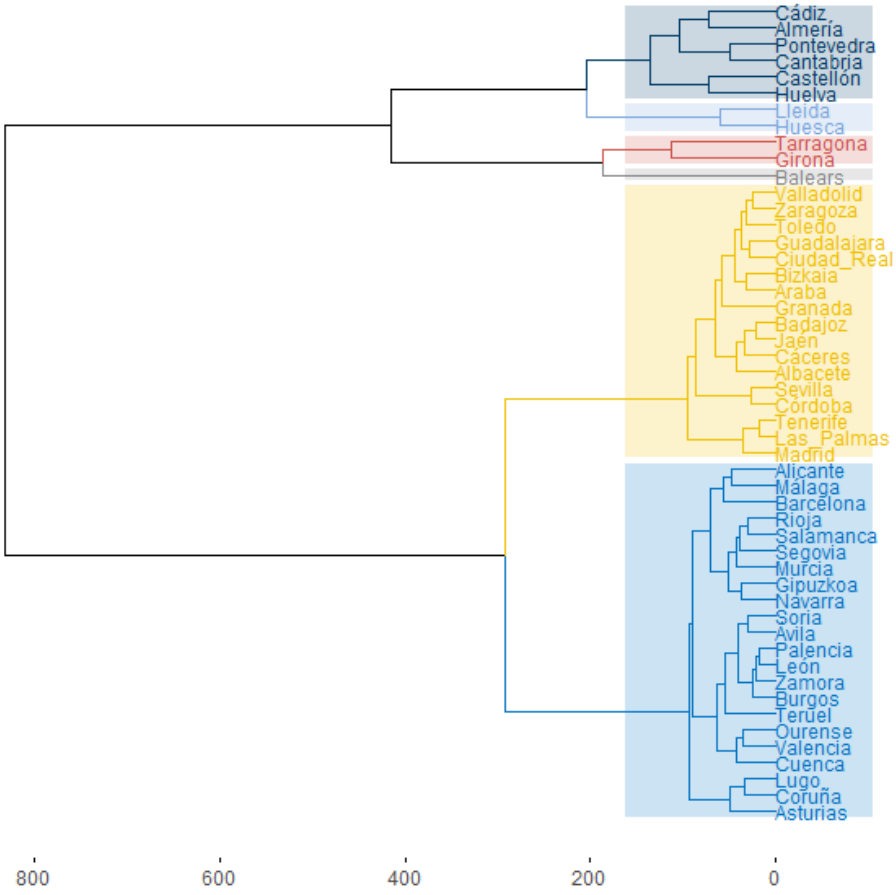
Fuente: Elaboración propia.

En primer lugar, se aprecian dos grupos muy diferenciados y geográficamente afines (Huesca y Lleida, por una parte, y Balears, Girona y Tarragona, por otra). En segundo lugar, aparecen otros tres de mayor tamaño mostrando una mayor continuidad en su aglomeración respecto a los dos factores. Dentro de estos tres grupos destaca el central, muy cohesionado y relativamente poco identificado con el primer factor. El paso de cinco a seis grupos separa Balears y el paso de seis a siete particiona uno de los grupos contiguos en dos. Se ha considerado que esta segunda partición es poco informativa pero no así la primera, debido a tres características especiales de Balears: su insularidad, su peculiar tendencia y su destacado peso[7].

[7] Baleares representa, en promedio, el 23 % del total de pernoctaciones, considerablemente más que las dos siguientes provincias (Las Palmas y Tenerife, con un 9 % y un 8 %, respectivamente).

Figura 6.

Estacionalidad: agrupación jerárquica (Ward). Dendrograma



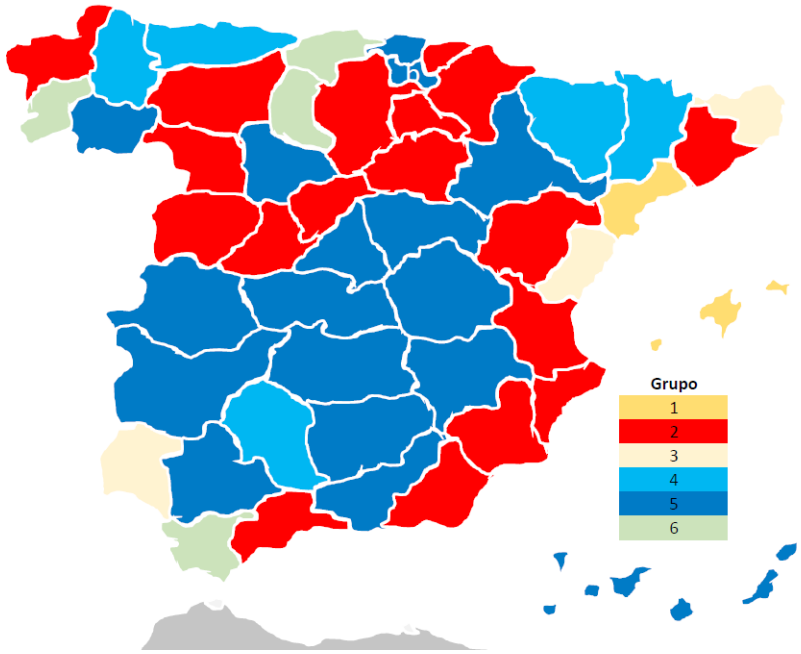
Fuente: Elaboración propia.

Considerando seis grupos, el dendrograma correspondiente ofrece la siguiente aglomeración:

En la figura 6 destacan dos grandes grupos con una distancia de fusión similar y cuatro grupos relativamente pequeños y heterogéneos, a tenor de su mayor distancia de fusión. Dentro de estos grupos, las provincias pirenaicas (Huesca y Lleida) muestran una mayor afinidad.

Figura 7.

Estacionalidad: agrupación particional (k-medias). Distribución geográfica



Fuente: Elaboración propia.

Considerando $G = 6$, el algoritmo de las k-medias descrito en la sección anterior da lugar a la distribución geográfica de los grupos que puede observarse en el mapa de la figura 7.

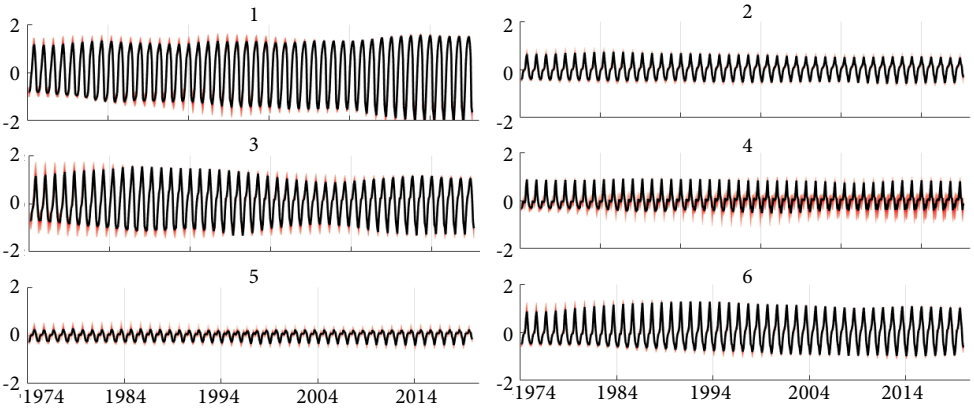
Las agrupaciones generadas son muy similares a las generadas con el método jerárquico de Ward, destacando algunas variaciones: Balears ha sido fusionada con Tarragona en un grupo y las provincias pirenaicas (Huesca y Lleida) han sido agrupadas junto a otras tres relativamente distantes (Asturias, Lugo y Huelva).

Para interpretar adecuadamente la pauta territorial es conveniente examinar los perfiles estacionales de los seis grupos, cuyo promedio temporal aparece en la figura 8.

El grupo 1 (Baleares y Tarragona) se caracteriza por su intenso patrón estacional, que ha ido ganando amplitud con el tiempo y cuyo perfil intraanual muestra una extensa estación alta (desde junio hasta septiembre). El segundo grupo es más numeroso (19 provincias) y su patrón estacional es bastante estable. Este grupo posee un perfil concentrado en el trimestre estival y, especialmente, en el mes de agosto. Geográficamente, abarca casi todas las provincias costeras mediterráneas pero no se identifica exclusivamente con esa característica, ya que

Figura 8.

Factores estacionales agrupados



Fuente: Elaboración propia.

la mayor parte del grupo está formado por provincias interiores, fundamentalmente del cuadrante noroccidental de la península.

El grupo 3 es un grupo pequeño (Girona y Tarragona), geográficamente afín y con una estacionalidad evolutiva. Su perfil intraanual marca un máximo en agosto, en fuerte contraste con los meses valle (diciembre y enero). El cuarto grupo es territorialmente heterogéneo (dos provincias pirenaicas, dos cantábricas y una interior), factor que puede explicar su heterogeneidad transversal. El perfil intraanual muestra el contraste básico entre verano e invierno de los dos grupos anteriores, si bien de forma menos acusada.

El grupo 5, uno de los dos más grandes y relativamente homogéneo internamente, se caracteriza por su reducido rango de variación estacional y su localización geográfica predominantemente interior, con la notable excepción de las dos provincias canarias. Finalmente, el sexto grupo es muy heterogéneo geográficamente. Su pauta estacional es bastante intensa y evolutiva.

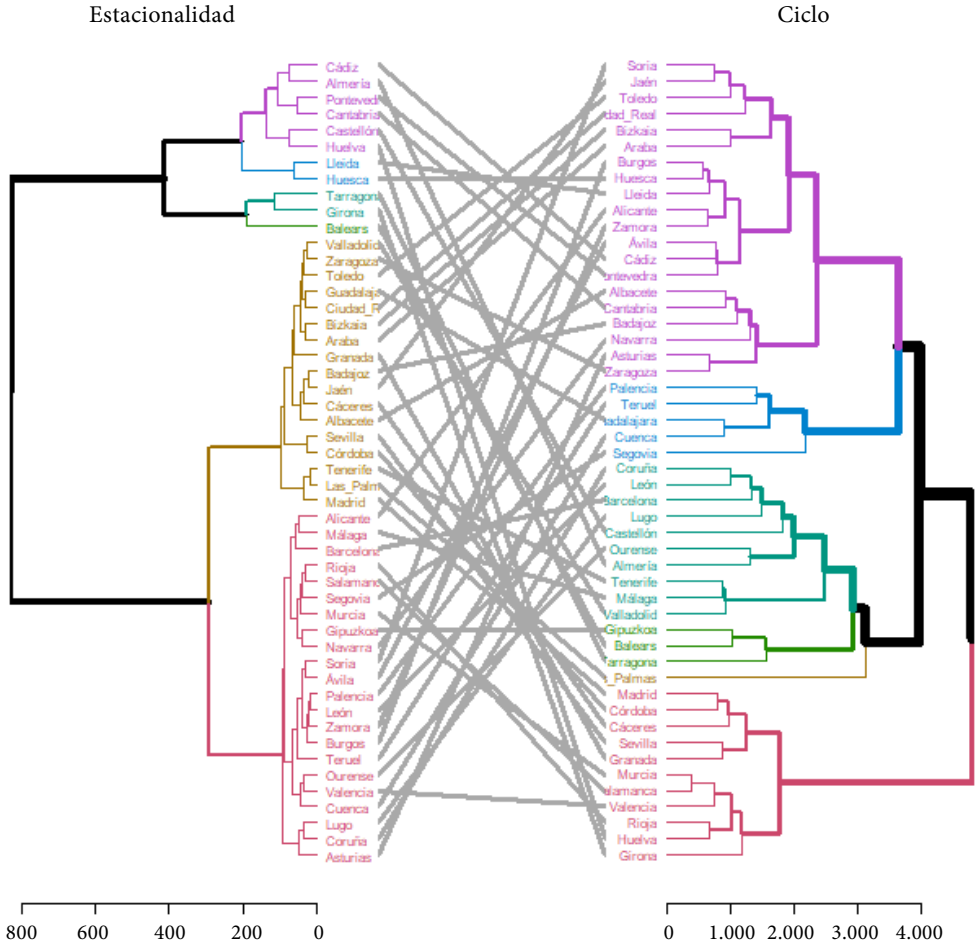
6.2. Conformidad cíclica

Una vez realizada la agrupación de las series provinciales según su pauta estacional, se puede responder a las preguntas planteadas en la introducción, comprobando si esta agrupación se reproduce al considerar su comportamiento cíclico.

En primer lugar, se ha realizado una comparación directa entre los dengrogramas obtenidos mediante el método de Ward aplicado a las matrices de distancia estacional y cíclica. La figura 9 representa ambos grafos y su correspondencia:

Figura 9.

Agrupación jerárquica: correspondencia entre dendrogramas



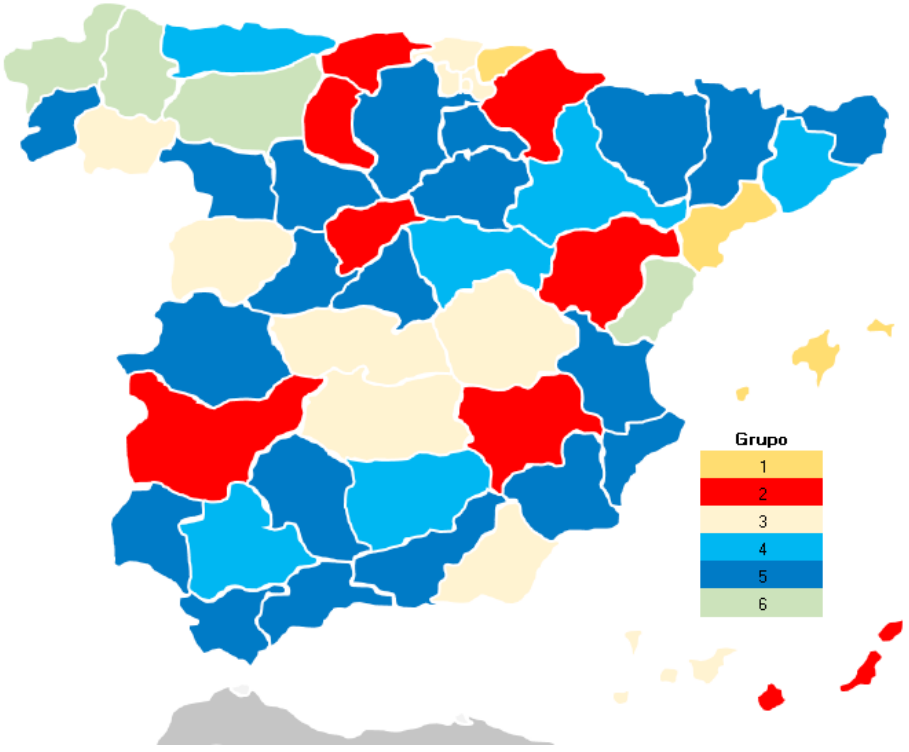
Fuente: Elaboración propia.

La coincidencia entre ambos grafos es moderada, a tenor del valor intermedio del índice de entrelazamiento (*entanglement*): 0,56. Por otra parte, la comparación geográfica entre ambas agrupaciones confirma el resultado anterior, pudiendo apreciarse una menor conexión territorial entre los seis grupos cíclicos así como un solapamiento muy contenido entre ambos. Esta menor conexión se muestra en el mapa de la figura 10.

Otra forma de examinar la conformidad entre ambas agrupaciones consiste en comparar el factor común de las señales cíclicas de cada uno de los seis grupos considerados con el

Figura 10.

Ciclo: agrupación jerárquica (Ward). Distribución geográfica



Fuente: Elaboración propia.

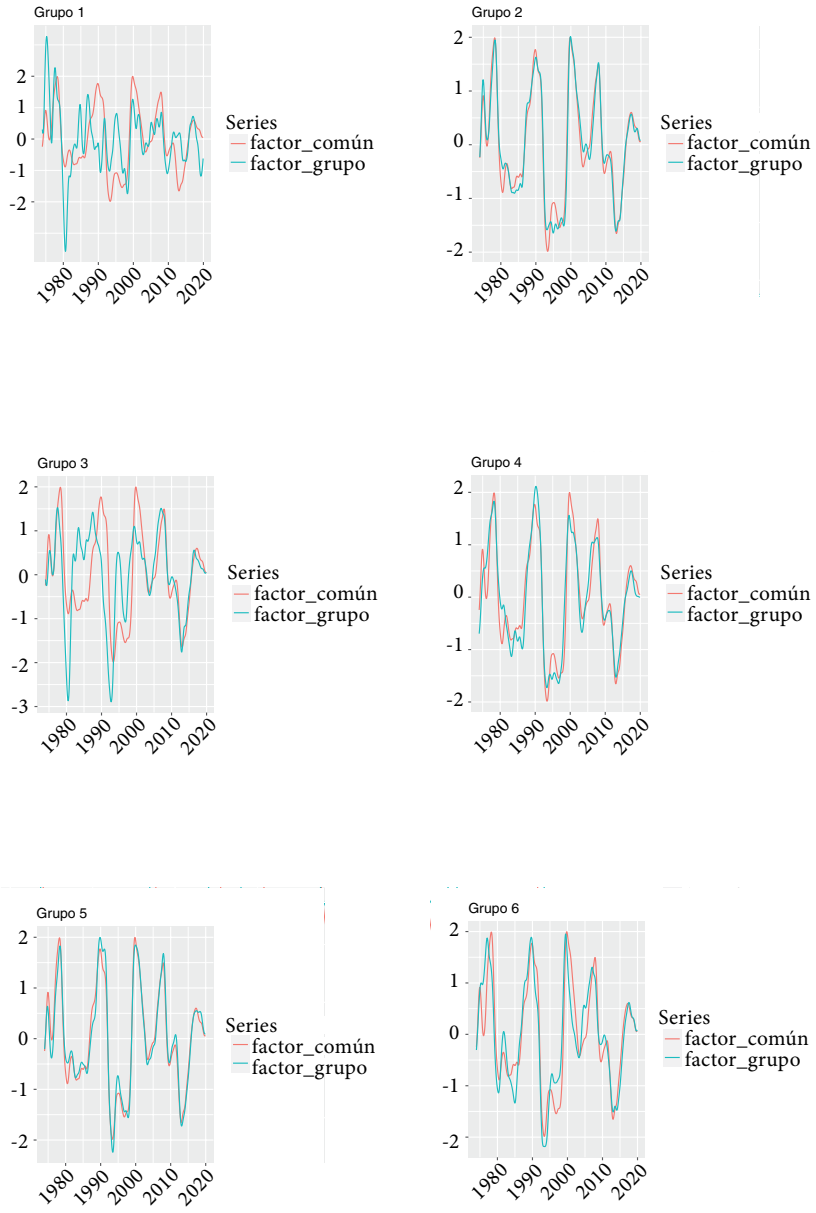
factor común de las cincuenta series provinciales. Los primeros representan la pauta cíclica específica de cada grupo mientras que el segundo sintetiza el ciclo nacional. La figura 11 muestra las series temporales correspondientes.

La comparación entre ambos factores muestra, por lo general, una elevada conformidad. Únicamente las provincias de los grupos 1 (Baleares y Tarragona) y 3 (Girona, Castellón y Huelva) presentan una correlación moderada con el factor nacional, sugiriendo un solapamiento entre las pautas estacionales más específicas de ambos grupos y la idiosincrasia de su comportamiento cíclico. La información proporcionada por las correspondientes funciones de correlación cruzada confirma el diagnóstico gráfico anterior, tal y como puede apreciarse en la figura 12.

En general, predomina una pauta dinámica esencialmente coincidente entre el factor nacional y el específico de cada grupo. Nuevamente, sólo el grupo 3 (Girona, Castellón y Huelva) muestra un cierto adelanto respecto al ciclo común a las cincuenta provincias españolas.

Figura 11.

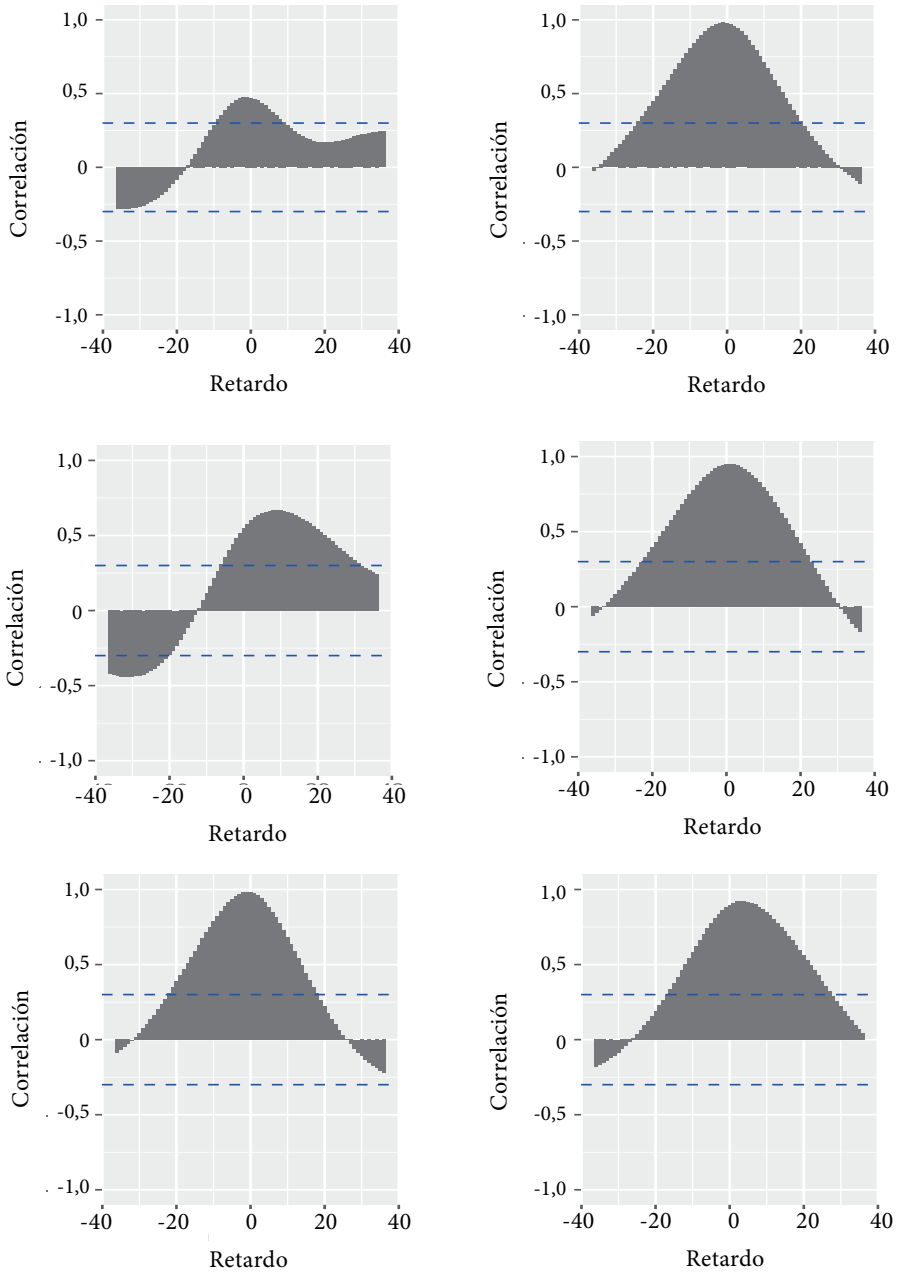
Factores cíclicos específicos vs. factor común



Fuente: Elaboración propia.

Figura 12.

Factores cíclicos específicos vs. factor común



Fuente: Elaboración propia.

7. CONCLUSIONES

El análisis realizado en este trabajo permite identificar una aglomeración de los factores estacionales modulada, principalmente, por la amplitud de sus perfiles intraanuales. Esta aglomeración se concentra en dos grandes grupos que muestran una notable conexión territorial. Se identifican también otros cuatro grupos de menor tamaño y más heterogéneos. En general, la pauta geográfica de los grupos es relativamente compleja, descartándose agrupaciones basadas en rasgos geográficos simples (por ejemplo, costa vs. interior).

La traslación de esta agrupación a los componentes cíclicos muestra un solapamiento muy moderado, de forma que series con comportamientos estacionales diferentes no poseen, por lo general, un patrón cíclico igualmente diferente. Reforzando el resultado anterior, el comportamiento cíclico a nivel provincial muestra un elevado grado de comunalidad, aportando la información estacional un elemento diferenciador marginal. Estos rasgos son bastante robustos frente a la muestra y al cálculo de la matriz de distancia (Frutos y Quilis, 2000).

A nivel teórico, la evidencia empírica obtenida resulta consistente con una visión del proceso de optimización de los agentes económicos en el que la estacionalidad aparece como una restricción exógena, de manera que los agentes la descuentan de forma sistemática a la hora de tomar sus decisiones.

En este sentido, la práctica habitual de analistas del ciclo y coyunturistas, consistente en trabajar con series ajustadas de estacionalidad e ignorando por tanto la pauta estacional, puede considerarse apropiada.

Desde un punto de vista metodológico, la combinación secuencial de técnicas muy diversas ha ofrecido un resultado coherente, permitiendo la aplicación de métodos esencialmente estáticos en un contexto de series temporales múltiples estacionales. En este sentido, el recurso a métodos univariantes de extracción de señales ha sido esencial, al permitir un adecuado tratamiento de la variedad subyacente en los datos y reducir sustancialmente la complejidad del análisis. Estos métodos, muy ligados al ajuste estacional, ofrecen una base sólida y muy consolidada para mitigar la "maldición de la dimension" inherente a la modelización de un vector de series temporales, especialmente si son estacionales.

Por otra parte, el procedimiento utilizado para el cálculo de las matrices de distancia (DTW) posibilita un enlace muy adecuado entre la primera etapa (univariante, basada en modelos dinámicos) y la tercera (multivariante, centrada en técnicas estáticas). El método DTW, al tener en cuenta la naturaleza dinámica de los objetos cuya distancia se desea medir y basarse en la semejanza entre sus perfiles, es especialmente apropiado para el contexto de agrupación de componentes subyacentes de un vector de series temporales. Adicionalmente, el método DTW debidamente modificado puede operar sobre vectores de series de dimensión muy elevada (Rakthanmanon *et al.,* 2012).

La escalabilidad del proceso está asegurada para las dos primeras etapas pero, para la tercera, requiere utilizar como procedimiento de aglomeración uno de tipo no jerárquico. La combinación del método *k-means* con procedimientos de optimización a gran escala (por ejemplo, basados en algoritmos genéticos) es una línea prometedora.

Finalmente, este trabajo puede ser ampliado en diversas direcciones. Combinar la metodología esencialmente exploratoria de este trabajo con un enfoque confirmatorio basado en modelos factoriales dinámicos (Nieto, Peña y Saboyá, 2016) o en modelos estructurales multivariantes (Harvey y Koopman, 1997) es una de ellas.

Otro desarrollo interesante consiste en aplicar la metodología utilizada en este trabajo a series económicas diarias, cuya estructura estacional múltiple plantea importantes retos estadísticos (Cuevas, Ledo y Quilis, 2020).

Referencias

Barsky, R. B. y Miron, J. A. (1989). The seasonal cycle and the business cycle. *Journal of Political Economy,* 97(3), pp. 503–534.

Beaulieu, J., MacKie-Mason, K. y Miron, J. A. (1992). Why do countries and industries with large seasonal cycles also have large business cycles? *Quarterly Journal of Economics,* 107(2), pp. 621–656.

Bógalo, J. y Quilis, E. M. (2000). Estimación del ciclo económico mediante filtros de Butterworth. Instituto Nacional de Estadística. *Boletín Trimestral de Coyuntura,* 87, pp. 73-93.

Box, G. E. P. y Jenkins, G. M. (1976). *Time Series Analysis, forecasting and control.* Holden Day.

Caiado, J., Crato, N. y Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis,* 50, pp. 2668–2684.

Caiado, J., Crato, N. y Poncela, P. (2020). A fragmented-periodogram approach for clustering big data time series. *Advances in Data Analysis and Classification,* 14, pp. 117–146.

Cecchetti, S., Kashyap, A. y Wilcox, D. (1997). Interaction between the seasonal and business cycles in production and inventories. *American Economic Review,* 87(5), pp. 84–92.

Cuevas. A., Ledo, R. y Quilis, E. M. (2020). Nowcasting the Spanish economy using very high frequency tax data. *SSRN Working Paper.*

Diebold, F. X. (2020). On the origin(s) of the term 'Big Data'. *arXiv Working Paper.*

Everitt, B. S., Landau, S., Leese M. y Stahl D. (2011). *Cluster Analysis.* John Wiley and Sons.

Faber, V. (1994). Clustering and the continuous k-means algorithm. *Los Alamos Science,* 22, pp. 138–144.

Franses, P. H. y de Bruin, P. (2000). Seasonal adjustment and the business cycle in unemployment. *Studies in Nonlinear Dynamics & Econometrics,* 4(2), pp. 1-14.

Frutos, R. y Quilis, E. M. (2000). Estacionalidad y ciclos en las series de pernoctaciones. Instituto Nacional de Estadística. *Boletín Trimestral de Coyuntura,* 76, pp. 65-75.

Galeano, P. y Peña, D. (2000). Multivariate analysis in vector time series. *Resenhas do Instituto de Matematica e Estatstica da Universidade de Sao Paulo,* 4(4), pp. 383-403.

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics,* 31(22), pp. 3718–3720.

Geremew, M. y Gourio, F. (2018). Seasonal and business cycles of U.S. employment. Federal Reserve Bank of Chicago. *Economic Perspectives,* 3, pp. 1-28.

Guerrero, V. M. (1993). Time-series analysis supported by power transformations. *Journal of Forecasting,* 12, pp. 37–48.

Harvey, A. C. y Koopman, S. J. (1997). Multivariate structural time series models. En: C. Heij, H. Schumacher, B. Hanzon, y C. Praagman (eds.). *System Dynamics in Economic and Financial Models.* John Wiley and Sons.

Hyndman, R. J. y Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software,* 27(3), pp. 1-22.

INE (2019). *Encuesta de Ocupación Hotelera (EOH).* Metodología.

INE (2020). Un retrato de nuestros turistas. *Cifras INE,* julio.

Kassambara, A. (2017). *Practical Guide to Cluster Analysis* in R. STHDA.com.

Kassambara, A. (2020). *Factoextra R package: easy multivariate data analyses and elegant visualization.* STHDA.com.

Kolanovic, M. y Krishnamachari, R. T. (2017). *Big Data and AI Strategies.* JP Morgan, Global Quantitative & Derivatives Strategy.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. y Hornik, K. (2019). Cluster: Cluster analysis basics and extensions. R package version 2.1.0. CRAN.

Maravall, A. (1987). Descomposición de series temporales. Especificación, estimación e inferencia. *Estadística Española,* 29(114), pp. 11–69.

Montero, P. y Vilar, J. A. (2014). TSclust: an R package for time series clustering. *Journal of Statistical Software,* 62(1), pp. 1–43.

Nieto, F. H., Peña, D. y Saboyá, D. (2016). Seasonality in multivariate time series. *Statistica Sinica,* 26(4), pp. 1389-1410.

Oppenheim, A. V. y Schaffer, R. W. (1989). *Discrete Time Signal Processing.* Prentice Hall.

Peña, D., Tiao, G. C. y Tsay, R. S. (2001). *A Course in Time Series Analysis.* John Wiley and Sons.

Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of Time Series Analysis,* 11(2), pp. 153–164.

Pollock, D .S . G. (1999). *A Handbook of Time Series Analysis, Signal Processing and Dynamics.* Academic Press.

Prescott, E. C. (1986). Theory ahead of business cycle measurement. Federal Reserve Bank of Minneapolis. *Quarterly Review,* 10(4), pp. 9–22.

Proakis, J. G. y Manolakis, D. K. (2006). *Digital Signal Processing.* Pearson New International.

Quilis, E. M. (2019). *FactorLib: a Matlab library for static factor analysis.* Matlab, Central File Exchange.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J. y Keogh, E. (2012). Searching and mining trillions of time series subsequences under Dynamic Time Warping (DTW). SIGKDD, pp. 262–270.

Rani, S. y Sikka, G. (2012). Recent techniques for clustering of time series data: a survey. *International Journal of Computer Applications,* 52(15), pp. 1–9.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics,* 20, pp. 53–65.

Sakoe, H. y Chiba, S. (1978), Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, IEEE Transactions on Speech and Signal Processing,* 26(1), pp. 43–49.

Sardá-Espinosa, A. (2019) Time series Clustering in R using the dtwclust package. *The R Journal,* 11(1), pp. 22–43.

Sathi, A. (2012). *Big Data Analytics.* MC Press.

Sax, C. y Eddelbuettel, D. (2018). Seasonal adjustment by X-13ARIMA-SEATS in R. *Journal of Statistical Software,* 87(11), pp. 1-17.

Todd, R. M. (1990). Periodic linear-quadratic methods for modeling seasonality. *Journal of Economic Dynamics and Control,* 14(3–4), pp. 763–795.

U.S. Census Bureau, Time Series Research Staff (2017). X-13ARIMA-SEATS Reference Manual. U.S. Census Bureau.

Wang, X., Smith, K. y Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery,* 13, pp. 335-364.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* 58(301), pp. 236-244.

CAPÍTULO VIII

Una aplicación del análisis de series temporales funcionales a los precios horarios de la electricidad en el mercado MIBEL

Pedro Galeano*

Actualmente, muchas medidas se registran de manera prácticamente continua a lo largo del tiempo dando lugar a conjuntos de observaciones que tienen forma de funciones (curvas) relativamente suaves y que son observadas con alta frecuencia. El estudio de datos con tales características se puede realizar mediante técnicas para el análisis de series temporales funcionales, un área de la estadística que ha recibido gran atención en las dos últimas décadas. En este capítulo se realiza una aplicación del análisis de series temporales funcionales a las curvas de rendimientos intradía acumulados de los precios horarios de la electricidad de España en el Mercado Ibérico de la Electricidad (MIBEL).

Palabras clave: análisis de datos funcionales, bandas de predicción, precios horarios de la electricidad, predicción, series temporales funcionales.

* El autor de este capítulo agradece el apoyo financiero de la Agencia Estatal de Investigación (PID2019-108311GB-I00/AEI/10.13039/501100011033).

1. INTRODUCCIÓN

La gran mayoría de métodos y procedimientos estadísticos tradicionales para el análisis de series temporales fueron diseñados para series con decenas, centenas o incluso algunos miles de observaciones. Estos métodos y procedimientos tratan de compensar la relativa falta de información con ciertos supuestos que permiten obtener predicciones relativamente fiables, lo que ha sido de gran ayuda para el desarrollo del conocimiento en un gran número de disciplinas. Si bien es cierto que el uso de modelos probabilísticos hace que el análisis de series temporales se pueda considerar un subcampo de la estadística, una parte fundamental de su desarrollo ha sido llevada a cabo por investigadores y científicos en otras áreas, muy especialmente en economía. Por ejemplo, los modelos de heterocedasticidad condicional, como el modelo GARCH de Engle (1982) y Bollerslev (1986) y el modelo de volatilidad estocástica de Taylor (1986), entre muchos otros modelos que permiten analizar la varianza condicionada de las series temporales, fueron desarrollados para entender y predecir la volatilidad de rendimientos financieros. También, las técnicas de cointegración de Granger (1981) y Engle y Granger (1987), entre otros, que permiten buscar correlaciones entre series temporales que ayuden a mejorar sus predicciones, fueron desarrolladas para entender la relación entre variables económicas, como el consumo y los ingresos, a largo plazo. Otro ejemplo son los contrastes de raíces unitarias, como los propuestos por Dickey y Fuller (1979), Sargan y Bhargava (1983), Said y Dickey (1984) y Kwiatkowski *et al.* (1992), que permiten determinar cuándo una serie temporal es estacionaria o no, fueron desarrollados en el contexto de variables económicas, como el PIB.

Durante los últimos veinte años han aparecido nuevas fuentes de información, como internet, los teléfonos inteligentes, las redes sociales y los sensores automáticos que producen grandes bases de datos estructuradas y no estructuradas con tamaños masivos y que pueden tener forma de textos, imágenes, vídeos y/o sonidos (véase Galeano y Peña, 2019). En particular, la monitorización hace que actualmente muchas medidas se registren prácticamente de manera continua a lo largo del tiempo, generando series temporales de gran longitud y con características especiales propias, como puede ser la presencia de ruido en forma de fluctuaciones que distorsionan las tendencias de la serie. Medidas de este estilo pueden ser, por ejemplo: (1) señales vitales humanas como la temperatura corporal, la presión arterial y las frecuencias cardíaca y respiratoria; (2) factores medioambientales como la temperatura, la velocidad del viento, la precipitación, y los niveles de contaminación de determinados contaminantes, como el monóxido de carbono o el óxido de azufre; y (3) el precio de activos o materias primas que se negocian en mercados financieros en tiempo real, como el precio de las acciones de una empresa o el precio de la energía eléctrica.

Debido a la gran longitud que pueden tener este tipo de series temporales y a que se pueden considerar varias de estas series de manera simultánea, los métodos y procedimientos más habituales para el análisis de la dependencia transversal y temporal pueden ser inadecuados y/o ineficientes. Por ello, están apareciendo nuevas metodologías entre las que se encuentra el análisis de series temporales funcionales. Este tipo de análisis surge cuando una serie temporal que se observa muy frecuentemente durante un extenso período de tiempo, se puede dividir en subseries de la misma longitud definidas en intervalos de tiempo conse-

cutivos definidos periódicamente, que pueden sufrir algún tipo de transformación posterior para alcanzar alguna propiedad como puede ser la estabilidad. Por ejemplo, una variable que se observa cada minuto durante varios meses o años se puede separar en subseries definidas en días o semanas consecutivas, dependiendo de sus características, que a su vez se pueden transformar en subseries estables o que comiencen con el mismo valor. Las subseries obtenidas de esta manera suelen mostrar dos características fundamentales. En primer lugar, tienden a tener un patrón común y formas relativamente suaves por lo que pueden ser estudiadas con las herramientas del análisis de datos funcionales (FDA, es su acrónimo en inglés). El FDA consiste en una colección de métodos y procedimientos para el análisis estadístico de muestras de funciones y es una de las áreas que mayor interés ha suscitado en la comunidad estadística en los últimos años. En la práctica, y aunque la frecuencia de toma de datos pueda ser muy alta, los procesos continuos se observan en un número discreto de puntos. Sin embargo, las técnicas de suavizado permiten reproducir las funciones que han sido observadas parcialmente obteniendo una única serie temporal funcional compuesta por una sucesión de funciones observadas durante el mismo período de tiempo, por ejemplo, en días, semanas, o meses sucesivos. Esta serie temporal funcional tiene una serie de características con dimensión subyacente infinita. Esto puede parecer problemático ya que es bien sabido que la maldición de la dimensión conlleva importantes problemas estadísticos. Sin embargo, los métodos de FDA permiten extraer información de la riqueza estructural inherente a los datos funcionales. En segundo lugar, las subseries obtenidas suelen ser dependientes debido a la manera en la que son definidas. Los procedimientos para el análisis de series temporales funcionales tratan de describir y entender dicha dependencia que puede ser utilizada, por ejemplo, para realizar predicciones de funciones futuras con bandas de predicción asociadas. De esta manera, las predicciones obtenidas corresponden a cada uno de los períodos en que se ha dividido la serie original. Por ejemplo, para subseries formadas por las observaciones correspondientes a los minutos de todo un día, la predicción funcional a un paso corresponderá a una predicción de todos los momentos del día siguiente al último observado, y no solamente para los minutos del día siguiente.

Existe una literatura creciente sobre el análisis de series temporales funcionales que, de alguna manera, extiende el análisis de series temporales univariantes y multivariantes al campo funcional. Aquí se incluyen: (1) las técnicas de suavizado propuestas por Ramsay y Silverman (2005), que permiten eliminar el ruido y suavizar los datos, lo que es beneficioso siempre que el proceso subyacente sea la principal preocupación; (2) los procedimientos de predicción basados en componentes principales funcionales propuestos por Hyndman y Ullah (2007), Hyndman y Shang (2009) y Aue, Norinho y Hörmann (2015), entre otros, que son extensiones de los procedimientos de reducción de la dimensión en el análisis de series temporales; (3) los modelos autorregresivos y de media móvil funcionales propuestos por Bosq (2000), Bosq y Blanke (2007), Klepsch, Klüppelberg y Wei (2017) y Li, Robinson y Shang (2020), entre otros, que son extensiones de los modelos autorregresivos y de media móvil del análisis de series temporales univariantes y multivariantes; (4) los procedimientos de estimación y predicción no paramétricos de Besse, Cardot y Stephenson (2000) y Antoniadis, Paparoditis y Sapatinas (2006), entre otros; (5) los modelos de heterocedasticidad condicional funcional propuestos por Hörmann, Horváth, y Reeder (2013) y Aue, Horváth y Pellatt (2017), entre otros; y (6) los modelos factoriales dinámicos funcionales propuestos por Hays,

Shen y Huang (2012), Kowal, Matteson y Ruppert (2012) y Kokoszka, Miao y Zhang (2015), entre otros. Estos y otros procedimientos han sido utilizados en un gran número de aplicaciones. Por ejemplo, Ramsay y Silverman (2005) detectaron diferentes tendencias estacionales y a largo plazo en el índice de fabricación de productos perecederos en USA mediante simples técnicas de suavizado. Hyndman y Ullah (2007) y Hyndman y Shang (2009), entre muchos otros, utilizaron los métodos de predicción basados en componentes principales funcionales para predecir curvas de mortalidad anual por edad de hombres y mujeres franceses, y para predecir tasas de fertilidad anual en Australia en función de la edad. Shen (2009) y Aue, Norinho y Hörmann (2015) también utilizaron componentes principales funcionales para predecir tasas de llegada de llamadas diarias a un servicio de atención al cliente de un banco y concentraciones de polución en Graz, Austria, respectivamente. Klepsch, Klüppelberg y Wei (2017) y Li, Robinson y Shang (2020) utilizaron modelos ARMA funcionales para predecir datos de tráfico en Alemania y tasas de fertilidad en varios países, respectivamente. Besse, Cardot y Stephenson (2000) y Antoniadis, Paparoditis y Sapatinas (2006) utilizaron métodos no paramétricos para predecir el efecto anual del fenómeno de *El Niño* y tasas de audiencias televisivas, respectivamente. Hörmann, Horváth, y Reeder (2013) y Aue, Horváth y Pellatt (2017) analizaron rendimientos diarios mediante modelos de heterocedasticidad condicional funcional. Por último, Hays, Shen y Huang (2012) y Kowal, Matteson y Ruppert (2012) utilizaron modelos factoriales dinámicos funcionales para predecir curvas de rendimientos financieros y señales cerebrales, respectivamente.

El objetivo principal de este capítulo es el de presentar una aplicación del análisis de series temporales funcionales de las curvas de rendimientos intradía acumulados de los precios horarios de la electricidad para España en el Mercado Ibérico de la Electricidad (MIBEL). Para ello, la serie temporal de precios horarios, observada durante un período extenso de años, se transforma en una sucesión de subseries definidas mediante una estandarización de los precios en cada uno de los días de negociación. A continuación, cada subserie se transforma en un dato funcional genuino mediante técnicas de suavizado y todas las subseries suavizadas en forma de función dan lugar a una única serie temporal funcional. Mediante técnicas del análisis funcional de datos, se concluirá que dicha serie temporal funcional es predecible por lo que se obtendran predicciones a un día vista de las curvas, junto con sus respectivas bandas de predicción. Para realizar los cálculos y gráficos mostrados en este artículo se han utilizado fundamentalmente las librerías fda y ftsa del *software* estadístico R. En el capítulo, se informará de las funciones utilizadas especificamente en cada uno de los cálculos realizados.

El resto de este capítulo se estructura como sigue. La sección segunda presenta la serie de precios horarios de la electricidad para España en el Mercado Ibérico de la Electricidad (MIBEL) y cómo, a partir de esta serie, se obtienen sus curvas de rendimientos intradía acumulados mediante técnicas de suavizado, dando lugar a una serie temporal funcional. La sección tercera realiza un análisis descriptivo de las curvas de rendimiento y comprueba la estabilidad temporal de las mismas. La sección cuarta presenta los componentes principales funcionales de las curvas de rendimiento, que no solo son una herramienta muy útil para obtener las principales fuentes de variación de las curvas, si no que permiten obtener predicciones de las curvas futuras. De hecho, la sección quinta realiza predicción a un día

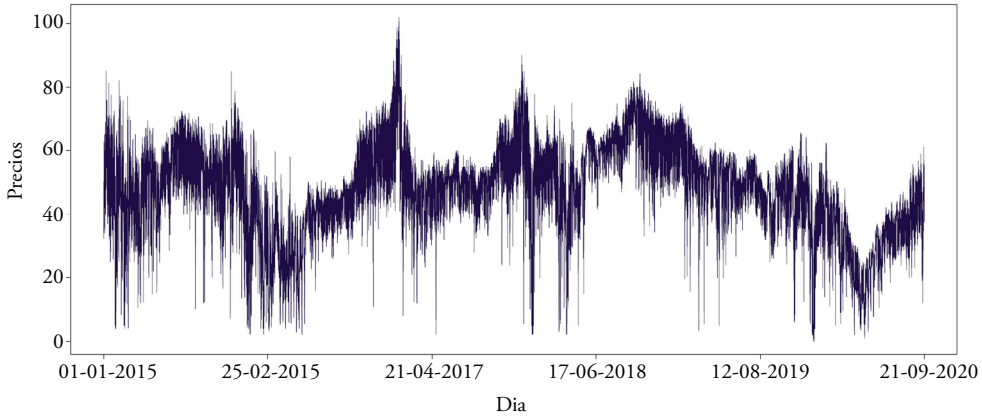
vista de las curvas observadas mediante componentes principales funcionales e introduce un procedimiento simple para la obtención de bandas de predicción, que son el equivalente de los intervalos y regiones de predicción en series temporales univariantes y multivariantes. Finalmente, la sección sexta presenta algunas conclusiones.

2. CURVAS DE RENDIMIENTOS INTRADÍA ACUMULADOS

La figura 1 muestra los precios de cierre cada hora de la electricidad para España en el Mercado Ibérico de la Electricidad (MIBEL) en los 2100 días comprendidos entre el 1 de

Figura 1.

Precios horarios de la electricidad para España en el Mercado Ibérico de la Electricidad (MIBEL)



Fuente: Elaboración propia.

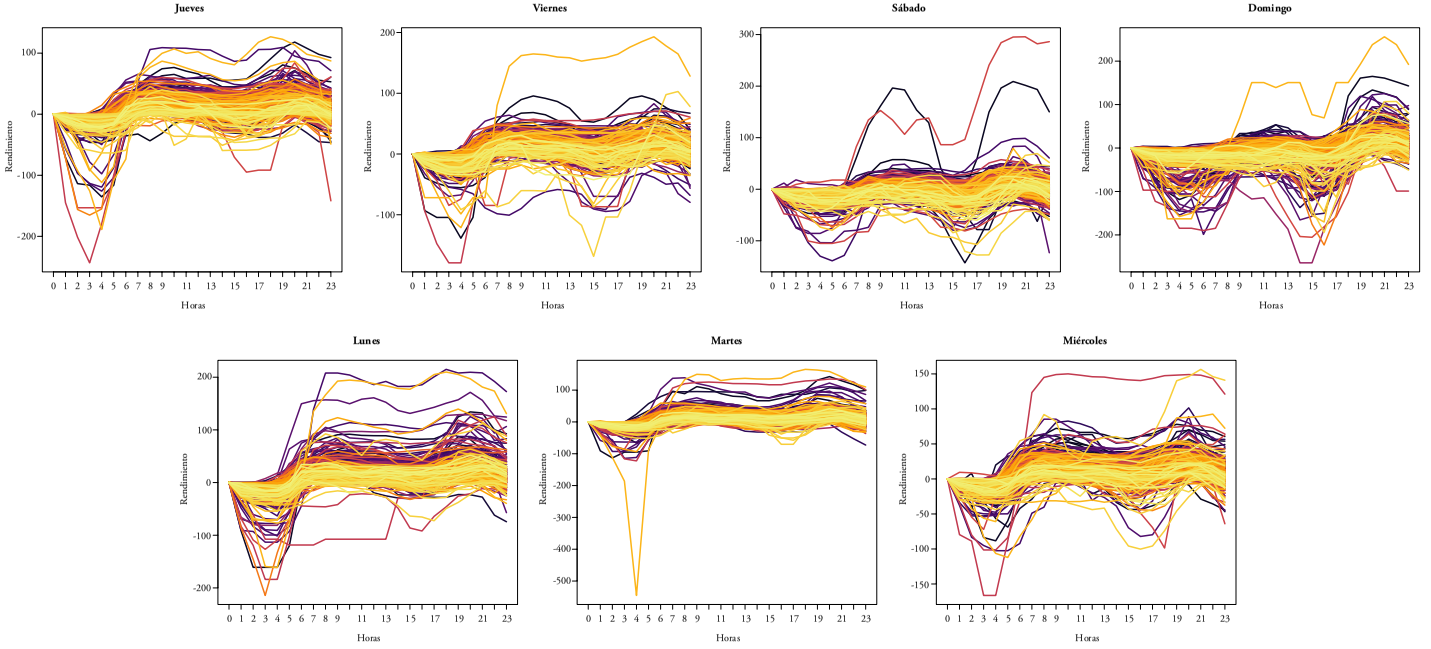
enero de 2015 y el 30 de septiembre de 2020, que han sido obtenidos de la página web de Red Eléctrica Española, https://www.esios.ree.es/es. Dado que el intervalo temporal entre observaciones es de una hora, la serie temporal de precios tiene 2.100 × 24 = 50.400 observaciones. Como se puede observar, la serie es claramente no estacionaria con los típicos crecimientos y decrecimientos habituales en series de precios y con la presencia de algunos valores extremos. Tal como hacen Gabrys, Horváth y Kokoszka (2010) con series de precios de gran longitud como esta, se definen los rendimientos intradía acumulados como:

$$r_i(t_j) = 100 \times [\ln p_i(t_j) - \ln p_i(t_1)], \qquad [1]$$

donde $p_i(t_j)$ es el precio en la hora t_j del i-ésimo día de negociación, para $i = 1, ..., 2100$ días, y t_1 y t_{24} son las 00:00 y las 23:00 horas del día, respectivamente. Dado que,

Figura 2.

Series de rendimientos intradía acumulados de los precios horarios de la electricidad desde el 1 de enero de 2015 al 30 de septiembre de 2020 divididas en los siete días de la semana



Fuente: Elaboración propia.

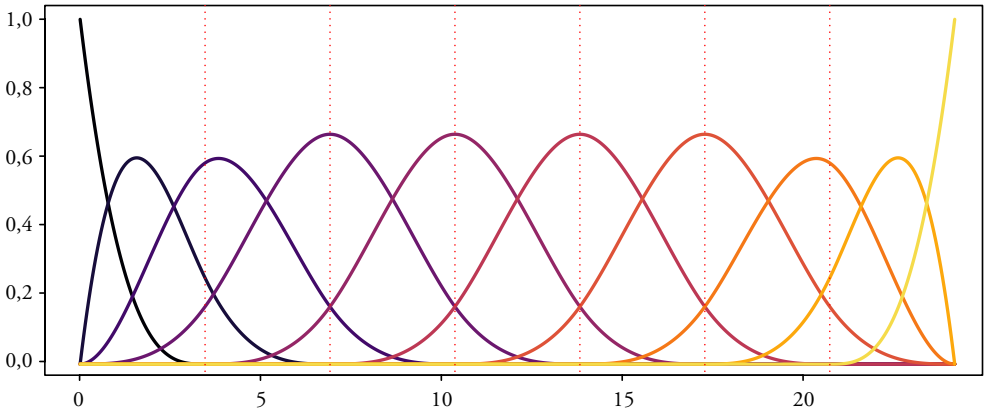
$$r_i(t_j) \simeq 100 \times \frac{p_i(t_j) - p_i(t_1)}{p_i(t_1)},$$

y teniendo en cuenta que $p_i(t_1)$ es un valor fijo para cada día i, los rendimientos intradía acumulados se pueden ver como una sucesión de series temporales diarias de 24 observaciones con el mismo origen, 0, y que resultan ser una especie de estandarización de los precios en cada uno de los días de negociación, ya que $p_i(\cdot)$ y $r_i(\cdot)$ tienen diferentes escalas y orígenes.

En particular, las 2.100 series de 24 rendimientos intradía acumulados obtenidas a partir de la serie de precios de la figura 1, se muestran en la figura 2. Las series se han dividido en los siete días de la semana, empezando por el jueves al ser el día correspondiente al 1 de enero de 2015, y se han dibujado con una escala de color de tal manera que las series de pasado más lejano tienen color obscuro y las series de pasado más cercano tienen color más claro. Nótese que muchas series de color claro ocultan a series de color más obscuros simplemente porque la función plot de R dibuja la curva del primer día y sucesivamente va añadiendo las curvas correspondientes al resto de días. Aunque es difícil extraer conclusiones específicas a partir de los gráficos, parece apreciarse que la mayoría de rendimientos intradía crecen en las primeras horas de la mañana y últimas de la tarde y decrecen en el resto de momentos del día. Este patrón común parece estar presente en todos los días de la semana. Además, existe un número determinado de series que parece tener un comportamiento diferente posiblemente debido a ser días festivos o de índole especial, e incluso algunas observaciones concretas de algunas series parecen mostrar también comportamientos extremos o atípicos. De hecho, la figura 2 sugiere que los rendimientos están afectados por fluctuaciones de precios o correcciones a corto plazo que pueden distorsionar las tendencias comunes.

Figura 3.

Funciones de una base de B-splines en el intervalo [0,24)



Fuente: Elaboración propia.

El proceso de suavizado que se describe a continuación permite, por un lado, suprimir la influencia del ruido más obvio en las series de rendimientos intradía y, por otro lado, transformar dichas series en genuinos datos funcionales. Para ello, nótese que si bien se disponen únicamente de los valores en [1], los rendimientos existen en cualquier momento del día. Por lo tanto, con los dos objetivos previos en mente, se supone que los rendimientos intradía acumulados en [1] se pueden escribir como

$$r_i(t_j) = \theta_i(t_j) + e_{ij},$$

donde θ_i es una función continua y suave definida en todo tiempo t $\in$ [0,24) y e_{ij} es un término de ruido. El conjunto de funciones:

$$\{\theta_i(t) : \ t \in [0,24), i = 1,...,2100\}, \qquad [2]$$

recibe el nombre de curvas de rendimientos intradía acumulados (CIDRs, es su acrónimo en inglés). Con el propósito de distinguir entre una CIDR como objeto funcional y el valor que toma dicha CIDR en un momento específico del día, se denota por θ_i a la i-ésima CIDR como objeto funcional y por $\theta_i(t)$ al valor puntual de la i-ésima CIDR en un tiempo t $\in$ [0,24) específico. Ahora, el paso clave para obtener las CIDR en [2] a partir de las series de rendimientos en [1] es expresar $\theta_i(t)$ mediante una expansión de funciones base, es decir, mediante una combinación lineal de funciones fijas determinadas que permiten aproximar cualquier función continua y suave. Los sistemas de funciones base más conocidos son la base de Fourier y las bases de B-splines (ver, por ejemplo, los capítulos 3 a 6 de Ramsay y Silverman, 2005). Por un lado, la base de Fourier es una sucesión de senos y cosenos con frecuencias enteras y está diseñada para el análisis de funciones periódicas, es decir, funciones que tienen el mismo valor al principio y al final del intervalo de observación. Por otro lado, las bases de B-splines están formadas por sucesiones de polinomios definidos en subintervalos consecutivos y están diseñadas para el análisis de cualquier tipo de función continua. Por lo tanto, para obtener las CIDRs se utiliza una base de B-splines con las 10 funciones definidas en el intervalo [0,24) que aparecen en la figura 3 y que permiten escribir $\theta_i(t)$ como sigue:

$$\theta_i(t) = \sum_{m=1}^{10} c_{im}\beta_m(t) = c_i'\beta(t), \qquad i = 1,\ldots,2100, \qquad [3]$$

donde $c_i = (c_{i1},...,c_{i10})'$ es el vector de coeficientes que minimizan la suma de cuadrados dada por,

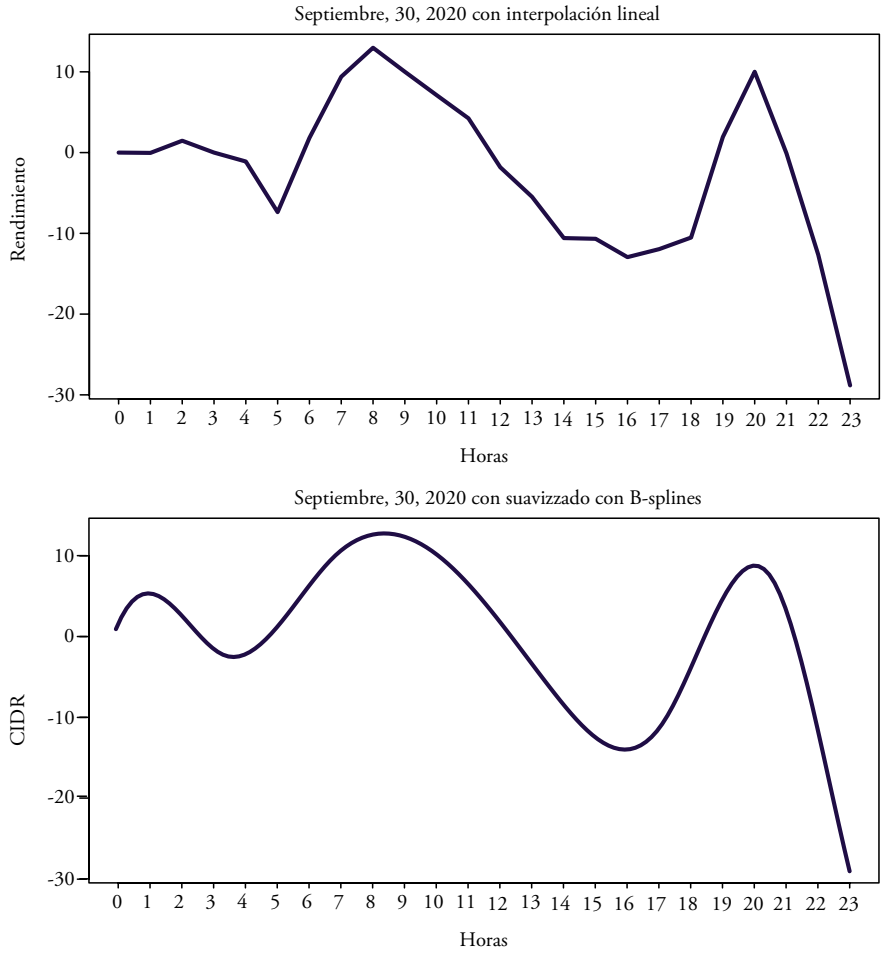
$$SC(c_{i1},\ldots,c_{i10}) = \sum_{j=1}^{24}\left(r_i(t_j) - \sum_{m=1}^{10} c_{im}\beta_m(t_j)\right)^2, \qquad i = 1,\ldots,2100,$$

y $\beta(t) = (\beta_1(t),..., \beta_{10}(t))'$ es el vector de funciones base evaluadas en tiempo t $\in$ [0,24), respectivamente. Por ejemplo, la parte derecha de la figura 4 muestra los 24 rendimientos intradía acumulados correspondientes al 30 de septiembre de 2020 interpolados linealmente, mientras que la parte izquierda muestra la CIDR asociada suavizada con la base de 10 B-splines

de la figura 3. Como se puede apreciar, la CIDR preserva la forma de los rendimientos horarios observados. Para realizar el suavizado se han utilizado las funciones create.bspline.basis y smooth.basis de la librería fda de R.

Figura 4.

CIDR para los precios horarios de la electricidad del 30 de septiembre de 2020



Fuente: Elaboración propia.

Es muy importante resaltar el cambio de enfoque que supone el disponer de las CIDRs suaves en [3] frente a disponer de la serie temporal de precios original o de las series de rendimientos discretos en [1]. Nótese que inicialmente se ha transformado una serie temporal de 50.400 observaciones en las 2.100 series temporales de 24 observaciones en [1], que a su vez se han transformado en las 2.100 funciones suaves en [2] escritas en términos de los

B-splines como aparecen en [3]. Por lo tanto, de aquí en adelante el objeto de estudio son estas 2.100 funciones definidas en todo momento del día, y no los rendimientos horarios que han dado lugar a dichas funciones. Esto supone un cambio substancial, ya que, por ejemplo, en lugar de predecir los rendimientos acumulados en cada hora en punto del 1 de octubre de 2020, el objetivo es predecir la CIDR completa correspondiente al 1 de octubre de 2020, lo que permitiría conocer la predicción en cualquier momento del día. Para ello, como se detalla en la sección quinta, se hace uso de la expansión en [3], del vector de coeficientes $c_i = (c_{i1},...,c_{i10})'$, para $i = 1,...,2100$, y del vector de funciones base evaluadas en tiempo $t \in [0,24)$, $\beta(t) = (\beta_1(t),..., \beta_{10}(t))'$.

Las CIDRs presentadas en esta sección dan lugar a muchas preguntas. Por ejemplo, es razonable preguntarse si las CIDRs de los precios de la electricidad son estacionarias bajo el punto de vista funcional. Aquí, como se describirá en la sección tercera, la estacionariedad funcional se refiere a la estabilidad de las CIDRs durante todo el período de observación. Conocer si las CIDRs son estacionarias o no es importante ya que ayuda a entender la dinámica de las funciones y a seleccionar un buen procedimiento para obtener predicciones de CIDR futuras. También parece adecuado preguntarse cuáles son las principales fuentes de variabilidad de las CIDRs que puedan ayudar a explicar las tendencias locales que parecen observarse en las figuras 2 y 4. Por último, también parece adecuado preguntarse si las CIDRs de los precios de la electricidad son funciones predecibles, es decir, si es posible predecir la CIDR de un día de negociación cualquiera con las CIDRs de los días de negociación previos. Para responder a estas y otras preguntas se realiza un análisis descriptivo funcional de las CIDRs en la sección tercera y se presentan los componentes principales funcionales en la sección cuarta.

3. ANÁLISIS DESCRIPTIVO FUNCIONAL DE LAS CURVAS DE RENDIMIENTOS INTRADÍA ACUMULADOS

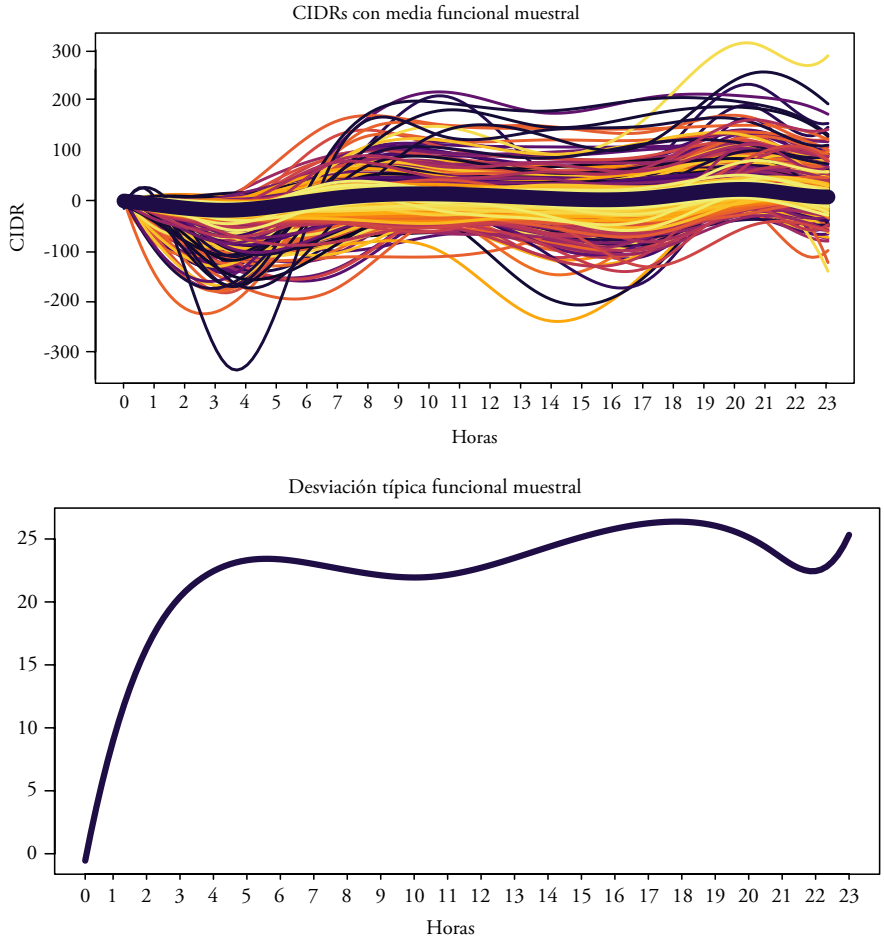
Las 2.100 CIDRs en [2] escritas mediante las expansiones en [3] forman una serie temporal funcional. Cada CIDR es una observación de la serie temporal funcional y se espera que las CIDRs de diferentes días tengan algún tipo de dependencia. Esta sección presenta un breve análisis descriptivo de la serie de CIDRs con el objetivo de entender sus principales características. Por sencillez en la exposición, dentro de lo posible se evita entrar en detalles técnicos complejos que los lectores interesados pueden consultar en referencias más generales sobre datos y series temporales funcionales, como pueden ser Ramsay y Silverman (2005), Ferraty y Vieu (2006), Hörmann y Kokoszka (2010), Horváth y Kokoszka (2012), Hsing y Eubank (2015), Srivastava y Klassen (2016) o Kokoszka y Reimherr (2017), entre otros.

En primer lugar, es importante conocer si las CIDRs son estacionarias bajo el punto de vista funcional. Aquí es necesario hacer hincapié que cuando se habla de estacionariedad desde un punto de vista funcional se refiere a la estabilidad entre las observaciones funcionales durante todo el período de observación y no a la estabilidad temporal de las observaciones puntuales, como es común en el análisis de series temporales univariantes y multivariantes.

Por ejemplo, tanto la figura 2 como la parte derecha de la figura 4 parecen indicar que las 24 observaciones discretas diarias tienen un comportamiento no estacionario. Sin embargo, la figura 2 parece indicar que las tendencias locales diarias parecen ser comunes, lo que sugiere que las CIDRs, que se muestran en el gráfico de la derecha de la figura 5, tienen un comportamiento estacionario. La estacionariedad de una serie temporal funcional se puede estudiar mediante un contraste de hipótesis propuesto por Horváth, Kokoszka y Rice (2014). La idea fundamental del estadístico de dicho contraste es comparar las sumas parciales de las funciones θ_1 ...,θ_{2100} con la suma total de todas las funciones, lo que permite conocer la

Figura 5.

Media funcional muestral y desviación típica funcional muestral para las CIDRs del precio de la electricidad



Fuente: Elaboración propia.

evolución de la serie temporal funcional. Bajo la hipótesis de estacionariedad, las diferencias entre las sumas parciales y la suma total no deben ser grandes, lo que sí debería ocurrir bajo la hipótesis alternativa de no estacionariedad. El estadístico del contraste proporciona un valor que cuantifica dichas diferencias que, en el caso concreto de las CIDRs, da lugar a un *p*-valor de 0.2347, lo que sugiere que no hay evidencia suficiente para rechazar la hipótesis nula de estacionariedad del conjunto de CIDRs. Este *p*-valor se ha obtenido gracias a la función T_stationary de la librería ftsa de R.

La condición de estacionariedad permite obtener un amplio conjunto de características de interés de las CIDRs. Las dos medidas más sencillas son la media funcional muestral, dada por:

$$\bar{\theta}(t) = \frac{1}{2100} \sum_{i=1}^{2100} \theta_i(t), \qquad t \in [0,24) \qquad [4]$$

que proporciona el promedio de los valores de las CIDRs en cualquier momento t ∈ [0,24), y la varianza funcional muestral dada por:

$$\hat{\sigma}^2(t) = \frac{1}{2099} \sum_{i=1}^{2100} \left(\theta_i(t) - \bar{\theta}(t)\right)^2, \qquad t \in [0,24) \qquad [5]$$

que proporciona la varianza de los valores de las CIDRs en cualquier tiempo t ∈ [0,24). Estas medidas se pueden calcular muy fácilmente tras substituir $\theta_1,...,\theta_{2100}$ en las ecuaciones [4] y [5] con sus expansiones en términos de $\beta_1,...,\beta_{10}$ dadas en [3], dando lugar a:

$$\bar{\theta}(t) = \bar{c}'\beta(t), \qquad t \in [0,24) \qquad [6]$$

y

$$\hat{\sigma}^2(t) = \beta(t)' S_0 \beta(t), \qquad t \in [0,24) \qquad [7]$$

donde $\bar{c} = \frac{1}{2100}\Sigma_{i=1}^{2100} c_i$ y $S_0 = \frac{1}{2099}\Sigma_{i=1}^{2100}\left(c_i - \bar{c}\right)\left(c_i - \bar{c}\right)'$, respectivamente, es decir, el vector de medias muestral y la matriz de covarianzas muestral de los vectores de coeficientes $c_1,...,c_{2100}$. El gráfico de la derecha de la figura 5 muestra las 2.100 CIDRs junto con su media funcional muestral. Más concretamente, la media funcional muestral es la línea más gruesa que parece resumir razonablemente bien el comportamiento más común de las curvas, creciendo en las primeras horas de la mañana y últimas de la tarde y decreciendo en el resto de momentos del día. Por otro lado, el gráfico de la izquierda de la figura 5 muestra la desviación típica muestral de las CIDRs, es decir, la raíz cuadrada de la varianza funcional muestral en cada tiempo t ∈ [0,24). Como se puede observar, la variabilidad alcanza máximos alrededor de las 5:00 horas de la mañana, las 18:00 horas de la tarde y a última hora del día. El cálculo de la media funcional muestral y la varianza funcional muestral se ha realizado con las funciones mean.fd y var.fd, respectivamente, de la librería fda de R.

Más información sobre la relación de las CIDRs en diferentes momentos del día se puede obtener a partir de la covarianza funcional muestral dada por

$$\hat{\gamma}_0(t,s) = \frac{1}{2099}\sum_{i=1}^{2100}\left(\theta_i(t)-\bar{\theta}(t)\right)\left(\theta_i(s)-\bar{\theta}(s)\right), \qquad t,s\in[0,24) \qquad [8]$$

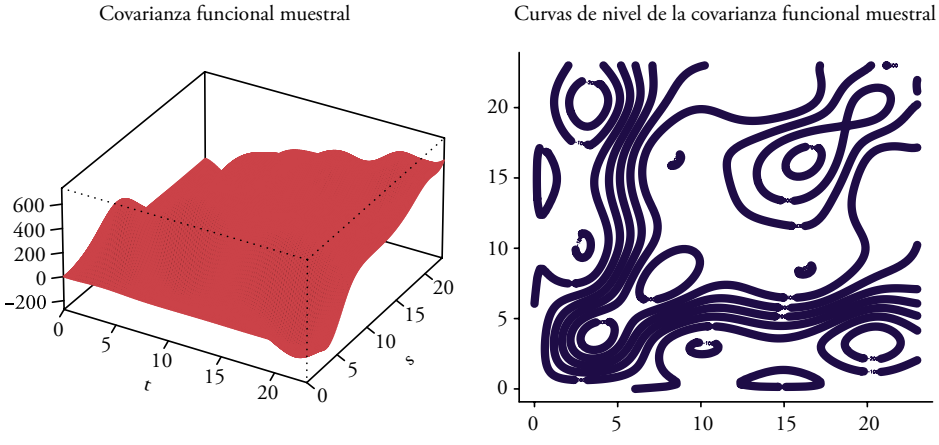
que proporciona la covarianza existente entre los valores que toman las CIDRs en dos tiempos cualesquiera *t, s* ∈ [0,24). Valores altos de $\hat{\gamma}_0(t,s)$ indican que los valores de las CIDR en los tiempos *t* y *s* tienden a estar simultáneamente por encima o por debajo de sus valores medios en dichos tiempos. En particular, nótese que $\hat{\gamma}_0(t,t) = \hat{\sigma}^2(t)$. Tras substituir la expansión [3] y la expresión [6] en la expresión [8], es fácil comprobar que:

$$\hat{\gamma}_0(t,s) = \beta(t)' S_0 \beta(s), \qquad t,s\in[0,24)$$

lo que permite un cálculo muy rápido del valor de $\hat{\gamma}_0(t,s)$ para cualquier par de tiempos *t, s* ∈ [0,24). El gráfico de la derecha de la figura 6 muestra la covarianza funcional muestral de las CIDRs en el cuadrado [0,24) × [0,24), mientras que el gráfico de la izquierda de la figura 6 muestra las curvas de nivel asociadas a dichas covarianzas. Como se puede observar, la mayoría de las covarianzas son positivas. Si se excluyen las varianzas, las mayores covarianzas se localizan entre las 9:00 horas de la mañana y las 16:00 horas de la tarde. Además, parece haber covarianzas negativas entre las horas previas y posteriores al cambio de día. El cálculo de la covarianza funcional muestral se ha realizado con la función var.fd de la librería fda de R, mientras que el cálculo de las curvas de nivel asociadas se ha realizado con la función contour de la librería graphics de R.

Figura 6.

Covarianza funcional muestral para las CIDRs del precio de la electricidad



Fuente: Elaboración propia.

Por último, es de especial interés conocer la relación que existe entre las CIDRs de días diferentes. Se pueden utilizar dos medidas para ello. La primera de ellas es la autocovarianza funcional muestral dada por:

$$\hat{\gamma}_l(t,s) = \frac{1}{2099} \sum_{i=1}^{2100-l} \left(\theta_{i+l}(t) - \overline{\theta}(t)\right)\left(\theta_i(s) - \overline{\theta}(s)\right), \qquad t,s \in [0,24) \qquad [9]$$

para l = 1,2,..., que proporciona la autocovarianza existente entre los valores que toman las CIDRs separadas l días en dos tiempos cualesquiera t, $s \in$ [0,24). En este caso, valores altos de $\hat{\gamma}_l(t,s)$ implican que los valores de las CIDRs separadas l días en los tiempos t y s tienden a estar simultáneamente por encima o por debajo de sus valores medios en dichos tiempos. Tras substituir la expansión [3] y la expresión [6] en la expresión [9], es fácil comprobar que:

$$\hat{\gamma}_l(t,s) = \beta(t)' S_l \beta(s), \qquad t,s \in [0,24) \qquad [10]$$

donde $S_l = \frac{1}{2099} \Sigma_{i=1}^{2100-t} \left(c_{i+l} - \overline{c}\right)\left(c_i - \overline{c}\right)'$ es la matriz de autocovarianzas muestral de orden l de los vectores de coeficientes $c_1,...,c_{2100}$. Si bien analizar las autocovarianzas muestrales puede resultar muy informativo, puede ser a su vez complicado determinar hasta que punto existe una dependencia fuerte o débil entre las CIDRs en días diferentes. Por ello, es conveniente definir autocorrelaciones funcionales muestrales a través de las autocovarianzas funcionales muestrales en [10]. La opción más popular es la propuesta por Horváth, Rice y Whipple (2016), quienes definieron el coeficiente de autocorrelación funcional dado por:

$$\hat{r}_l = \frac{\left(\int_0^{24}\int_0^{24} \hat{\gamma}_l(t,s)^2 \, ds dt\right)^{1/2}}{\int_0^{24} \hat{\sigma}^2(t)\, dt}, \qquad l = 1,2,\ldots \qquad [11]$$

tal que $\hat{r}_l \in [0,1]$. Como se puede ver, el coeficiente $\hat{r}_l$ en [11] es la ratio entre las magnitudes de $\hat{\gamma}_l$ y $\hat{\sigma}^2$. Por lo tanto, $\hat{r}_l$ no está definido para un par de tiempos (t, s), como la autocovarianza funcional en [9], si no que se trata de un valor entre 0 y 1 libre de escala y que mide la autocorrelación entre las CIDRs para el retardo l. Como en el caso de series temporales univariantes, valores de $\hat{r}_l$ cercanos a 1 indican autocorrelaciones fuertes y valores de r_l cercanos a 0 indican autocorrelaciones débiles. El conjunto de valores $\hat{r}_1, \hat{r}_2, ...$, forma la llamada función de autocorrelación funcional, que es el equivalente a la función de autocorrelación en el análisis de series temporales univariantes, y que es útil para identificar qué retardos de la serie de CIDRs están asociados a autocorrelaciones elevadas. Substituyendo en [11] los valores de $\hat{\sigma}^2(t)$ y $\hat{\gamma}_l(t,s)$ en términos del vector $\beta(t)$ y las matrices S_0 y S_1 en [7] y [10], respectivamente, es relativamente sencillo comprobar que:

$$\hat{r}_l = \frac{Tr\left(S_l B S_l' B\right)^{1/2}}{Tr\left(S_0 B\right)}, \qquad [12]$$

donde $B = \int_0^{24} \beta(t)\beta(t)' \, dt$ es la matriz de dimensión 10 × 10 con elementos:

$$B_{mk} = \int_0^{24} \beta_m(t)\beta_k(t)\, dt = \langle \beta_m, \beta_k \rangle, \qquad m,k = 1,\ldots,10 \qquad [13]$$

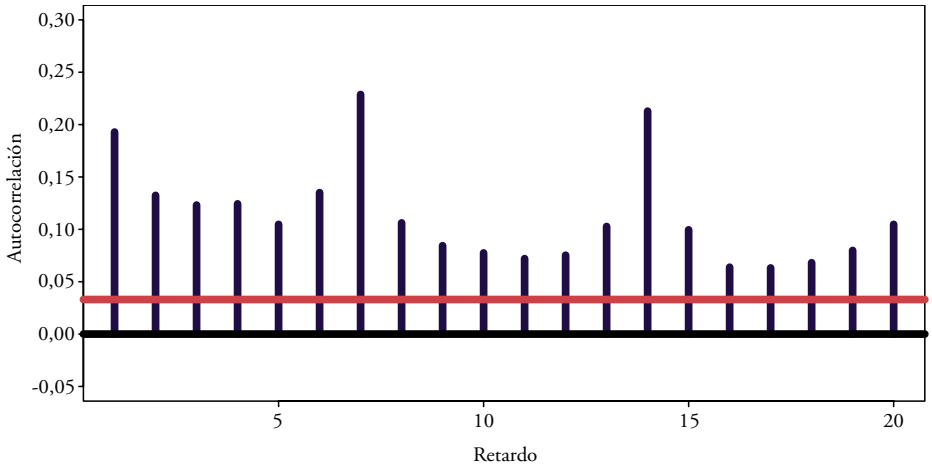
es decir, el producto interior entre las funciones base β_m y β_k. En particular, si $m = k$, entonces:

$$B_{mm} = \int_0^{24} \beta_m(t)^2 \, dt = \langle \beta_m, \beta_m \rangle = \| \beta_m \|^2, \qquad m = 1, \ldots, 10 \qquad [14]$$

es decir, la norma al cuadrado de la función base β_m. La función inprod de la librería fda de R proporciona el valor de la matriz B tras realizar integración numérica de las integrales en [13] y [14] mediante la regla trapezoidal, por lo que el usuario no tiene que realizar el cálculo de ninguna integral. El proceso de cálculo es extremadamente rápido por lo que el coste computacional es muy bajo. En particular, la figura 7 muestra la función de autocorrelación funcional muestral en [12] para las ÇIDRs de los precios de la electricidad. Más concretamente el gráfico muestra los valores de $\hat{r}_l$, para l = 1,...,20,, junto con una línea a la altura del intervalo de confianza al 95 % de dichas autocorrelaciones bajo la hipótesis de que la serie es ruido blanco funcional. El gráfico informa de dos aspectos interesantes. En primer lugar, existen autocorrelaciones significativas lo que parece indicar que las CIDRs para los precios de la electricidad pueden ser predecibles a partir de las CIDRs previas. En cualquier caso, las autocorrelaciones no son muy altas, por lo que las predicciones pueden no ser extremadamente buenas. En segundo lugar, las autocorrelaciones para l = 7 y l = 14 son significativas, indicando que existe un cierto efecto estacional debido a que las curvas representan los días de la semana. En cualquier caso, este efecto puede no parece ser lo suficientemente grande como para que el proceso generador sea no estacionario, ya que ninguna de estas autocorrelaciones está cerca de 1.

Figura 7.

Función de autocorrelación funcional muestral para las CIDRs del precio de la electricidad



Fuente: Elaboración propia.

4. COMPONENTES PRINCIPALES FUNCIONALES DE LAS CURVAS DE RENDIMIENTOS INTRADÍA ACUMULADOS

Todas las medidas definidas en la sección tercera resumen las características más importantes de la serie temporal funcional de CIDRs, $\theta_1,...,\theta_{2100}$. Sin embargo, la herramienta más útil para entender el comportamiento de las CIDRs y para realizar predicciones de futuras CIDRs es el análisis de componentes principales funcionales (FPCA, es su acrónimo en inglés). FPCA es el análogo funcional de la conocida técnica de reducción de dimensión en el análisis estadístico multivariante y aquí será útil para determinar las tendencias comunes que están presentes en la dinámica de las CIDRs. Más concretamente, los componentes principales funcionales (FPCs) son funciones que resumen las principales fuentes de variación de las CIDRs con respecto a la media funcional muestral en [6]. Más concretamente, el FPCA permite representar las CIDRs mediante una expansión de la forma,

$$\theta_i(t) = \overline{\theta}(t) + \sum_{m=1}^{10} s_{im}\psi_m(t), \qquad t \in [0,24)$$

donde los FPC, $\psi_1,...,\psi_{10}$, son funciones ortogonales suaves de norma uno, y las puntuaciones asociadas a los FPC, $s_{im} = \langle \theta_i - \overline{\theta}, \psi_m \rangle$, para $m = 1,...,10$, son valores reales de media muestral 0, y de tal manera que cada CIDR se puede aproximar con mínima pérdida de información mediante los primeros M < 10 FPC como sigue:

$$\theta_i(t) \simeq \overline{\theta}(t) + \sum_{m=1}^{M} s_{im}\psi_m(t). \qquad [15]$$

La expresión [15] muestra que $\psi_1,...,\psi_M$ representan las tendencias comunes que dominan el comportamiento diario de las CIDRs.

Para calcular los FPC se procede de manera similar a como se calculan sus homólogos multivariantes. En particular, el primer FPC se define como la función continua y suave ψ_1 definida en el intervalo [0,24) tal que la varianza de sus puntuaciones asociadas dada por,

$$\frac{1}{2099}\sum_{i=1}^{2100} s_{i1}^2$$

sea máxima sujeto a que $||\psi_1|| = 1$. Para resolver el problema, nótese que, por un lado, dadas las expansiones [3] y [6], la *i*-ésima CIDR centrada i - $\theta_i - \overline{\theta}$ se puede escribir como:

$$\theta_i(t) - \overline{\theta}(t) = \tilde{c}_i'\beta(t), \qquad t \in [0,24) \qquad [16]$$

donde $\tilde{c}_i = c_i - \overline{c}$, para $i = 1,...,2100$, mientras que, por otro lado, la primera FPC ψ_1 se puede escribir en términos de la base de B-splines mediante la expansión:

$$\psi_1(t) = \sum_{m=1}^{10} v_{1m}\beta_m(t) = v_1'\beta(t), \qquad t \in [0,24) \qquad [17]$$

donde $v_1 = (v_{11},...,v_{1,10})'$ es un vector de coeficientes a determinar. Por lo tanto, substituyendo [16] y [17] en la definición de las puntuaciones de la primera FPC, se obtiene que,

$$s_{i1} = \left\langle \theta_i - \overline{\theta}, \psi_1 \right\rangle = \tilde{c}_i' B v_1,$$

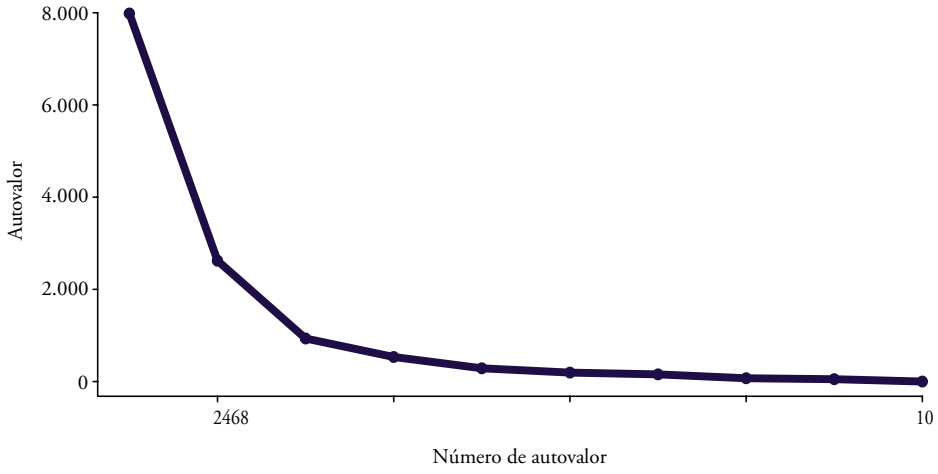
donde B es la matriz de productos interiores de los elementos de la base de B-splines $\beta_1,...,\beta_{10}$ introducida en la sección tercera, dando lugar a que la varianza de las puntuaciones se puedan escribir de manera sencilla mediante la expresión:

$$\frac{1}{2099} \sum_{i=1}^{2100} s_{i1}^2 = v_1' B S_0 B v_1.$$

Además, tras substituir [17] en la definición de $||\psi_1||$, se obtiene que la restricción se puede escribir como $\|\psi_1\|^2 = v_1' B v_1 = 1$. En resumen, el problema de optimización se reduce a obtener el vector v_1 tal que el valor de $v_1' B S_0 B v_1$ sea máximo sujeto a que $v_1' B v_1 = 1$, cuya solución viene dada por $v_1 = B^{1/2} d_1$, donde d_1 el autovector ligado al mayor autovalor a_1 de la matriz $B^{1/2} S_0 B^{1/2}$. De esta manera, se obtiene el vector de coeficientes v_1 en términos de la base de B-splines del primer FPC en [17]. Nótese además que las puntuaciones asociadas a la primera FPC tiene media muestral 0 y varianza muestral a_1.

Figura 8.

Autovalores asociados a los componentes principales funcionales de las CIDRs para los precios de la electricidad



Fuente: Elaboración propia.

El segundo FPC se define como la función continua y suave ψ_2 definida en el intervalo [0,24) tal que la varianza de sus puntuaciones asociadas sea máxima sujeto a que $||\psi_2|| = 1$ y a que ψ_2 sea ortogonal a ψ_1, es decir, que $\langle \psi_1, \psi_2 \rangle = 0$. Como en el caso anterior, la segunda FPC ψ_2 se puede escribir en términos de la base de B-splines mediante la expansión:

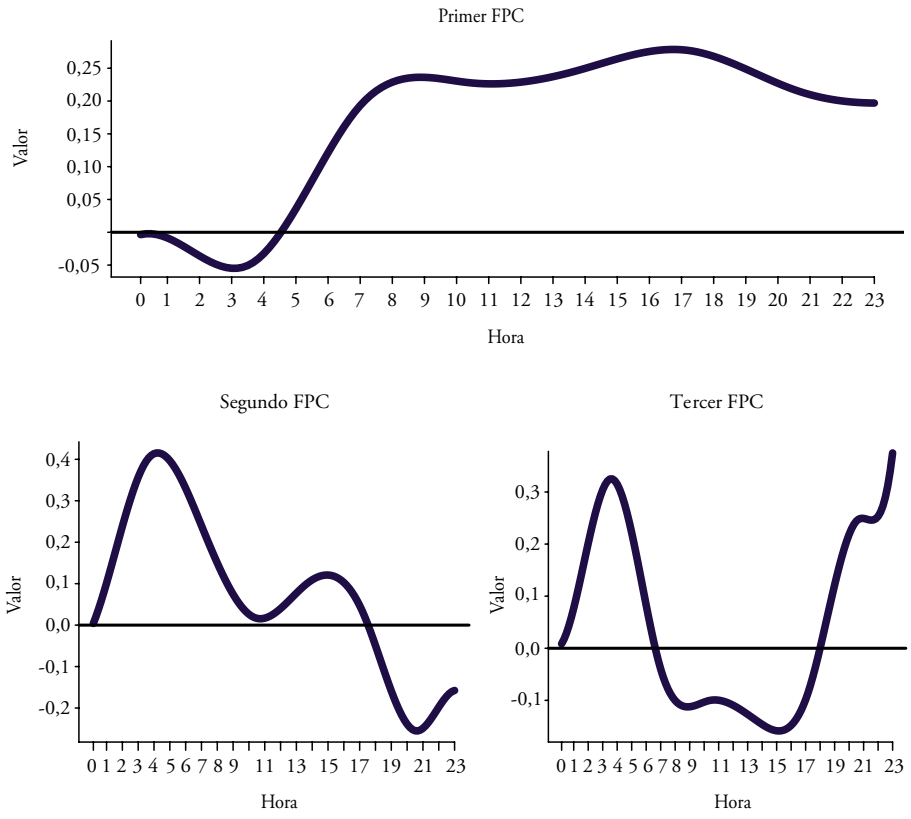
$$\psi_2(t) = \sum_{m=1}^{10} v_{2m} \beta_m(t) = v_2' \beta(t), \qquad t \in [0,24) \qquad [18]$$

donde $v_2 = (v_{21},...,v_{2,10})'$ es un vector de coeficientes a determinar. Procediendo de manera similar al cálculo del primer FPC, se llega a que la solución al problema de optimización viene dada por que $v_2 = B^{1/2}d_2$ es el vector de coeficientes del segundo FPC en [18], donde d_2 es el autovector ligado al segundo mayor autovalor a_2 de la matriz $B^{1/2}S_0B^{1/2}$. Como es previsible, el resto de FPC se obtiene de manera similar por lo que los FPC de las CIDRs son las funciones $\psi_1,..., \psi_{10}$ que tienen expansiones de B-splines con vectores de coeficientes $v_m = B^{-1/2}d_m$, para $m = 1,... 10$, donde $d_1,..., d_{10}$ son los autovectores de la matriz $B^{1/2}S_0B^{1/2}$ asociados a los autovalores $a_1,..., a_{10}$. Además las puntuaciones del m-ésima FPC tienen media muestral 0 y varianza muestral a_m, lo que permite utilizar el criterio clásico de acumulación de varianzas para seleccionar un valor apropiado de M.

La figura 8 muestra los 10 autovalores $a_1,..., a_{10}$ asociados a los FPC de las CIDRs de tal manera que la acumulación de las varianzas del primero hasta el quinto FPC es capaz de

Figura 9.

Tres primeros componentes principales funcionales de las CIDRs para los precios de la electricidad



Fuente: Elaboración propia.

explicar el 62,21 %, 82,65 %, 89, 92 %, 94,07 % y 96,32 %, respectivamente, de la suma de las varianzas de todos los FPC. A la vista de la figura y de estos porcentajes, los primeros $M = 3$ FPC, que se muestran en la figura 9, proporcionan las principales fuentes de variación de las CIDRs. Los tres FPC parecen tener una relación evidente con varios momentos del día. Por ejemplo, el primer FPC está fundamentalmente relacionado con la variabilidad creada desde las 7:00 horas de la mañana hasta las 23:00 horas, es decir, el período de mayor actividad. A continuación, el segundo FPC contrapone la variabilidad creada durante la madrugada y la variabilidad creada a última hora de la tarde. De hecho, en la covarianza funcional muestral de las CIDRs que se muestra en la figura 6, se puede observar una relación negativa entre ambos momentos del día. Por último, el tercer FPC contrapone la variabilidad creada de madrugada y a última hora del día con la variabilidad creada desde la mañana hasta media tarde, que es el período ligado a la jornada laboral más habitual. El cálculo de los FPC de las CIDRs se ha llevado a cabo mediante la función pca.fd de la librería fda de R.

5. PREDICCIÓN DE LAS CURVAS DE RENDIMIENTOS INTRADÍA ACUMULADOS

Una vez descritas las características más importantes de las CIDRs como serie temporal funcional, el objetivo de esta sección es realizar predicción para las mismas. Existen dos opciones principales para la predicción de series temporales funcionales: (1) mediante los modelos autorregresivos y de media móvil funcional; y (2) mediante los procedimientos basados en componentes principales funcionales.

Los modelos autorregresivos y de media móvil funcionales son extensiones directas de los modelos autorregresivos y de media móvil univariantes y multivariantes. Dada la popularidad de este tipo de modelos para el análisis de series temporales univariantes y multivariantes, no es de extrañar que sus extensiones fuesen los primeros modelos que fueron utilizados para el análisis de series temporales funcionales. Más concretamente, Bosq (1991) propuso el modelo autorregresivo funcional de orden 1, denotado por FAR(1), que fue extendido al modelo autorregresivo funcional de orden p, denotado por FAR(p), por Bosq (2000) y Bosq y Blanke (2007). Más recientemente, Klepsch y Klüppelberg (2016), Klepsch, Klüppelberg y Wei (2017) y Li, Robinson y Shang (2020) han propuesto, respectivamente, el modelo de media móvil funcional, denotado por FMA(q), el modelo autorregresivo y de media móvil funcional, denotado por FARMA(p,q), y el modelo autorregresivo fraccionalmente integrado y de media móvil funcional, denotado por FFARIMA(p,d,q). Expresiones, características y propiedades de todos estos modelos se pueden encontrar en los artículos mencionados y no se presentan aquí por su elevada complejidad técnica. De entre todos los modelos anteriores, el más popular en la práctica para el ajuste de series temporales funcionales es el modelo FAR(1). Esto se debe a dos razones fundamentales. En primer lugar, los modelos FAR(p) han sido analizados y utilizados desde hace alrededor de dos décadas (véanse las referencias mencionadas en la introducción), mientras que los modelos FMA(q), FARMA(p,q) y FFARIMA(p,d,q) han sido propuestos muy recientemente. Por lo tanto, las ventajas del uso de estos modelos frente al uso de los modelos FAR(p) no son claras todavía. Se menciona este punto porque, tal como

Kokoszka y Reimherr (2013) hicieron notar, al contrario que la autorregresión escalar y vectorial, solo valores muy pequeños de p, 1 o 2 como mucho, son relevantes en un modelo FAR(p). Esto es debido a que cada una de las curvas se considera un único objeto funcional y, por lo tanto, la dependencia entre funciones estacionarias temporalmente alejadas debe ser escasa o nula, salvo que exista algún tipo de dependencia a muy largo plazo, quizás debida a algún tipo de efecto estacional. En segundo lugar, este tipo de modelos se define a partir de operadores complejos que requieren complicadas condiciones de existencia. Más aún, la estimación de dichos operadores está basada en formas regularizadas de las ecuaciones de Yule-Walker que son relativamente sencillas de manejar solamente en el caso del modelo FAR(1), por lo que utilizar este modelo permite simplificar el procedimiento inferencial. Por estas y otras razones que se detallarán más adelante, los procedimientos basados en componentes principales funcionales son actualmente más populares para la obtención de predicciones de series temporales funcionales reales (véanse de nuevo las referencias mencionadas en la introducción). De hecho, para llevar a cabo predicciones de las CIDRs, que tienen un cierto efecto estacional, parece más conveniente el uso de este tipo de procedimientos que se presentan a continuación.

La predicción con componentes principales funcionales es bastante sencilla, estando a la vez también ligada, como los modelos autorregresivos y de media móvil funcionales, al uso de métodos para la predicción de series temporales univariantes y multivariantes. La idea de este método está basada en utilizar la expansión [15] de las CIDRs en términos de los primeros FPC. Más concretamente, dicha expansión permite escribir las CIDRs como:

$$\theta_i(t) = \overline{\theta}(t) + s_i'\psi(t) + \varepsilon_i(t), \qquad [19]$$

donde $s_i = (s_{i1}, s_{i2}, s_{i3})'$ es el vector que contiene las puntuaciones de la i-ésima CIDR asociadas a las tres primeras FPC, $\psi(t) = (\psi_1(t), \psi_2(t), \psi_3(t))'$ es el vector de FPC evaluados en tiempo $t \in [0, 24)$, y ε_i es una función de media funcional muestral cero que no debe contener información útil de cara a la predicción de las CIDRs futuras. Ahora, dada la serie de CIDRs $\theta_1,...,\theta_{2100}$, el objetivo es obtener la predicción de la serie funcional de CIDRs a un día vista mediante la expansión [19]. En general, la CIDR correspondiente al día siguiente del último observado, es decir θ_{2100}, se puede predecir mediante:

$$\hat{\theta}_{2101}(t) = \overline{\theta}(t) + \hat{s}_{2101}'\psi(t), \qquad [20]$$

donde $\hat{s}_{2101} = \left(\hat{s}_{2101,1}, \hat{s}_{2101,2}, \hat{s}_{2101,3}\right)'$ es el vector de predicciones de las puntuaciones asociadas a los tres primeros FPC de las CIDRs. Por lo tanto, el problema de predecir la CIDR θ_{2100} se transforma en el problema de predecir el vector de puntuaciones dado por $s_{2101} = (s_{2101,1}, s_{2101,2}, s_{2101,3})'$. El resto de la sección se dedica a presentar dos maneras diferentes de obtener la predicción $\hat{s}_{2101}$ y, por tanto, de obtener la predicción $\hat{\theta}_{2101}$ en [20]. Además se presenta un procedimiento sencillo para obtener una banda de confianza para dicha predicción funcional.

En primer lugar, el procedimiento propuesto por Hyndman y Ullah (2007) y posteriormente analizado y utilizado en Hyndman y Shang (2009), Shen (2009) y Shang (2017), entre otros, consiste en obtener las predicciones $\hat{s}_{2101,1}$, $\hat{s}_{2101,2}$ y $\hat{s}_{2101,3}$, de manera

independiente. Es decir, las predicciones $\hat{s}_{2101,1}$, $\hat{s}_{2101,2}$ y $\hat{s}_{2101,3}$ resultan de predecir a un día vista las tres series temporales univariantes de puntuaciones $s_{1m},...,s_{2100,m}$, para $m = 1, 2$ y 3. Esta decisión es justificable en el sentido de que las puntuaciones de los FPC son incorreladas. En principio, para realizar predicciones univariantes, se puede emplear cualquier procedimiento apropiado para las características de las series de puntuaciones. Si las series de puntuaciones se pueden ajustar bien mediante modelos ARIMA, lo normal sería utilizar los procedimientos automáticos de predicción con modelos ARIMA univariantes donde el modelo se selecciona mediante un criterio de selección de modelos, como el criterio de Akaike (AIC), el criterio de Akaike corregido (AICc) o el criterio Bayesiano (BIC), ver Brockwell y Davis (2017), Shumway y Stoffer (2017) y Hyndman y Athanasopoulos (2018), para referencias sobre este tipo de procedimientos. Sin embargo, si las series de puntuaciones son relativamente complejas y los modelos ARIMA no son adecuados, se pueden utilizar otros procedimientos como pueden ser los diferentes métodos de alisado exponencial, modelos de series temporales no lineales, redes neuronales, o cualquier otra alternativa.

En segundo lugar, el procedimiento propuesto por Aue, Norinho y Hörmann (2015), consiste en obtener la predicción del vector $\hat{s}_{2101} = \left(\hat{s}_{2101,1}, \hat{s}_{2101,2}, \hat{s}_{2101,3}\right)$, de manera conjunta. Es decir, se obtiene la predicción a un día vista de la serie temporal trivariante $s_i = (s_{i1}, s_{i2}, s_{i3})'$, donde $i = 1,...,2100$. Esta decisión también es justificable ya que, si bien las puntuaciones de los FPC son incorreladas, pueden existir autocovarianzas cruzadas entre diferentes puntuaciones diferentes de 0. Para realizar las predicciones multivariantes, Aue, Norinho y Hörmann (2015) proponen utilizar modelos vectoriales autorregresivos (VAR) para los que también existen procedimientos automáticos de predicción basados en criterios de selección de modelos, véase, por ejemplo, Lütkepohl (2006) y Tsay (2017).

Una vez presentados los dos procedimientos basados en FPC, es importante resaltar dos ventajas adicionales que tienen estos métodos con respecto a los métodos basados en modelos ARMA funcionales. La primera ventaja está basada en que Aue, Norinho y Hörmann (2015) demostraron que las predicciones obtenidas con su procedimiento y utilizando un modelo FAR(p) son asintóticamente equivalentes. Por lo tanto, dado que la obtención de las predicciones con FPC es más sencilla, parece entonces razonable utilizar este método. La segunda ventaja está basada en que Liebl (2013) demostró que la descomposición en [15] es perfectamente válida incluso si la serie temporal funcional es no estacionaria. Por lo tanto, las predicciones con FPC se pueden utilizar sin mayor complicación en el caso no estacionario. Más detalles sobre este último punto se pueden encontrar en Lansangan y Barrios (2009) y Shen (2009).

En series temporales univariantes y multivariantes, los intervalos y las regiones de predicción, respectivamente, juegan un papel fundamental para medir la incertidumbre asociada a las predicciones puntuales. Además, son útiles para comparar predicciones de diferentes métodos y explorar diferentes escenarios basados en diferentes supuestos. En el caso de series temporales funcionales, los intervalos o regiones de predicción se substituyen con bandas de predicción que tienen en cuenta la estructura funcional de dicha predicción. Existen diferentes maneras de obtener estas bandas de predicción. Por un lado, Aue, Norinho y Hörmann (2015) propusieron un método simple basado en predicciones

dentro de la muestra. Por otro lado, Shang (2017) y Paparoditis y Shang (2020) propusieron procedimientos bootstrap para obtener dichas bandas. Por sencillez en la exposición y dado que los procedimientos bootstrap son computacionalmente más costosos, a continuación, se describe el procedimiento propuesto por Aue, Norinho y Hörmann (2015) adaptado a la obtención de una banda de predicción a un paso para las CIDRs de los precios de la electricidad en España. Para ello, se dispone de los FPC ψ_1, ψ_2, ψ_3, de sus series de puntuaciones asociadas $s_{1m},\ldots,s_{2100,m}$, para $m = 1, 2, 3$, y de un procedimiento para las predicciones univariantes para las series de puntuaciones individuales o de un procedimiento para la predicción del vector de series de puntuaciones. El procedimiento funciona siguiendo los cinco pasos que se describen a continuación:

1. Utilizar el procedimiento de predicción univariante o multivariante para construir predicciones a un paso de las CIDRs $\theta_L,\ldots,\theta_{2100}$, donde L es un cierto entero pequeño.
2. Obtener residuos funcionales basados en las predicciones anteriores, $\hat{\theta}_L,\ldots,\hat{\theta}_{2100}$, dados por $\hat{\varepsilon}_i = \theta_i - \hat{\theta}_i$, para $i = L+1,\ldots,2100$.
3. Obtener la desviación típica funcional muestral de los residuos funcionales dada por:

$$\hat{\sigma}_\varepsilon(t) = \sqrt{\frac{1}{2100-L-1}\sum_{i=L}^{2100}\left(\hat{\varepsilon}_i(t) - \bar{\hat{\varepsilon}}(t)\right)^2},$$

   donde $\bar{\hat{\varepsilon}}$ denota la media funcional muestral de $\hat{\varepsilon}_L,\ldots,\hat{\varepsilon}_{2100}$.

4. Encontrar dos valores q_L y q_U tales que el $\alpha \times 100\%$ de los residuos funcionales verifiquen:

$$q_L\hat{\sigma}_\varepsilon(t) \le \hat{\varepsilon}_i(t) \le q_U\hat{\sigma}_\varepsilon(t), \qquad t \in [0,24].$$

5. La banda de confianza para la predicción $\hat{\theta}_{2100}$ está dada por:

$$\hat{\theta}_{2101}(t) + q_L\hat{\sigma}_\varepsilon(t) \le \hat{\theta}_{2101}(t) \le \hat{\theta}_{2101}(t) + q_U\hat{\sigma}_\varepsilon(t), \qquad t \in [0,24].$$

A continuación, se considera la predicción de las CIDRs a un paso utilizando los procedimientos basados en FPC descritos previamente. Para ello, se realizan dos ejercicios diferentes. En el primer ejercicio, se obtienen predicciones funcionales para las últimas 100 CIDRs observadas basadas en las CIDRs previas como sigue:

1. Para cada $i = 2001,\ldots,2100$, se utilizan las primeras i-1 CIDRs para obtener los primeros tres FPC y las series de puntuaciones asociadas.
2. Se aplican los procedimientos de Hyndman y Ullah (2007) y de Aue, Norinho y Hörmann (2015) para obtener sendas predicciones a un paso de la CIDR el día de negociación i utilizando la expresión [20].
3. Para cada predicción obtenida se calculan el error cuadrático integrado medio de predicción (MISFE) y el error absoluto integrado medio de predicción (MIAFE) dados por:

$$MISFE_i = \int_0^{24} \left(\theta_i(t) - \hat{\theta}_i(t)\right)^2 dt,$$

y

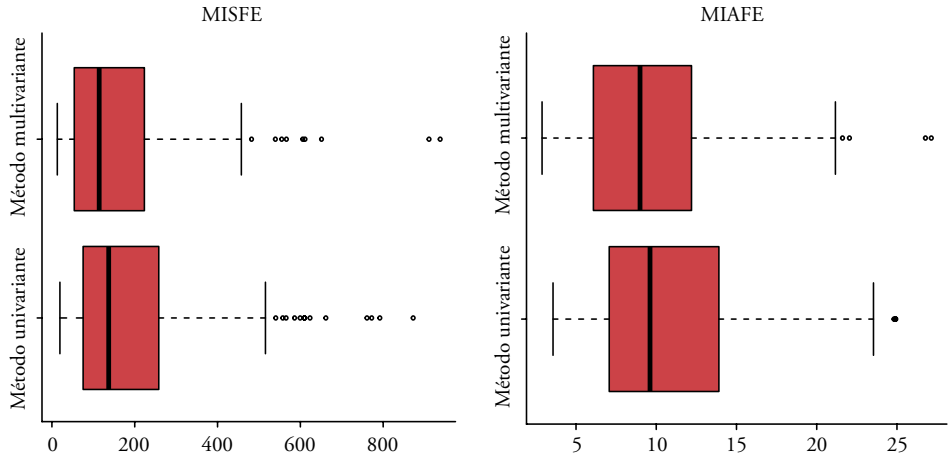
$$MIAFE_i = \int_0^{24} |\theta_i(t) - \hat{\theta}_i(t)| \, dt,$$

respectivamente, que permitirán determinar que método de predicción es más apropiada para cada CIDR $\theta_{2001},...,\theta_{2100}$, y más en general para las 100 CIDRs en su conjunto.

Por un lado, para obtener las predicciones con el método de Hyndman y Ullah (2007), se utiliza la función forecast.ftsm de la librería ftsa de R, que permite realizar las predicciones de las series de puntuaciones mediante una selección automática de modelos ARIMA, modelos de espacio de estado para suavizado exponencial o paseos aleatorios con deriva. Para las CIDRs, se ha seleccionado la predicción con modelos ARIMA. Por otro lado, para obtener las predicciones con el método de Aue, Norinho y Hörmann (2015), se utiliza la función farforecast de la misma librería, que solo permite realizar las predicciones del vector de series de puntuaciones mediante una selección automática de modelos VAR. Los 100 valores obtenidos de ambas medidas del error de predicción se pueden observar en la figura 10, que muestra que el procedimiento de Aue, Norinho y Hörmann (2015) es ligeramente superior al procedimiento de Hyndman y Ullah (2007), como se esperaba. De hecho, las medias muestrales de los valores del criterio MISFE son 216.7 y 178.6, respectivamente, para los procedimientos de Hyndman y Ullah (2007) y Aue, Norinho y Hörmann (2015), mientras que las medias respectivas para el criterio MIAFE son 11.129 y 9.867, respectivamente.

Figura 10.

Error cuadrático integrado medio de predicción (MISFE) y error absoluto integrado medio de predicción (MIAFE) para las 100 últimas CIDRs para los precios de la electricidad en España

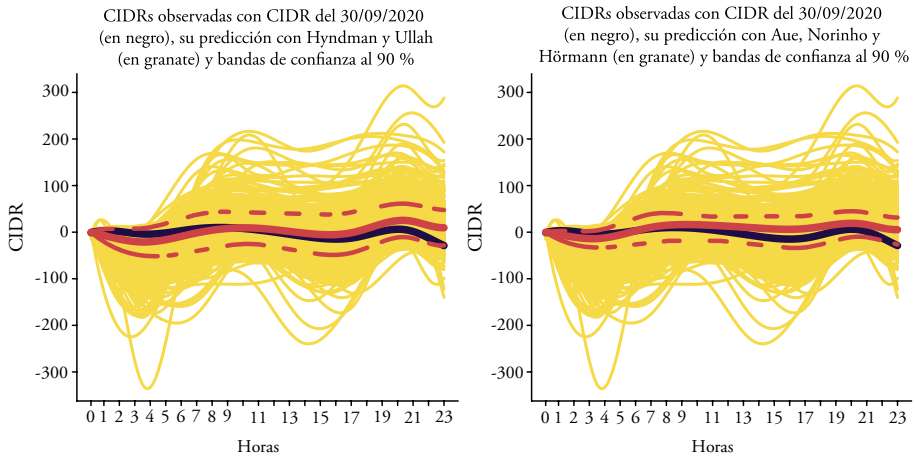


Fuente: Elaboración propia.

En el segundo ejercicio, se obtiene la predicción funcional y la banda de predicción asociada para la última CIDR observada correspondiente al 30 de septiembre de 2020 con los dos procedimientos considerados en el ejercicio anterior. Los valores del MIAFE para la predicción con el procedimiento de Hyndman y Ullah (2007) y con el procedimiento de Aue, Norinho y Hörmann (2015) son 12.220 y 12.513, respectivamente, que ocupan los lugares 65 y 78 de entre los valores del MIAFE obtenidos para los últimos 100 días. Por lo tanto, las predicciones obtenidas no son de las mejores de entre el conjunto de predicciones de los últimos 100 días observados. Más concretamente, los valores del criterio MISFE para esta CIDR son 228.7 y 221,5, respectivamente, para los procedimientos de Hyndman y Ullah (2007) y Aue, Norinho y Hörmann (2015), mientras que los valores respectivos para el criterio MIAFE son 12.22 y 12.51, respectivamente. Esto indica que ninguna de las dos predicciones para la última CIDR domina a la otra. La figura 11 muestra la verdadera CIDR correspondiente al 30 de septiembre de 2020 (en negro), las predicciones con los dos métodos utilizados (en granate), una en cada gráfico de la figura, sus correspondientes bandas de predicción asociadas al 90 % (en granate y línea discontinua), y el resto de CIDRs (en amarillo). Como se puede apreciar, las dos predicciones son bastante parecidas, por lo que no sorprende los valores parecidos obtenidos de los criterios MIAFE y MISFE, y bastante cercanas a la verdadera CIDR teniendo en cuenta los valores del resto de CIDRs. En ambos casos, la banda de predicción correspondiente contiene a toda la CIDR del 30 de septiembre de 2020 durante todo el día.

Figura 11.

A la derecha, predicción funcional para la CIDR del 30 de septiembre de 2020 junto con una banda de predicción asociada para el método de Hyndman y Ullah (2007), y a la izquierda, predicción funcional para la CIDR del 30 de septiembre de 2020 junto con una banda de predicción asociada para el método de Aue, Norinho y Hörmann (2015)



Fuente: Elaboración propia.

6. CONCLUSIONES

Este capítulo presenta una aplicación sencilla del análisis de series temporales funcionales a los precios horarios de la electricidad de España en el mercado MIBEL. Más concretamente, el análisis descrito analiza las curvas de rendimientos intradía acumulados de los precios horarios de la electricidad. El objetivo fundamental del capítulo es el de llamar la atención sobre un conjunto de técnicas que están apareciendo en el área del análisis de datos funcionales y que están despertando el interés de un número importante de estadísticos y de profesionales en otras áreas. Evidentemente, todavía hay un gran espacio para la mejora y para el desarrollo de técnicas más completas y que cubran aspectos relevantes como, por ejemplo, el análisis conjunto de varias series temporales funcionales. Además, el análisis de las curvas de rendimientos intradía acumulados de los precios horarios de la electricidad puede ser más completo teniendo en cuenta aspectos que no se describen en este artículo meramente introductorio.

Referencias

Antoniadis, A., Paparadotis, E. y Sapatinas, T. (2006). A functional wavelet-kernel approach for time series prediction. *Journal of the Royal Statistical Society, Series B,* 68, pp. 837–857.

Aue, A., Horváth, L. y Pellatt, D. F. (2017). Functional generalized autoregressive conditional heteroskedasticity. *Journal of Time Series Analysis,* 38, pp. 3–21.

Aue, A., Norinho, D. D. y Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association,* 110, pp. 378–392.

Besse, P. C., Cardot, H. y Stephenson, D. B. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics,* 27, pp. 673–687.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics,* 31, pp. 307–327.

Bosq, D. (1991). Modelization, nonparametric estimation and prediction for continuous time processes. En G. Roussas (eds.), *Nonparametric Functional Estimation and Related Topics. NATO ASI Series (Series C: Mathematical and Physical Sciences),* 335, pp. 509–529.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications.* Berlin: Springer.

Bosq, D. y Blanke, D. (2007). *Inference and Prediction in Large Dimensions.* Chichester: Wiley.

Brockwell, P. J. y Davis, R. A. (2017). *Introduction to Time Series and Forecasting* (third edition). New York: Springer.

Dickey, D. A. y Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association,* 74, pp. 427–431.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica,* 50, pp. 987–1007.

Engle, R. F. y Granger, C. W. J. (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica,* 55, pp. 251–276.

Ferraty, F. y Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice.* New York: Springer.

Gabrys, R., Horváth, L. y Kokoszka, P. (2010). Tests for error correlation in the functional linear model. *Journal of the American Statistical Association,* 105, pp. 1113–1125.

Galeano, P. y Peña, D. (2019). Data science, big data and statistics. *Test,* 28, pp. 289–329.

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics,* 23, pp. 121–130.

Hays, S., Shen, H. y Huang, J. Z. (2012). Functional dynamic factor models with application to yield curve forecasting. *Annals of Applied Statistics,* 6, pp. 870–894.

Hörmann, S., Horváth, L. y Reeder, R. (2013). A functional version of the ARCH model. *Econometric Theory,* 29, pp. 267–288.

Hörmann, S. y Kokoszka, P. (2010). Weakly dependent functional data. *Annals of Statistics,* 38, pp. 1845–1884.

Horváth, L. y Kokoszka, P. (2012). *Inference with Functional Data with Applications.* New York: Springer.

Horváth, L., Kokoszka, P. y Rice, G. (2014). Testing stationarity of functional time series. *Journal of Econometrics,* 179, pp. 66–82.

Horváth L., Rice, G. y Whipple, S. (2016). Adaptive bandwidth selection in the estimation of the long run covariance of functional time series. *Computational Statistics & Data Analysis,* 100, pp. 676–693.

Hsing, T. y Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators.* West Sussex: John Wiley & Sons.

Hyndman, R. J. y Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (second edition). Melbourne: OTexts.

Hyndman, R. J. y Shang, H. L. (2009). Forecasting functional time series (with discusssions). *Journal of the Korean Statistical Society,* 38, pp. 199–221.

Hyndman, R. J. y Ullah, S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis,* 51, pp. 4942–4956.

Klepsch, J. y Klüppelberg, C. (2016). An innovations algorithm for the prediction of functional linear processes. *Journal of Multivariate Analysis,* 155, pp. 252–271.

Klepsch, J., Klüppelberg, C. y Wei, T. (2017). Prediction of functional ARMA processes with an application to traffic data. *Econometrics and Statistics,* 1, pp. 128–149.

Kokoszka, P., Miao, H. y Zhang, X. (2015). Functional dynamic factor model for intraday price curves. *Journal of Financial Econometrics,* 13, pp. 456–477.

Kokoszka, P. y Reimherr, M. (2013). Determining the order of the functional autoregressive model. *Journal of Time Series Analysis,* 34, pp. 116–129.

Kokoszka, P. y Reimherr, M. (2017). *Introduction to Functional Data Analysis.* CRC press, Boca Ratón.

Kowal, D. R., Matteson, D. S. y Ruppert, D. (2017). A Bayesian multivariate functional dynamic linear model. *Journal of the American Statistical Association,* 112, pp. 733–744.

Kwiatkowski, D., Philips, P. C. B., Schmidt, P. y Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics,* 54, pp. 159–178.

Lansangan, J. R. G. y Barrios, E. B. (2009). Principal components analysis of nonstationary time series data. *Statistics and Computing,* 19, pp. 173–187.

Li, D., Robinson, P. M. y Shang, H. L. (2020). Long-range dependent curve time series. *Journal of the American Statistical Association,* 115, pp. 957–971.

Liebl, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *The Annals of Applied Statistics,* 7, pp. 1562–1592.

Lütkepohl, H. (2006). *New Introduction to Multiple Time Series Analysis.* New York: Springer.

Paparoditis, E. y Shang, H. L. (2020). Bootstrap prediction bands for functional time series. https://arxiv.org/abs/2004.03971v1

Ramsay, J. O. y Silverman, B. W. (2005). *Functional Data Analysis,* 2nd edition. New York: Springer.

Said, E. S. y Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika,* 71, pp. 599–607.

Sargan, J. D. y Bhargava, A. (1983). Testing residuals from the least squares regression for being generated by the Gaussian random walk. *Econometrica,* 51, pp. 153–174.

Shang, H. L. (2017). Forecasting intraday S&P 500 index returns: A functional time series approach. *Journal of Forecasting,* 36, pp. 741–755.

Shen, H. (2009). On modeling and forecasting time series of smooth curves. *Technometrics,* 51, pp. 227–238.

Shumway, R. H. y Stoffer, D. S. (2017). *Time Series Analysis and its Applications* (fourth edition). New York: Springer.

Srivastava, A. y Klassen, E. (2016). *Functional and Shape Data Analysis.* New York: Springer.

Taylor, S. J. (1986). *Modeling Financial Time Series.* Chichester: Wiley.

Tsay, R. S. (2017). *Multivariate Time Series Analysis: With R and Financial Applications.* New Jersey: Wiley.

CAPÍTULO IX

Predicción y clasificación basada en distancias parcialmente observadas

Aldo R. Franco Comas*
Andrés M. Alonso Fernández

El procedimiento de vecinos más cercanos, k-NN en inglés, se utiliza para la clasificación de nuevas observaciones empleando la matriz de distancias entre las observaciones a clasificar y las observaciones en la muestra de entrenamiento. En este trabajo, desarrollamos un procedimiento k-NN para situaciones donde no es posible calcular todas las distancias entre las nuevas observaciones y las observaciones en la muestra de entrenamiento. Por otra parte, k-NN depende de la distancia utilizada y proponemos un procedimiento para el aprendizaje de la distancia a partir de las distancias en la muestra de entrenamiento. El procedimiento propuesto se ilustra utilizando conjuntos de datos reales.

Palabras clave: aprendizaje de distancias, clustering, curvas de oferta, k-NN, matrices de distancia.

* Este trabajo ha sido parcialmente financiado por la Agencia Estatal de Investigación (PID2019-108311GB-I00/AEI/10.13039/501100011033) y por la Comunidad de Madrid en el marco del convenio plurianual con la Universidad Carlos III Madrid en su línea de actuación "Excelencia para el Profesorado Universitario" del V Plan Regional de Investigación Científica e Innovación Tecnológica 2016-2020. Los autores agradecen los comentarios y mejoras sugeridas por los editores del libro, Daniel Peña, Pilar Poncela y Esther Ruiz.

1. INTRODUCCIÓN

El algoritmo *k-NN (k- Nearest Neighbors* o k-Vecinos más Cercanos) es probablemente el procedimiento de clasificación y regresión más fácil de entender, incluso por personas que estén alejadas del área de conocimiento del aprendizaje estadístico. En lo que sigue, se presentan las características básicas del procedimiento k-NN que es la herramienta básica de este trabajo.

Los datos de entrada tienen la forma (x_i, y_i) siendo x_i un vector de longitud p donde están representadas p variables o atributos para el caso i-ésimo, mientras que y_i es la variable respuesta o etiqueta. La respuesta puede ser un valor numérico como por ejemplo el salario, la estatura o el peso; también puede ser un valor categórico como el tipo de contrato laboral o la tipología de riesgo de accidentes. Igualmente se necesita una distancia entre un caso x_i y otro x_j denotada por d (x_i, x_j). Para cada nuevo caso x_0, se calculan las distancias d (x_0, x_i) para i = 1,2,...,n, siendo n el número total de casos en el conjunto de entrenamiento y se buscan los k casos más cercanos según esa distancia. En la clasificación k-NN, la salida generalmente es una clase o etiqueta. Al nuevo objeto se le asigna la clase más común entre sus k vecinos más cercanos. Si k = 1, entonces el nuevo objeto simplemente se asigna a la clase del vecino más cercano. En la regresión k-NN, la salida usualmente es el promedio de los valores de los k vecinos más cercanos. Existen otros procedimientos de asignación/clasificación y predicción como vecinos más cercanos ponderados donde se da mayor peso a los casos cercanos en función de la distancia al caso nuevo (ver Biccego y Loog, 2016).

Como ventajas principales del k-NN tenemos: 1) No paramétrico: no hace suposiciones explícitas sobre la forma funcional de los datos, evitando los inconvenientes de alejarse de la distribución o modelo subyacente; 2) Precisión satisfactoria: se obtienen precisiones altas, sin embargo puede no ser competitiva en comparación con otros métodos de aprendizaje supervisado, como pueden ser las redes neuronales o las máquinas de vector soporte (SVM) que utilizan procesos de entrenamiento mucho más costosos computacionalmente; 3) Evoluciona constantemente: dado que es un aprendizaje basado en casos, el algoritmo se adapta inmediatamente a medida que recopilamos nuevos datos de entrenamiento. Esto permite que el algoritmo responda rápidamente a los cambios en la entrada durante el uso en tiempo real; 4) Fácil de implementar para problemas de múltiples clases: la mayoría de los algoritmos son fáciles de implementar para problemas binarios, sin embargo requieren un esfuerzo extra a la hora de ser implementados para múltiples clases, como puede ser el caso de las SVM, mientras que k-NN se ajusta a múltiples clases sin dificultad; 5) Clasificación y regresión: una de las mayores ventajas del k-NN es que se puede usar tanto para problemas de clasificación como de regresión; 6) Variedad de distancias: ofrece una alta flexibilidad a la hora de elegir cualquier distancia al construir el modelo.

Mientras que sus desventajas son: 1) Datos no balanceados: no funciona bien en datos no balanceados, es decir, cuando hay alguna clase claramente menos representada; 2) Sensibilidad a valores atípicos: es sensible a los valores atípicos, ya que simplemente elige a los vecinos según los criterios de distancia; 3) Presencia de valores perdidos: no funciona cuando hay valores perdidos en las variables independientes o regresoras; 4) Necesita carac-

terísticas homogéneas: es recomendable que las características tengan la misma escala para evitar que alguna variable sea determinante por el hecho de utilizar unidades de medida diferente del resto de las variables; 5) Maldición de la dimensión: bajo un amplio conjunto de condiciones, a medida que aumenta la dimensionalidad, la distancia al punto de datos más cercano se acerca a la distancia al punto de datos más lejano (ver Beyer *et al.,* 1999); 6) Computacionalmente costoso: puede ser muy fácil de implementar, pero a medida que crece el conjunto de datos, la velocidad del algoritmo disminuye rápidamente.

Algunas de estas desventajas son compartidas por otros procedimientos de clasificación/regresión. En este trabajo nos centraremos en la desventaja del coste computacional que aparece cuando tenemos un problema donde n es muy grande, como ocurre en el contexto de datos masivos (*big data*), y calcular todas las distancias para predecir no es factible. Esta situación puede darse, por ejemplo, cuando, por razones de tiempo de respuesta, por un tiempo de cómputo excesivo (en conjuntos de datos extremadamente grandes) o pruebas destructivas (análisis de ADN), no es posible calcular todas las distancias de las nuevas observaciones a todas las observaciones en la muestra de entrenamiento. En este trabajo, propondremos una solución para los casos en que no sea posible calcular todas las distancias del caso a clasificar a todos los elementos del conjunto de entrenamiento.

1.1. Objetivos

El objetivo fundamental de este trabajo es proponer un procedimiento k-NN donde no sea necesario calcular todas las distancias cada vez que tengamos que predecir un caso nuevo. Dado que solamente podemos calcular un porcentaje de las distancias, las restantes deben ser imputadas, este es el problema al cual nos enfrentamos. Compararemos distintos enfoques que se han propuesto para completar matrices de distancias (Dhillon, Sra y Tropp, 2005; De Soete, 1984; Lapointe y Kirsch, 1995) en combinación con el procedimiento k-NN en términos del error del vector de distancias y del orden de vecinos resultantes. Los métodos también se compararan por el coste computacional aún sabiendo que esto último depende del ordenador, lenguaje y forma de implementación. Todos los experimentos se realizaron usando R (R Development Core Team, 2008) y los códigos desarrollados están disponibles mediante solicitud a los autores.

1.2. Estructura

El resto del trabajo se estructura en las siguientes secciones. En la sección segunda se hace una revisión de la literatura sobre algoritmos que resuelven el problema de completamiento de matrices de distancias y se proponen dos procedimientos de inicialización. Se realizan comparaciones mediante simulación para determinar el método que obtiene mejores resultados. En la sección tercera se utiliza el método propuesto en el conjunto de datos MNIST (LeCun y Cortes, 2010) para un problema de clasificación de múltiples clases (10 clases) y en datos del mercado eléctrico español para un problema de regresión. La sec-

ción cuarta presenta las conclusiones y extensiones de trabajo. Las definiciones básicas que serán de utilidad para la formulación del problema de completamiento de matrices de distancias y la evaluación de los procedimientos se presentan en un apéndice.

2. METODOLOGÍA: ANTECEDENTES Y PROPUESTA

En esta sección, primero, se realiza una breve revisión bibliográfica de artículos que abordan problemas similares al nuestro como lo es el problema de la métrica más cercana o el problema de la inferencia filogenética. Se propone una solución basada en procedimientos clúster y se compara el desempeño de las distintas opciones mediante un estudio de simulación.

2.1. El problema de la métrica más cercana

Supongamos que tenemos una matriz $\boldsymbol{D}$ de dimensión $n \times n$ donde representamos las distancias entre n individuos. Por tanto, los elementos de $\boldsymbol{D}$ deberían cumplir las desigualdades triangulares pero, por errores de medición o incluso por omisión de algunas mediciones, estas desigualdades no se verifican. El *problema de la métrica más cercana* consiste en encontrar una matriz $\boldsymbol{M}$ cuyos elementos cumplan las correspondientes desigualdades triangulares y que esté próxima a la matriz $\boldsymbol{D}$. Este problema ha sido estudiado en Dhillon, Sra y Tropp (2003 y 2005) y Brickell *et al.* (2008), y puede formularse como sigue:

> Si tenemos n puntos, podemos representar las medidas entre cada par de puntos en una matriz simétrica $\boldsymbol{D}$ cuya entrada (i, j) representa la "distancia" entre los objetos i y j. Buscamos aproximar esta matriz por otra matriz, $\boldsymbol{M}$, cuyas entradas satisfacen las desigualdades triangulares. Es decir, $m_{ik} \leq m_{ij} + m_{jk}$ por cada tupla (i, j, k). En otras palabras se requiere una matriz de distancias $\boldsymbol{M}$ que sea la más cercana a una matriz de disimilitud dada $\boldsymbol{D}$ con respecto a alguna norma entre matrices. Específicamente, se busca una matriz de distancias $\boldsymbol{M}$ tal que,
>
> $$\boldsymbol{M} = \arg \min_{\boldsymbol{X} \in \mathcal{M}_n} \left\{ \left\| \boldsymbol{W} \odot (\boldsymbol{X} - \boldsymbol{D}) \right\| \right\}, \qquad [1]$$
>
> donde $\mathcal{M}_n$ es el conjunto de todas las matrices de distancias $n \times n$, $||\cdot||$ es una norma matricial, $\boldsymbol{W}$ es una matriz $n \times n$ de pesos, simétrica y no negativa, y $\odot$ denota la multiplicación elemento a elemento entre dos matrices. La matriz de pesos $\boldsymbol{W}$ refleja nuestra confianza en las entradas de la matriz $\boldsymbol{D}$. Por ejemplo, cuando cada d_{ij} representa una medida con varianza σ_{ij}^2, podríamos utilizar $w_{ij} = 1/\sigma_{ij}^2$ y si falta una entrada en $\boldsymbol{D}$ podríamos poner cero en el peso correspondiente.

El problema de la métrica más cercana se diferencia del escalado multidimensional métrico en que no busca una matriz de datos cuyas distancias (usualmente euclídeas) estén cerca de la matriz $\boldsymbol{D}$ ni impone ninguna hipótesis sobre el espacio subyacente más que requerir que sea un espacio métrico.

En Dhillon, Sra y Tropp (2003) se prueba que el problema [1] alcanza su mínimo en $\mathcal{M}_n$. Además, cada mínimo local es un mínimo global. Si, además, la norma es estrictamente convexa y la matriz de pesos no tiene ceros o infinitos fuera de su diagonal, entonces hay un mínimo global único. En principio, es posible utilizar cualquier norma entre matrices para este problema, pero lo analizaremos para las normas L_r. Con lo cual los problemas asociados son:

$$\min_{\mathbf{X}\in\mathcal{M}_n} [\sum_{j\neq k} |w_{jk}(x_{jk}-d_{jk})|^r]^{1/r} \quad \text{si} \quad 1\leq r<\infty \qquad [2]$$

y

$$\min_{\mathbf{X}\in\mathcal{M}_n} \max_{j\neq k} |w_{jk}(x_{jk}-d_{jk})| \quad \text{si} \quad r=\infty. \qquad [3]$$

Para resolver los problemas [2] y [3] puede parecer que se debería utilizar programación lineal cuando $r = 1$ ó ∞, y programación convexa para $r > 1$, pero resulta que los requisitos de tiempo y almacenamiento de estos enfoques son prohibitivos. En Brickell *et al.* (2008) se proponen algoritmos para resolver el problema [1] para las normas L_1, L_2, y L_∞ y siguiendo su propuesta, se planteará el algoritmo para $r = 2$ puesto que este caso resulta ser el más simple y juega un papel fundamental en la resolución de los problemas para $r = 1$ y $r = \infty$. El algoritmo *triangle fixing* va recorriendo todas las tuplas (i, j, k) y corrigiendo aquellas que no satisfacen la desigualdad triangular.

En el algoritmo 1, las $e_{ij} = m_{ij} - d_{ij}$ representan los cambios de las disimilaridades originales, d, a las distancias finales, m; $b_{ijk} = d_{ki} + d_{jk} - d_{ij}$ indica en qué medida se viola la desigualdad triangular, y z es tal que $e = -\boldsymbol{A}'z$ con $\boldsymbol{A}$ la matriz que codifica las desigualdades triangulares. Si en una tupla (i, j, k) se incumple la desigualdad triangular, tenemos que $e_{ij} - e_{jk} - e_{ki} > b_{ijk}$. Para resolver este incumplimiento se proyecta ortogonalmente el vector e sobre el conjunto de restricciones $\{x : x_{ij} - x_{jk} - x_{ki} \leq b_{ijk}\}$, lo que devuelve como solución a:

$$\begin{aligned} x_{ij} &\longleftarrow e_{ij} + \mu, \\ x_{jk} &\longleftarrow e_{jk} + \mu, \\ x_{ki} &\longleftarrow e_{ki} + \mu, \end{aligned} \qquad [4]$$

con $\mu = \dfrac{-b_{ijk} + e_{ij} - e_{jk} - e_{ki}}{3}$. Por lo tanto, solo tres componentes de e tienen que ser actualizadas, y sus valores están dados por [4]. El procedimiento se repite hasta que ninguna tupla reciba una actualización significativa.

En las simulaciones utilizaremos $r = 2$. El algoritmo para este valor está implementado en R en la librería *dtw (Dynamic Time Warping)* (Giorgino, 2009).

Llegados a este punto, nos preguntamos ¿qué relación tiene el problema de la métrica más cercana con nuestro problema? Si asumimos que conocemos todas las distancias entre los n objetos en la muestra de entrenamiento podemos crear una nueva matriz $(n + 1) \times (n + 1)$ donde la columna/fila $(n + 1)$ serían las distancias del punto x_0 a las restantes teniendo en cuenta que

Algoritmo 1.

Algoritmo Triangle Fixing para L_2

Entrada: Matriz de disimilitud, $\boldsymbol{D}$, y tolerancia, ε.

Salida: $\boldsymbol{M} = argmin_{\boldsymbol{X} \in \mathcal{M}_N} ||(\boldsymbol{X} - \boldsymbol{D})||_2$.

for $1 \leq i < j < k \leq n$ **do**

| $z_{ijk} \leftarrow 0$

end

for $1 \leq i < j \leq n$ **do**

| $e_{ij} \leftarrow 0$

end

$\delta \leftarrow 1 + \varepsilon$

while $\delta > \varepsilon$ **do**

  foreach *Incumplimiento de la desigualdad triangular para (i, j, k)* **do**

    $b_{ijk} \leftarrow d_{ki} + d_{kj} - d_{ij}$

    $\mu \leftarrow - \frac{1}{3}(b_{ijk} - e_{ij} + e_{jk} + e_{ki})$

    $\theta \leftarrow \text{mín } \{-\mu, z_{ijk}\}$

    $e_{ij} \leftarrow e_{ij} - \theta$

    $e_{jk} \leftarrow e_{jk} + \theta$

    $e_{ki} \leftarrow e_{ki} + \theta$

    $z_{ikj} \leftarrow z_{ikj} - \theta$

  end

  $\delta \leftarrow$ *Suma de todos los cambios en los valores de e.*

end

return $M = D + E$

tendremos un ℓ % de distancias conocidas y un $(1 - \ell)$ % de distancias sin calcular y podemos utilizar el algoritmo 1 para imputar esos valores. A estos valores faltantes tenemos que darle algún valor inicial pues el algoritmo triangle fixing necesita valores numéricos. Hemos considerado varias opciones para inicializar estos valores faltantes:

- $d(x_0, x_j) = 0$, siendo j aquel objeto del cual no tenemos la distancia. De este primer valor, no se reportan resultados pues el tiempo de ejecución era extremadamente alto al igual que las diferencias relativas entre matrices.
- $d(x_0, x_j) = \dfrac{\sum_{i=1}^{n} d(x_i, x_j)}{n}, i \neq j.$
- $d(x_0, x_j) = \text{mediana } (d(x_i, x_j)), \forall i \neq j.$

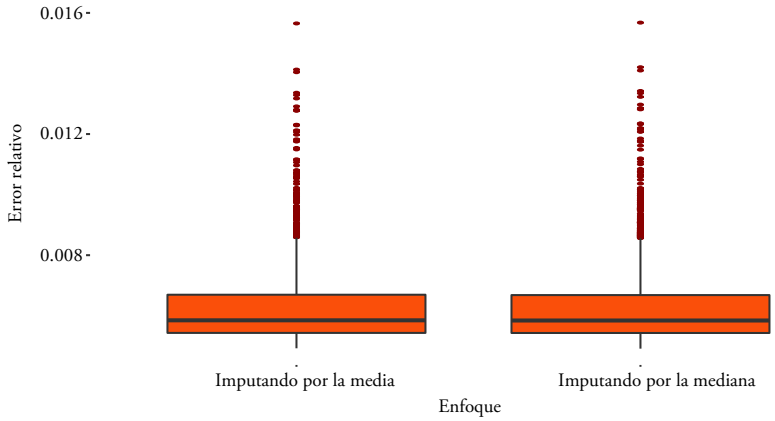
Para comparar las distintas propuestas de valores iniciales de las distancias desconocidas realizamos un estudio de simulación que consta de los siguientes pasos:

1. Se generan $n = 800$ puntos de una distribución normal multivariante con $\mu = 0_p$ y $\Sigma = I_p$, siendo $p = 20$.
2. Se calcula la matriz de distancias, usando la distancia euclidiana.
3. Se genera un nuevo punto x_0 y se calculan las n distancias.
4. Se asumen conocidas las distancias de x_0 a $n\ell = 80$ puntos, ($\ell = 0{,}1$) y las restantes distancias se imputan usando los métodos descritos anteriormente, con lo cual tenemos una matriz de disimilitud, D.
5. Se aplica el algoritmo triangle fixing a D y nos devuelve una matriz M.
6. Se calcula la diferencia relativa entre la matriz M y la matriz de distancias original al igual que la diferencia relativa entre el vector de distancias correspondientes a x_0 en la matriz original y en M. También se registra el tiempo de cómputo.
7. Se repiten los pasos 3 – 6, $N = 1.000$ veces.

Se pueden apreciar en la figura 1 las diferencias relativas entre las matrices que devuelve el algoritmo y las matrices originales. Estas grandes diferencias y altos costes en tiempo (como se puede apreciar en la figura 2) vienen dados por el hecho que estas imputaciones no verifican holgadamente las desigualdades triangulares, con lo cual el algoritmo tiene que iterar muchas veces. Por tanto debemos buscar una forma de mejorar estas aproximaciones iniciales.

Figura 1.

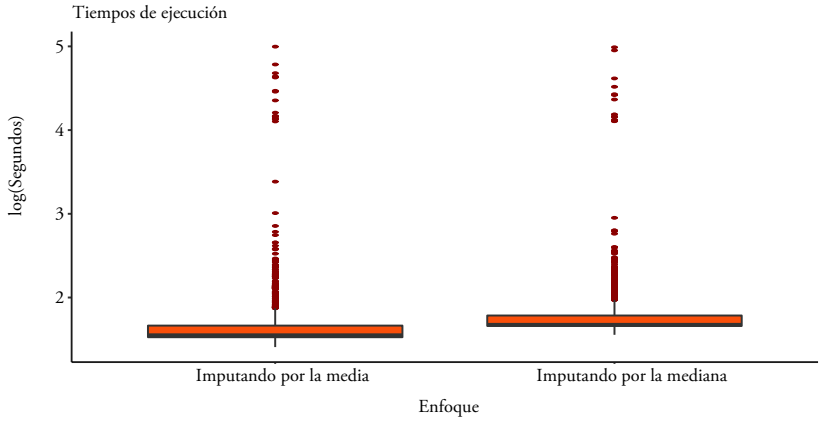
Diagramas de caja de las diferencias relativas entre la matriz de distancias original y la matriz imputada, *M*



Fuente: Elaboración propia.

Figura 2.

Diagramas de caja de los tiempos de ejecución para cada enfoque en escala logarítmica



Fuente: Elaboración propia.

2.1.1. Método de inicialización propuesto

Teniendo en cuenta las restricciones impuestas a nuestro problema, sabemos que las distancias $d(x_i, x_j)$ son conocidas $\forall i, j \in \{1,2,...,n\}$ y podemos calcular las distancias de $d(x_0, x_i)$ cuando $i \in I$ con $|I| \ll n$ pero no podemos calcular las $d(x_0, x_i)$ cuando $i \in I^c$. Los cardinales de los conjuntos I e I^c son aproximadamente iguales a ℓn y $(1 - \ell n)$, respectivamente.

Por otra parte, haciendo uso de las desigualdades triangulares, tenemos:

$$d(x_0, x_{i^*}) \leq d(x_0, x_i) + d(x_i, x_{i^*}) \quad \forall i \in I \text{ e } i^* \in I^c \qquad [5]$$

y, puesto que la desigualdad se tiene $\forall i \in I$, se concluye que:

$$d(x_0, x_{i^*}) \leq \min_{i \in I} \{d(x_0, x_i) + d(x_i, x_{i^*})\}. \qquad [6]$$

De igual forma, tenemos que:

$$d(x_i, x_{i^*}) \leq d(x_0, x_{i^*}) + d(x_0, x_i) \qquad [7]$$

y

$$d(x_0, x_i) \leq d(x_0, x_{i^*}) + d(x_i, x_{i^*}) \qquad [8]$$

por tanto, obtenemos que:

$$|d(x_0, x_i) - d(x_i, x_{i^*})| \leq d(x_0, x_{i^*}), \qquad \forall i \in I \text{ e } i^* \in I^c \qquad [9]$$

lo que implica que:

$$\max_{i \in I} | d(x_0, x_i) - d(x_i, x_{i^*}) | \leq d(x_0, x_{i^*}). \qquad [10]$$

Notar que las desigualdades [6] y [10] nos permiten acotar las distancias $d(x_0, x_{i^*})$ mediante:

$$\max_{i \in I} | d(x_0, x_i) - d(x_i, x_{i^*}) | \leq d(x_0, x_{i^*}) \leq \min_{i \in I} \{d(x_0, x_i) + d(x_i, x_{i^*})\} \qquad [11]$$

e imputaremos dichas distancias con la media de sus cotas:

$$d(x_0, x_{i^*}) = \frac{1}{2}\left(\min_{i \in I} \{d(x_0, x_i) + d(x_i, x_{i^*})\} + \max_{i \in I} | d(x_0, x_i) - d(x_i, x_{i^*}) | \right). \qquad [12]$$

Cuando asignamos los valores imputados usando la expresión [12] para posteriormente aplicar el algoritmo triangle fixing obtenemos que la matriz $\boldsymbol{D}$ coincide con la matriz $\boldsymbol{M}$ que devuelve dicho algoritmo. Con lo cual parece ser que esos valores imputados satisfacen todas las condiciones que debe verificar una matriz de distancia. Tenemos una prueba parcial de esta propiedad y hemos comprobado mediante simulación utilizando ℓ entre 0,1 y 0,4, pero no tenemos una demostración completa de esta conjetura.

Dado el coste computacional del algoritmo triangle fixing, imputaremos las distancias faltantes usando la expresión [12]. Hasta ahora hemos seleccionado el conjunto de puntos, I, tomando una muestra aleatoria en el conjunto de entrenamiento. En lo que sigue propondremos un método que nos ayude a mejorar estos resultados basándonos en métodos clúster.

2.2. Clúster o agrupamiento

El clustering o agrupamiento es una técnica de aprendizaje no supervisada que intenta encontrar relaciones entre variables pero no la relación que guardan con respecto a una variable objetivo, es decir sin hacer uso de las etiquetas. El conjunto de datos tiene que ser dividido automáticamente en clústeres, de manera que los objetos dentro del mismo clúster sean similares, mientras que los objetos de diferentes clústeres sean menos semejantes. No existe una definición general de clúster, lo que significa que diferentes enfoques pueden obtener diferentes clústeres del mismo conjunto de datos.

El motivo por el cual el agrupamiento será de gran importancia en este trabajo está dado por el hecho de que hasta ahora hemos seleccionado los puntos a los cuales se le calculan las distancias de una forma totalmente aleatoria. Pero usaremos la idea propuesta en Zhang *et al.* (2016), donde en un conjunto de datos muy grandes se decide aplicar k-NN por submuestras y cada submuestra es definida mediante un clúster. A continuación se puede ver un resumen de dicho algoritmo.

Algoritmo 2.

k-NN basado en clústeres

Entrada: Datos de entrenamiento y dato a clasificar x_0.

Salida: Etiqueta.

1. Producir K clústeres, C_1, C_2, ..., C_K.
2. Calcular la distancia de x_0 a todos los centroides, $d(x_0, C_i)\ \forall i = 1, ..., K$.
3. Buscar el centroide (C_i) más cercano a x_0.
4. Crear un nuevo conjunto de datos, correspondientes a los puntos que pertenecen a C_i.
5. Usar estos datos como datos de entrenamiento para predecir x_0.

Existen muchos métodos de clustering como, por ejemplo, la agrupación jerárquica, los métodos K-medias, *PAM* (*Partitioning Around Medoids*), y *DBSCAN* (*Density-Based Spatial Clustering of Application with Noise*). Estos son de los más conocidos y usados (Schubert *et al.,* 2017), sin embargo dado el impacto de los datos masivos existen trabajos que intentan paralelizar estos procedimientos y escalar a conjuntos de datos más grandes.

2.2.1. K-medias

Uno de los algoritmos más usados es el K-medias, el cual tiene como objetivo crear K grupos a partir de n observaciones, en el que cada una pertenece a un grupo cuyo valor medio es el más cercano. Sean $(x_1, x_2,..., x_n)$ las observaciones, el algoritmo construye una partición de dichas observaciones, $C = \{C_1, ..., C_K\}$, con el fin de minimizar la suma de los cuadrados dentro de cada grupo:

$$\arg\min_{C} \sum_{i=1}^{K} \sum_{x_j \in C_i} \| x_j - \mu_i \|_2^2, \qquad [13]$$

donde μ_i es la media de las observaciones en C_i.

Desafortunadamente, aunque el agrupamiento de K-medias es bastante eficiente en tiempo computacional se sabe que es sensible a los valores atípicos. Por esta razón, a veces se utiliza la agrupación de K-medoides, donde se consideran objetos representativos dentro de cada clúster en lugar de los centroides.

2.2.2. K-medoides

K-medoides es una familia de algoritmos que escogen puntos del conjunto de datos como centros y trabaja con una métrica arbitraria de distancias. Es más robusto ante atípicos que K-medias porque minimiza una suma de disimilaridades (entre pares de puntos) en vez de una suma de distancias euclidianas a una media. Un medoide puede ser definido, como el objeto de un grupo cuya disimilaridad media a todos los objetos en el grupo es mínima.

La implementación práctica más común de K-medoides es el algoritmo Partición Alrededor de Medoides (PAM) cuya idea es similar a K-medias en el sentido de asociar cada punto a un representante más cercano, pero sustituye la media del clúster en [13] por un elemento del clúster. Otros algoritmos K-medoides son CLARA y CLARANS (Schubert y Rousseeuw, 2019). Sin embargo, PAM tiene el inconveniente de que funciona de manera ineficiente para conjuntos de datos grande (Han, Kamber y Tung, 2001). Existe una versión, basada en K-medias, que se denomina fastkmed y que permite el manejo de grandes conjuntos de datos. El algoritmo fastkmed parte de la idea de seleccionar mediodes iniciales basándose en la suma de las distancias relativas al conjunto de datos.

El resultado de varios estudios de similación muestran que dicho método tiene mejor rendimiento que el agrupamiento K-medias. También se reduce significativamente el tiempo de ejecución con respecto al PAM con un rendimiento comparable, PAM es del orden de $O(k(n-k)^2)$ y el fastkmed del orden de $O(nk)$ muy similar al K-medias (Park y Jun, 2009).

Algoritmo 3.

Algoritmo fastkmed

Entrada: Seleccionar K de los n puntos como medoides.

Salida: K medoides.

1. Seleccione los medoides iniciales:
   - Calcular la distancia entre cada par de objetos.
   - Calcular v_j para el objeto j como sigue:

$$v_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{i=1}^{n} d_{il}}, \; j = 1, \ldots, n$$

   - Ordenar v_j en orden ascendente y seleccionar los k objetos con menores v_j como medoides iniciales.
   - Obtener la partición inicial asignando cada objeto al medoide más cercano.
   - Calcular la suma de distancias de todos los objetos a sus medoides.
2. Actualizar medoides:
   - Encontrar un nuevo medoide de cada grupo, que es el objeto que minimiza la distancia total a otro objeto en su clúster. Actualizar el medoide actual en cada grupo reemplazando con el nuevo medoide.
3. Asignar objetos a los medoides:
   - Asignar cada objeto al medoide más cercano y obtener el resultado del clúster.
   - Calcular la suma de la distancia de todos los objetos a sus medoides. Si la suma es igual a la anterior, detener el algoritmo y en caso contrario, volver al paso 2.

2.3. Procedimientos aditivo y ultramétrico

El problema de la inferencia filogenética a partir de conjuntos de datos que incluyen entradas incompletas o inciertas es uno de los temas más relevantes en biología sistemática (Lapointe y Makarenkow, 2004). El objetivo es inferir filogenias o relaciones de parentesco a partir de datos evolutivos, incluida información faltante, por ejemplo, cuando las secuencias de nucleótidos o proteínas observadas contienen huecos o entradas faltantes.

Entre los diferentes algoritmos de inferencia filogenética hay dos que se han usado para completar matrices de distancias. Estos métodos han sido propuestos por De Soete (1984) y Lapointe y Kirsch (1995) para ultramétricas, mientras que para matrices de distancias aditivas por Landry, Lapointe y Kirsch (1996). Ambos algoritmos están implementados en R, en la librería *ape* (*Analyses of Phylogenetics and Evolution*) desarrollada por Paradis y Schliep (2018). Estos algoritmos utilizan las imputaciones que se derivan de las definiciones 5.5 y 5.6, respectivamente.

Algoritmo 4.

Algoritmo para distancias ultramétricas

Entrada: Distancia parcial d en el conjunto de n observaciones.

Salida: Distancia completa o parcial d en el conjunto de n observaciones.

1. Contar el número de NA en d, se denota por c.

2. **foreach** $d(i, j)$ *que sea NA de* d **do**

   $MinMax$ es la máxima entrada de d.

   foreach k *tal que* $d(i, k)$ *y* $d(j, k)$ son valores conocidos de d ***do***

      $Max = \max(d(i, k); d(j, k))$

      if $Max = minMax$ **then**

         $Max = minMax$

      end

   end

   if *Si hay al menos un par conocido de entradas* $d(i, k)$ *y* $d(j, k)$ **then**

      $d(i, j) = MinMax$

      $c = c - 1$

   end

end

3. Si cambia c, ir al paso 2.

Algoritmo 5.

Algoritmo para distancias aditivas

Entrada: Distancia parcial d en el conjunto de n observaciones.

Salida: Distancia completa o parcial d en el conjunto de n observaciones.

1. Contar el número de NA en d, se denota por c.

2. **foreach** $d(i, j)$ que sea *NA de* d **do**

   $MinMax$ es la máxima entrada de d.

   foreach k, l tal que $d(i, k)$, $d(j, k)$, $d(i, l)$, $d(j, l)$ y $d(k, l)$ son valores conocidos de d ***do***

      $Max = \max(d(i, k) + d(j, l); d(i, l) + d(j, l) - d(k, l))$

      if $Max < minMax$ **then**

         $Max = minMax$

      end

   end

   if Si hay al menos las entradas $d(i, k)$, $d(j, k)$, $d(i, l)$, $d(j, l)$ *y* $d(k, l)$ **then**

      $d(i, j) = MinMax$

      $c = c - 1$

   end

end

3. Si cambia c, ir al paso 2.

Estos algoritmos tienen la ventaja de no necesitar ningún tipo de imputación inicial. De hecho los valores faltantes deben definirse como NA para poder hacer uso de estos métodos.

2.4. Estudio de simulación

En esta sección se compararán los diferentes métodos propuestos y para ello usaremos dos métricas, el MAE y el índice de Jaccard, es decir, el error absoluto medio entre la matriz real y la matriz obtenida y la concordancia entre el conjunto de vecinos real y el derivado a partir de la matriz de distancias obtenida. Las definiciones formales pueden verse en el apéndice. Haremos simulaciones donde compararemos los resultados observación a observación, es decir, en ningún momento utilizaremos las etiquetas que pueden tener estas observaciones. Este ejercicio se hará en la próxima sección para el conjunto de datos MNIST usando la mejor estrategia que salga de estas simulaciones.

El motivo por el cual usaremos ambas métricas viene dado por el hecho de que el índice de Jaccard presenta un inconveniente. Por ejemplo, supongamos que los vecinos más cercanos para una observación de validación x_0 son las observaciones $A = \{10, 15, 20, 30\}$ de la muestra de entrenamiento y obtenemos que un algoritmo nos devuelve los siguientes puntos más cercanos $B_1 = \{10, 15, 20, 35\}$ y otro algoritmo los puntos $B_2 = \{10, 15, 20, 40\}$ con lo cual se obtiene que $J(A, B_1) = J(A, B_2) = 3/5$. Vemos que el índice de Jaccard coincide y no permite diferenciar entre ambos resultados. Con el MAE seleccionaríamos el resultado con menor error de imputación de las distancias.

Primero, estudiaremos el número de clústeres que debemos escoger de tal manera que el índice de Jaccard sea mayor. Es importante destacar que este problema es diferente al que resuelven métodos de selección del número de clústeres presentes en un conjunto de datos.

Para encontrar un número "óptimo" de clústeres mediante simulación se realizan los siguientes pasos:

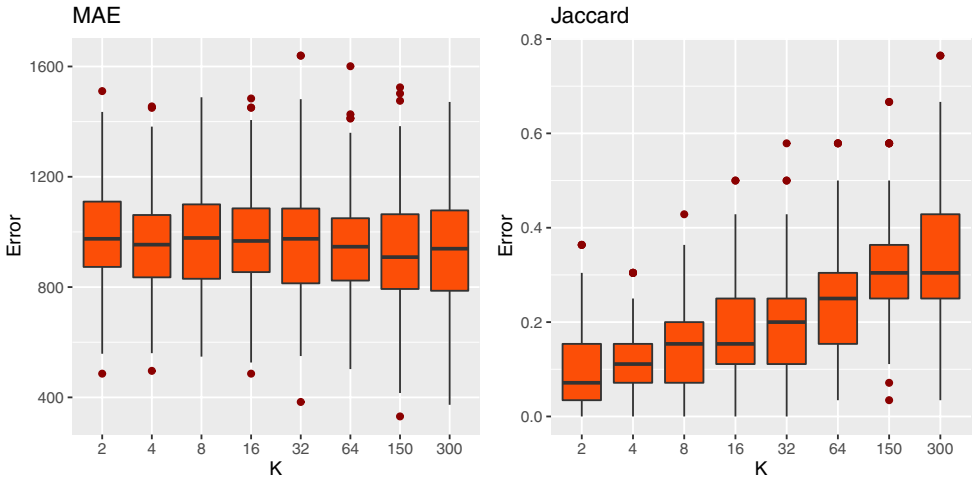
1. Se generan n = 3.000 puntos de una distribución normal multivariante con $\mu = 0_p$ y $\Sigma = I_p$, siendo $p = 50$.
2. Se calcula la matriz de distancias, usando la distancia euclidiana.
3. Se aplica fastkmed a dicha matriz de distancias para diferentes valores de $K = (2, 4, 8, 16, 32, 64, 150, 300)$, donde K es el número de clústeres. De tal manera se obtienen K mediodes $\{M_1, M_2, ..., M_K\}$.
4. Se genera un nuevo punto x_0.
5. Se calculan las distancias $d(x_0, M_i)$ con $i = 1, ..., K$ y se ordenan de menor a mayor.
6. En este punto ya hemos calculado K distancias con lo cual las restantes se harán seleccionando de cada clúster más cercano hasta que se hayan calculado el número máximo de distancias $n\ell = 300$, pues $\ell = 0{,}1$.
7. Para los diferentes valores de K expuestos se calcula el índice de Jaccard y el MAE entre el conjunto real de puntos más cercano y el conjunto de puntos más cercano que se obtiene imputando, para $k = 15$ vecinos.
8. Se repiten los pasos 4 - 7, $N = 200$ veces.

Se puede observar en la figura 3 que a medida que el número de clústeres aumenta el índice de Jaccard mejora, con lo cual a partir de este momento decidiremos usar como número de clústeres el valor correspondiente a $K = n\ell/2$ ($K = 150$ en la figura), pues presenta menor variabilidad que $n\ell$ que sería el valor correspondiente a $K = 300$. También observamos que el MAE decrece lentamente con el número de clústeres.

A continuación, compararemos los procedimientos de imputación usando puntos al azar o usando clústeres:

Figura 3.

Búsqueda de un valor "óptimo" del número de clústeres, *K*



Fuente: Elaboración propia.

1. Se generan n = 5.000 puntos de una distribución normal multivariante con $\mu = 0_p$ y $\Sigma = I_p$, siendo p =50.

2. Se calcula la matriz de distancias, usando la distancia euclidiana.

3. Se genera un nuevo punto x_0.

4. Se calculan las distancias de x_0 a los n puntos.

5. Se calculan las distancias de x_0 a $n\ell$ puntos al azar y también a la misma cantidad usando $K = n\ell/2$ clústeres. En ambos casos las restantes distancias se imputan usando la expresión [12].

6. Se ordenan dichas distancias y se extrae cuáles son los k puntos más cercanos, el máximo valor que toma k es $\sqrt{n}/2$.

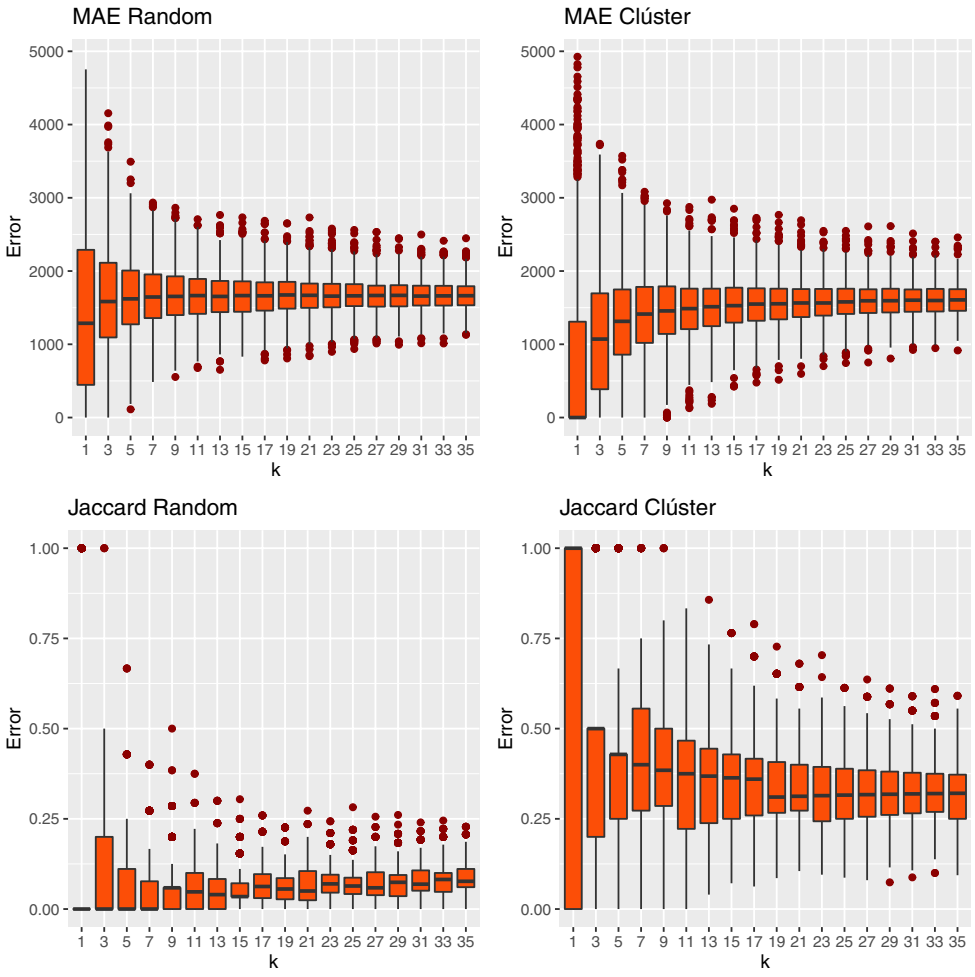
7. Para los diferentes valores de k y los dos procedimientos, se calcula el índice de Jaccard y el MAE entre el conjunto de puntos más cercano real y el conjunto de puntos más cercano que se obtiene mediante imputación.

8. Se repiten los pasos 4 – 7, N = 1.000 veces.

En la figura 4 se puede apreciar claramente que calcular distancias al azar tiene un desempeño inferior al procedimiento basado en clústeres tanto para el MAE como para el

Figura 4.

Comparación del procedimiento de imputación con selección al azar frente a selección usando clústeres



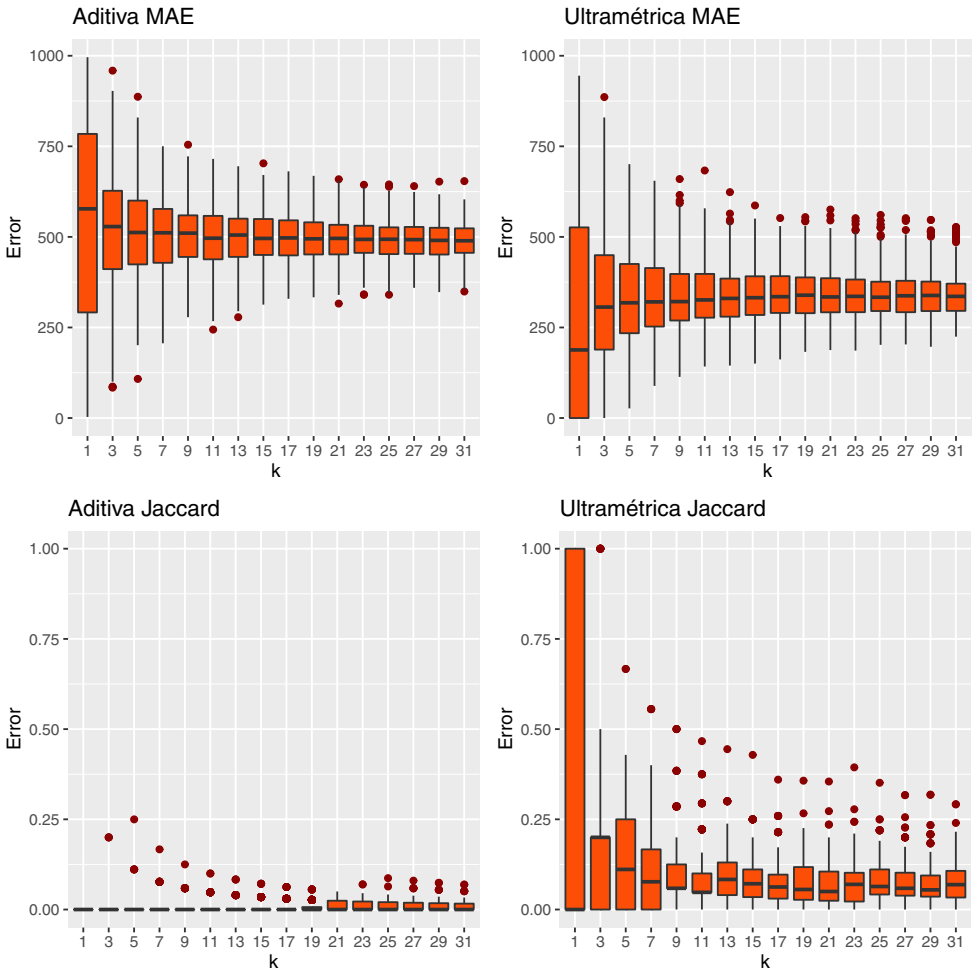
Fuente: Elaboración propia.

índice de Jaccard. También hemos realizado la comparación entre los métodos ultramétrico y aditivo, viendo claramente que el ultramétrico obtiene mejores resultados, ver figura 5.

Finalmente, se muestra una comparación entre el procedimiento ultramétrico y el procedimiento propuesto basado en clústeres, donde se puede apreciar que nuestro método tiene un mejor desempeño (ver figura 6).

Figura 5.

Comparación de los procedimientos aditivo y ultramétrico

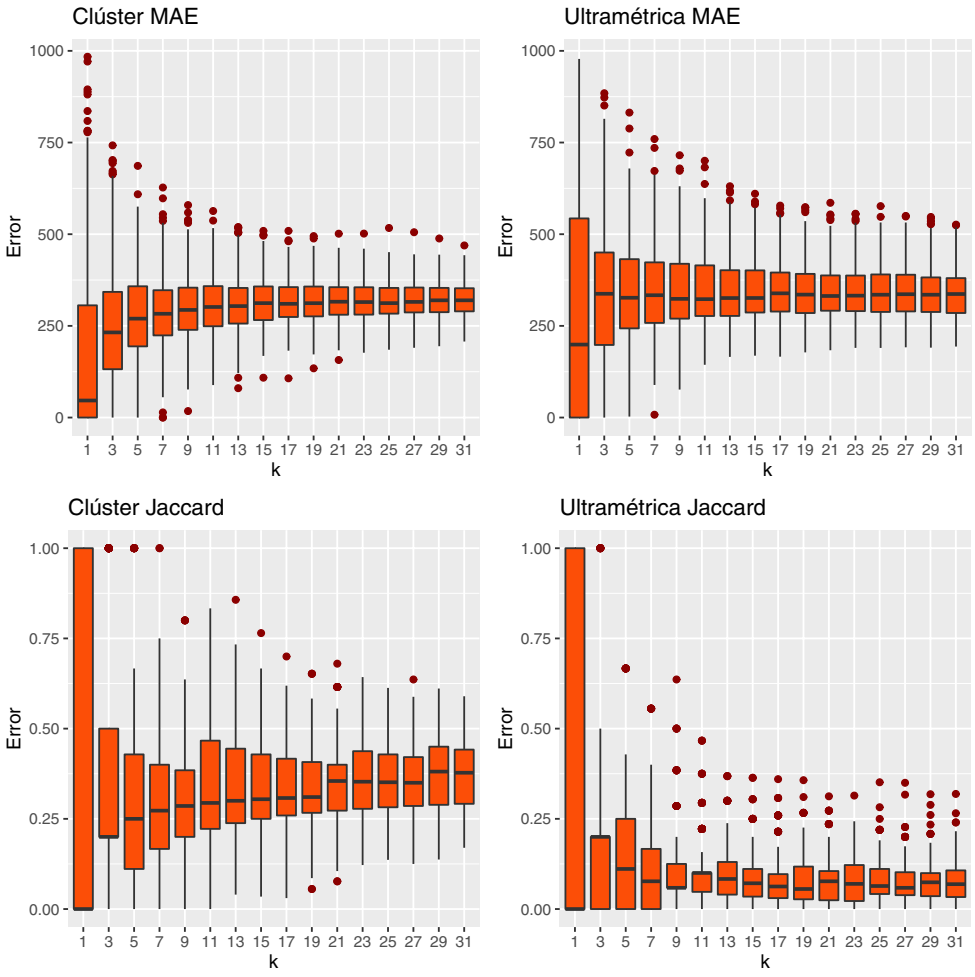


Fuente: Elaboración propia.

Se observa que, respecto al MAE, las diferencias entre ambos enfoques es pequeña. Sin embargo, analizando el índice de Jaccard se puede ver que el método propuesto es superior al resto. Al método propuesto, en adelante, lo denominaremos como "Imputación mediante clústeres".

Figura 6.

Comparación de los procedimientos ultramétrico y de imputación mediante clúster



Fuente: Elaboración propia.

3. EJEMPLOS CON CONJUNTOS DE DATOS REALES

En esta sección, ilustraremos el uso del procedimiento de imputación mediante clústeres en dos conjuntos de datos reales:

- MNIST: Es una base de datos de dígitos escritos a mano. Son 42.000 imágenes en blanco y negro, cada una con 28 pixeles de ancho y de alto.

- COEMS: Es una base de datos que contiene las curvas de oferta de electricidad horarias del mercado secundario en España en el período desde el 1 de enero de 2014 al 31 de diciembre de 2019.

3.1. MNIST

En esta sección, comprobaremos el procedimiento propuesto con el k-NN clásico usando un dataset real de clasificación. Usaremos la base de datos MNIST que consiste en imágenes en blanco y negro normalizadas cuyas dimensiones son de 28 × 28 píxeles en niveles de escala de grises y representan dígitos escritos a mano. El problema de clasificación consiste en, dada una nueva imagen, predecir qué número tiene escrito.

Figura 7.

Imágenes seleccionadas del conjunto de datos de entrenamiento del MNIST

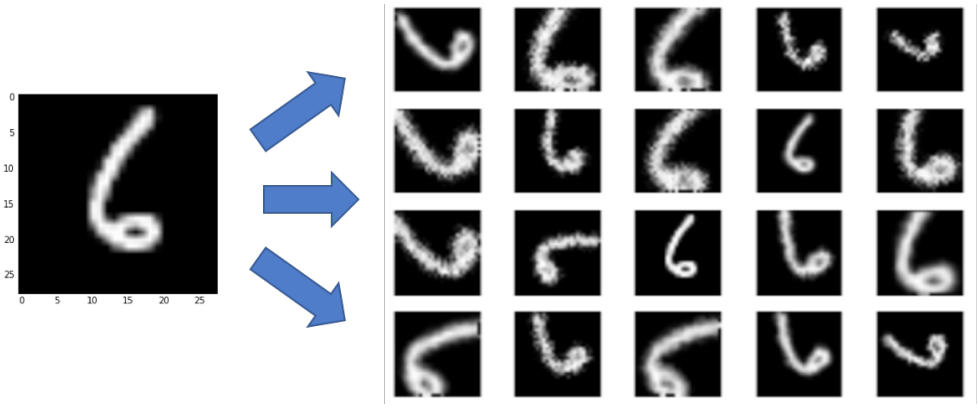
Fuente: https://commons.wikimedia.org/wiki/File:MnistExamples.png

La base de datos fue descargada de https://www.kaggle.com/c/digit-recognizer/data https://www.kaggle.com/c/digit-recognizer/data, consta de 42.000 observaciones, de las cuales dejaremos el 75 % para entrenar y validar y las restantes para probar nuestro método. Esta división se hizo usando un reparto estratificado entre las muestras de entrenamiento y de prueba (Kuhn *et al.,* 2019).

Las razones para utilizar un muestreo estratificado en lugar de un muestreo aleatorio simple son mantener una distribución similar de las etiquetas originales en el nuevo subconjunto; las mediciones se vuelven más manejables cuando la población se agrupa en estratos y, a menudo, es deseable tener estimaciones de los parámetros de la población para todos los subgrupos (Hyndman y Athanassopoulos, 2018).

Figura 8.

Ejemplo de *data augmentation* para una imagen seleccionada del MNIST

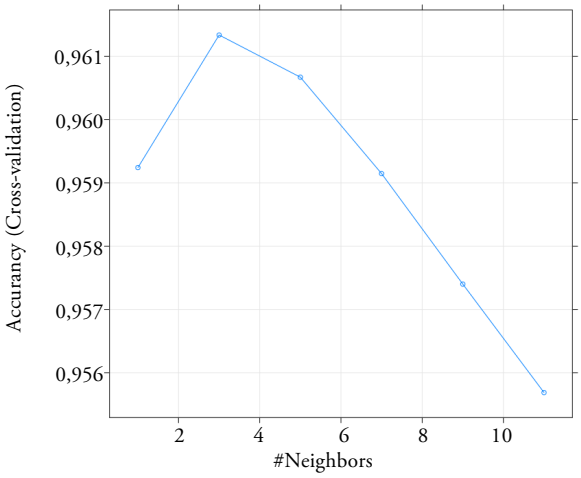


Fuente: https://www.snorkel.org/doks-theme/assets/images/2017-08-11-tanda/data_aug_basic.png

De esta manera, el conjunto de datos está balanceado con lo cual evitamos un punto débil que presenta k-NN. En caso de que fuera un conjunto de datos no balanceado se recomienda hacer *data augmentation*. Esta técnica es muy común a la hora de clasificar imágenes y no

Figura 9.

Selección del parámetro *k* del k-NN en datos de entrenamiento y validación del MNIST



Fuente: Elaboración propia.

es más que tomar una imagen y crear nuevas muestras rotando, dilatando, trasladando o agregando ruido blanco a la original, ver figura 8.

Para la búsqueda del valor óptimo de k, se utilizó el procedimiento de validación cruzada con cinco submuestras del conjunto de entrenamiento y validación donde se obtiene una precisión del 96,14 % siendo el mejor valor $k = 3$, ver figura 9.

Usando ahora este valor de k predecimos las etiquetas para los valores de la muestra de prueba usando el k-NN donde se obtiene una precisión del 96,98 % con la siguiente matriz de confusión:

Tabla 1.

Matriz de confusión del MNIST usando k-NN ($k = 3$)

	Referencia									
Predicción	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0	1.024	0	4	2	1	4	6	2	1	6
1	0	1.161	8	1	7	1	0	6	10	2
2	2	4	1.002	4	0	2	2	3	4	2
3	1	0	5	1.043	0	8	1	0	18	10
4	0	1	0	0	985	0	0	1	3	8
5	0	1	1	20	0	920	6	0	16	6
6	5	0	0	1	8	11	1.019	0	5	0
7	0	3	21	5	3	0	0	1.083	5	10
8	0	1	1	6	1	0	0	0	941	1
9	1	0	2	5	13	2	0	5	12	1.002

Fuente: Elaboración propia.

Observando la matriz anterior podemos notar que los casos donde más falla el procedimiento son en etiquetas similares, por ejemplo el 2 con el 7, y el 8 con el 3 y con el 5.

Para este ejercicio, trabajamos con $\ell = 0{,}25$, lo que significa que solo podemos calcular el 25 % de las distancias. Las restantes fueron imputadas usando el método propuesto, donde el número de clústeres es K = 3938. De esta forma se obtiene una precisión del 96,42 % y la matriz de confusión que aparece en la tabla 2.

La matriz de confusión es similar a la obtenida utilizando todos los datos (Tabla 1). Este ejercicio muestra que el procedimiento propuesto tiene un comportamiento similar al que se obtiene calculando todas las distancias.

Tabla 2.

Matriz de confusión del MNIST usando k-NN (k = 3) e imputación mediante clústeres

	Referencia									
Predicción	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0	1.022	0	9	3	2	4	11	0	4	4
1	1	1.166	10	4	10	2	1	10	11	1
2	1	1	990	7	0	0	1	6	7	0
3	0	1	2	1.030	0	7	0	0	14	3
4	1	2	2	0	971	0	2	6	4	6
5	0	0	0	14	0	920	2	0	19	2
6	6	0	4	3	7	8	1.016	0	11	1
7	2	0	21	13	4	2	0	1.067	5	12
8	0	0	3	7	0	1	1	0	923	2
9	0	1	3	6	24	4	0	11	17	1.016

Fuente: Elaboración propia.

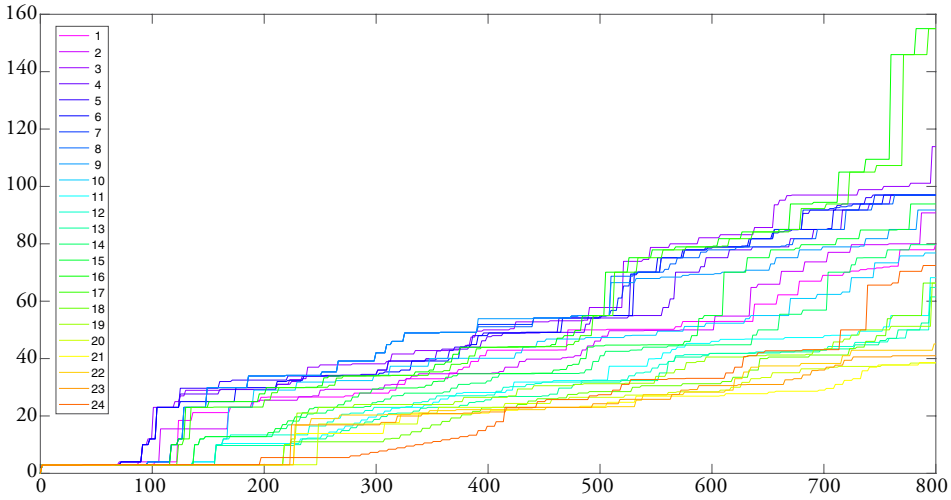
3.2. COEMS

El mercado eléctrico en España permite a los productores de electricidad ofrecer bloques de energía a diferentes precios, generalmente relacionados con sus costos marginales, en momentos específicos del día. El operador del sistema, Red Eléctrica de España (REE), recoge todas las ofertas (bloques de energía) y sus correspondientes precios de todos los participantes para formar la curva de oferta con la que se obtendrá el precio marginal de cada hora. La energía final producida por cada participante será remunerada a este precio marginal. Existen diferentes mercados: diario, intradiario, secundario y terciario. Estos mercados se diferencian tanto en el momento del día en que se realiza como en el horizonte de aplicación. Los datos están disponibles en [https://www.esios.ree.es/es/curvas-de-ofertas].

En este ejemplo estamos interesados en el mercado secundario donde el horizonte de aplicación (y por tanto, de predicción) son las 24 horas del día siguiente. Usaremos la información disponible hasta el momento antes de que se realice el mercado, por lo que podemos usar las curvas de oferta de ese mercado del día D-1, para predecir las curvas de oferta del día siguiente, D. También, en ese momento, están disponibles los precios del día D (mercado diario) y las predicciones de demanda y de producción de parques eólicos pero su incorporación en el modelo predictivo será objeto de investigación futura. A modo de ejemplo, en la figura 10 mostramos las curvas de oferta para las 24 horas del 1 de enero de 2014 en el rango de 0 a 800 MW. Utilizamos la gráfica de arco iris propuesta por Hyndman y Shan (2010), ya que nos permite darnos cuenta de que las curvas que corresponden a las horas

Figura 10.

Curvas de oferta de electricidad del mercado secundario, 1 de enero de 2014



Fuente: Elaboración propia.

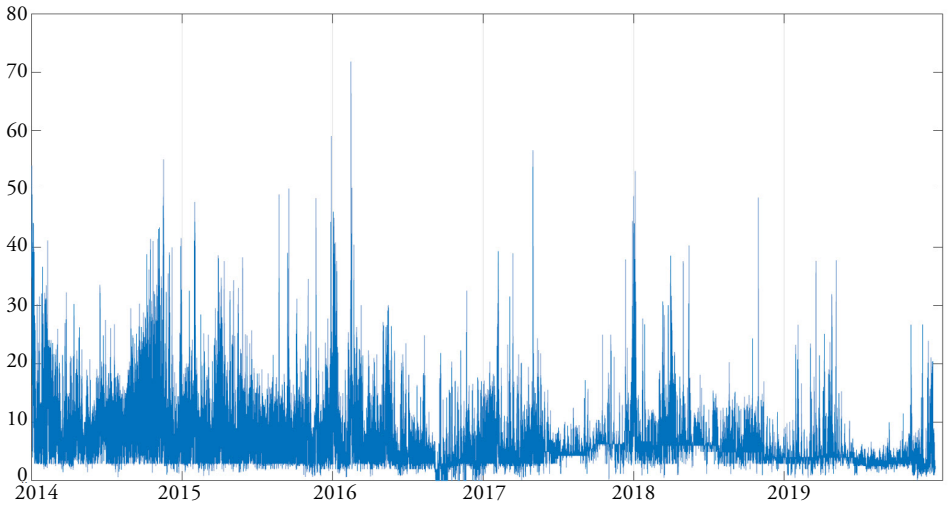
contiguas están cercanas lo que significa que los datos históricos serán útiles para predecir las curvas de oferta del día siguiente.

Estas curvas de ofertas son funciones escalonadas, no decrecientes y continuas por la izquierda. Tanto el número de escalones (ofertas con distinto precio) como la máxima oferta acumulada son dependientes del día/hora. En este ejemplo, usaremos 800 MW como valor máximo de la oferta acumulada para todas las curvas de oferta.

Otra representación interesante de estas curvas son las series de precios horarias para cantidades fijas. Es decir, si fijamos la cantidad de electricidad que queremos "comprar", qué precios tendremos que pagar por esa cantidad. Por ejemplo, supongamos que queremos obtener la serie temporal de precios al que podríamos comprar 400MW, para ello tendríamos que cortar las curvas de la figura 10 fijando el valor 400 en el eje horizontal y determinar los precios correspondientes. En la figura 11, se representa esta serie temporal. Esta representación de series de tiempo nos permite estudiar la dependencia temporal utilizando herramientas como las autocorrelaciones simples y parciales. En la figura 12, mostramos las funciones de autocorrelaciones simples de estas series temporales. Se observa que todas las series tienen una clara estacionalidad diaria pero no solo eso sino que la dependencia temporal varía en función de la cantidad a comprar. Por ejemplo, la dependencia es más fuerte en pequeñas o medianas cantidades (menos de 200MW), aumenta en cantidades intermedias (400MW a 600MW) y disminuye en cantidades grandes (más de 600MW). Estas diferencias en la dinámica sugieren una posible relación no lineal entre las curvas en D-1 y en D.

Figura 11.

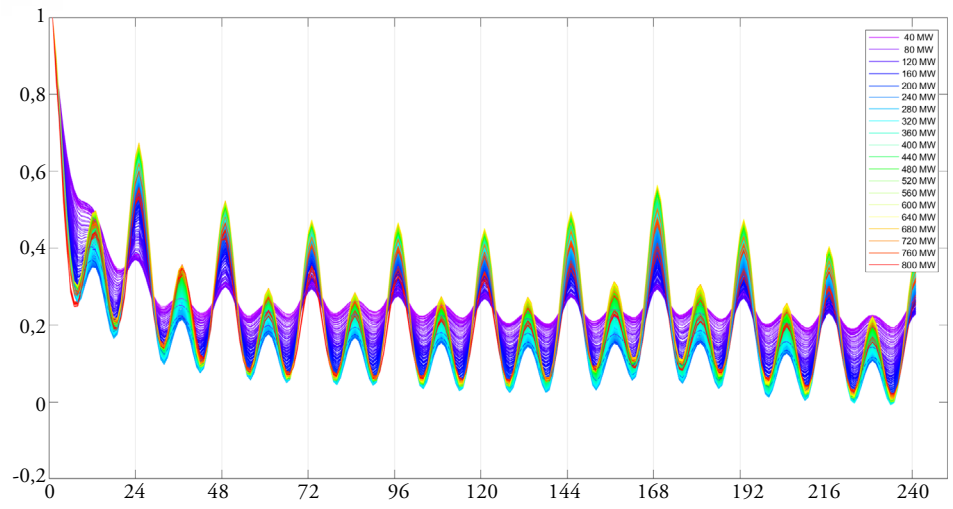
Serie temporal de los precios de 400MW en el mercado secundario, 2014-2019



Fuente: Elaboración propia.

Figura 12.

Funciones de autocorrelación simple de los precios de ofertas desde 40 MW a 800MW en el mercado secundario, 2014-2019



Fuente: Elaboración propia.

Un procedimiento simple de predicción basado en k-NN consiste en buscar días pasados que hayan tenido un comportamiento similar al día D-1 y tomar como predicciones las curvas del día siguiente de esos días similares. El procedimiento se formula como sigue:

- Dadas las 24 curvas de oferta del día D-1, $\boldsymbol{C_t} = \{C_{t-23}, C_{t-22}, ..., C_t\}$, queremos predecir las 24 curvas del día D, $\boldsymbol{C_{t+24}} = \{C_{t+1}, C_{t+2}, ..., C_{t+24}\}$.
- Sea $s^* = argmin s_{\leq t-24} d(\boldsymbol{C_t}, \boldsymbol{C_s})$, es decir, $\boldsymbol{C_{s^*}}$ son las 24 curvas consecutivas más cercanas a $\boldsymbol{C_t}$.
- La predicción se obtiene mediante $\hat{\boldsymbol{C}}_t = \{C_{s^*+1}, C_{s^*+2}, ..., C_{s^*+24}\}$.

En la formulación anterior faltaría por definir la distancia, d, entre $\boldsymbol{C_t}$ y $\boldsymbol{C_s}$. Esto nos condujo a la pregunta de cuál distancia es la importante para la predicción, d ($\boldsymbol{C_t}$, $\boldsymbol{C_s}$) o d ($\boldsymbol{C_{t+24}}$, $\hat{\boldsymbol{C}}_{t+24}$). Evidentemente, la segunda de estas distancias es más relevante porque mide el error de predicción. Sin embargo, esta distancia no se puede calcular *a priori* porque no conocemos $\boldsymbol{C_{t+24}}$ en el momento de la predicción. La cuestión es si podemos aprender de d ($\boldsymbol{C_t}$, $\boldsymbol{C_s}$) para predecir $d(\boldsymbol{C_{t+24}}, \boldsymbol{C_{s+24}})$. Notar que la predicción de $\boldsymbol{C_{t+24}}$ mediante el procedimiento k-NN será $\boldsymbol{C_{s+24}}$. En lo que sigue proponemos un procedimiento para el aprendizaje de la distancia (*distance learning*). Vamos a simplificar el problema y predecir $d(\boldsymbol{C_{t+h}}, \boldsymbol{C_{s+h}})$, es decir, la distancia entre las curvas de la hora h.

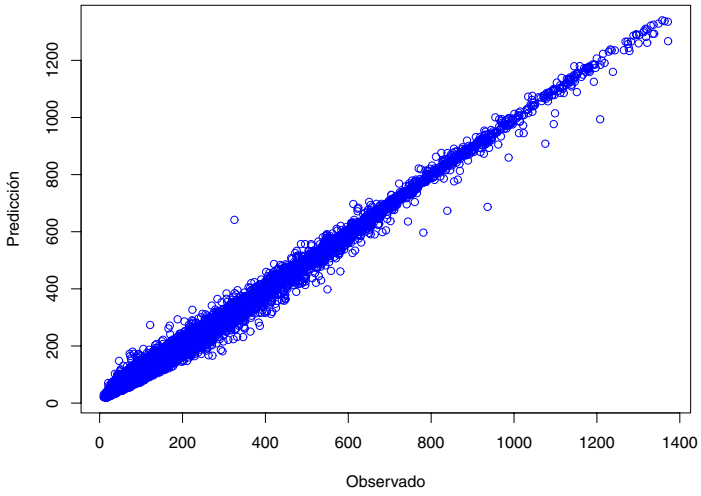
- Para dos tiempos en la muestra de entrenamiento, s y t con $s \leq t - 24$, podemos calcular las siguientes distancias:
  - $d(C_t, C_s)$, $d(C_{t-1}, C_{s-1})$ ..., $d(C_{t-23}, C_{s-23})$
  - $d(C_t, C_{s+h})$, $d(C_{t-1}, C_{s+h})$ ..., $d(C_{t-23}, C_{s+h})$
- Estas 48 variables serán el *input* de un procedimiento de bosque aleatorio (*random forest*) para predecir la distancia $d(C_{t+h}, C_{s+h})$.

La selección de las variables está motivada por la alta dependencia temporal entre las curvas tanto a corto plazo como en retardos múltiplos de 24. La elección del procedimiento de bosque aleatorio se basa en su versatilidad para modelizar relaciones no lineales e interacciones entre variables. Por simplicidad, hemos utilizado la distancia euclídea entre cada par de curvas, es decir, $d^2(C_t, C_s) = \int_0^{800} (C_t(q) - C_s(q))^2 dq$.

Para el entrenamiento del procedimiento se usan las 43.824 curvas horarias de los años 2014 a 2018 y, como conjunto de prueba, las 8.760 horas del año 2019. Se utilizan 25 réplicas bootstrap con 20000 observaciones para la selección del número de variables a incluir en el modelo. El modelo final utiliza 500 árboles y obtiene un ajuste satisfactorio en el conjunto de entrenamiento tal como se ilustra en la figura 13. Las variables más relevantes se muestran en la figura 14 que como vemos son tanto variables que captan la dependencia de corto plazo ($X25 = d(C_t, C_{s+1})$, $X1 = (d(C_t, C_s)$ y $X26 = d(C_{t-1}, C_{s+1}))$ como variables que captan la dependencia estacional diaria ($X48 = C_{t-23}, C_{s+1}$) y $X47 = d(C_{t-22}, C_{s+1})$).

Figura 13.

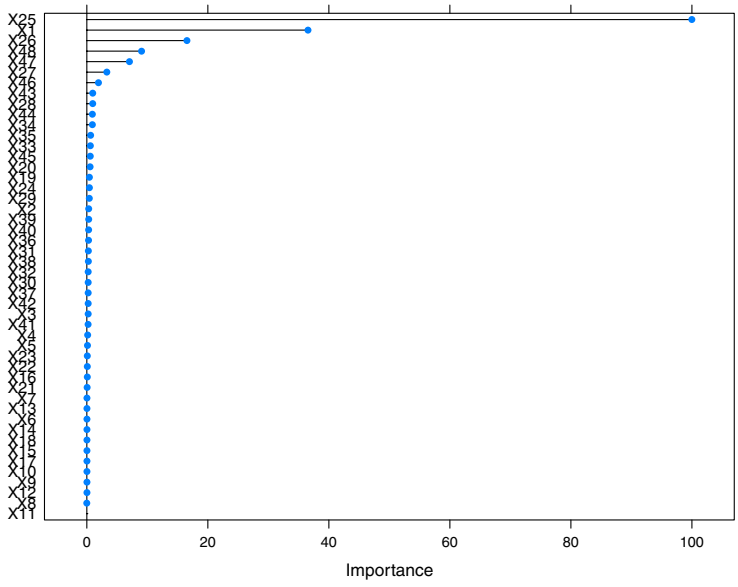
Ajuste del procedimiento de bosque aleatorio para *h* = 1, 2014-2018



Fuente: Elaboración propia.

Figura 14.

Importancia de las variables en el modelo de bosque aleatorio para *h* = 1



Fuente: Elaboración propia.

El mismo procedimiento se realiza para los distintos horizontes de predicción, $h = 1, 2, ..., 24$, es decir, se entrenan 24 modelos de bosque aleatorio. Estos modelos nos permiten realizar la predicción para el día siguiente utilizando las prediciones de las distancias para los distintos horizontes de predicción. Se utiliza como modelo de referencia aquel que elige

Tabla 3.

Medias de las distancias euclídeas entre la curva real y su predicción
(Errores estándar entre parentésis)

Método	*h = 1*	*h = 12*	*h = 24*
Referencia	88.586 (0.322)	79.068 (0.275)	73.555 (0.288)
k-NN + RF (1)	**59.245** (0.227)	87.614 (0.224)	73.347 (0.236)
k-NN + RF (12)	95.813 (0.265)	**64.543** (0.213)	87.332 (0.229)
k-NN + RF (24)	72.806 (0.224)	87.185 (0.205)	**57.811** (0.147)

Fuente: Elaboración propia.

la menor distancia euclídea agregada entre el último día disponible y días anteriores en la muestra de entrenamiento. En la tabla 3 se muestran las medias de las distancias euclídeas entre la curva real y su predicción en la muestra de prueba (8.760 horas del año 2019). En la tabla, Referencia corresponde al modelo de referencia, y k-NN + RF(h) corresponde a la combinación de k-NN con el procedimiento de bosque aleatorio entrenado para el horizonte h. Como vemos, el procedimiento propuesto mejora al modelo de referencia en todos los horizontes de predicción y se concluye que es relevante el entrenamiento para cada horizonte de predicción por separado.

Tabla 4.

Desempeño del procedimiento k-NN + RF(h) en el caso de matrices parcialmente observadas. Medias de las distancias euclídeas entre la curva real y su predicción
(Errores estándar entre parentésis)

Método	*h = 1*	*h = 12*	*h = 24*
k-NN + RF (1) ℓ = 25 %	**60.217** (0.227)	93.154 (0.239)	84.137 (0.244)
k-NN + RF (12) ℓ = 25 %	122.196 (0.312)	**64.405** (0.213)	102.161 (0.250)
k-NN + RF (24) ℓ = 25 %	87.253 (0.243)	101.508 (0.244)	**57.457** (0.150)

Fuente: Elaboración propia.

Por último, hemos repetido el ejercicio anterior para el caso de matrices parcialmente observadas, en particular, solo se calculan el 25 % ($\ell = 0{,}25$) de las distancias entre las curvas del conjunto de entrenamiento y las curvas del último día disponible. Los resultados se muestran en la tabla 4 y podemos concluir que los resultados son similares a los obtenidos usando todas las distancias. En este caso, es más evidente la importancia del entrenamiento del modelo para cada horizonte de predicción.

4. CONCLUSIONES Y EXTENSIONES

En este trabajo hemos estudiado modificaciones al procedimiento k-NN cuando no es factible calcular todas las distancias entre las nuevas observaciones y todas las observaciones en el conjunto de entrenamiento.

Hemos estudiado la combinación del algoritmo triangle fixing con varias propuestas de valores iniciales resultando que la propuesta basada en acotaciones obtiene mejores resultados y es computacionalmente factible en grandes conjuntos de datos.

Hemos comprobado que la selección de observaciones al azar en el conjunto de entrenamiento es inferior a una selección basada en clústeres tanto en índice de Jaccard como en el MAE. El procedimiento de imputación basado en clústeres también resulta superior a los procedimientos aditivos y ultramétricos que han sido propuestos en relación con la inferencia filogenética.

Finalmente, hemos utilizado el procedimiento propuesto en dos conjuntos de datos reales y hemos observado que los resultados son similares a los obtenidos con un k-NN que utiliza todas las distancias.

Como futuras líneas de investigación, tenemos:

- Implementar estos algoritmos en un lenguaje como C o C++ haciendo uso del paquete Rcpp (Eddelbuettel y Balamuta. 2018) lo que permitiría realizar experimentos en conjuntos de datos de mayor dimensión.

- Buscar un número "óptimo" de clústeres que tenga en cuenta la función de pérdida del problema a resolver con k-NN, es decir, que maximice la precisión en el caso del problema de clasificación o que minimice el error de predicción en el problema de regresión. También puede ser interesante estudiar la selección conjunta del parámetro k del k-NN y del parámetro K del procedimiento de imputación mediante clústeres.

- Estudiar extensiones al aprendizaje de distancia incorporando la componente temporal.

Referencias

Beyer, K., Goldstein, J., Ramakrishnan, R. y Shaft, U. (1999). When is "nearest neighbor" meaningful? *Database Theory- ICDT'99,* pp. 217-235. Springer Berlin Heidelberg.

Bicego, M. y Loog, M. (2016). Weighted k-nearest neighbor revisited. *23rd International Conference on Pattern Recognition (ICPR),* pp. 1642-1647. IEEE.

Brickell, J., Dhillon, I., Sra, S. y Tropp, J. A. (2008). The metric nearness problem. *SIAM Journal on Matrix Analysis and Applications,* 30, pp. 375-396.

Buneman, P. (1971). *The Recovery of Trees from Measures of Dissimilarity,* pp. 387-395. Edinburgh: University Press.

Dauben, J. W. (1990). *Georg Cantor: His Mathematics and Philosophy of the Infinite.* Princeton University Press.

De Soete, G. (1984). Additive-tree representations of incomplete dissimilarity data. *Quality and Quantity,* 18, pp. 387-393.

Dhillon, I., Sra, S. y Tropp, J. (2003). The metric nearness problem with applications. Technical report, University of Texas at Austin.

Dhillon, I., Sra, S. y Tropp, J. A. (2005). Metric nearness: Problem formulation and algorithms. *Advances in Neural Information Processing,* 17, pp. 361-368.

Eddelbuettel, D. y Balamuta, J. J. (2018). Extending R with C++: A Brief Introduction to Rcpp. *The American Statistician,* 72, pp. 28-36.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of Statistical Software,* 31, pp. 1-24.

Han, J., Kamber, M. y Tung, A. K. H. (2001). Spatial clustering methods in data mining. *Geographic data mining and knowledge discovery,* pp. 188-217.

Hartigan, J. A. (1967). Representation of similarity matrices by trees. *Journal of the American Statistical Association,* 62, pp. 1140-1158.

Heltshe, J. F. (1988). Jackknife estimate of the matching coeficient of similarity. *Biometrics,* 44, pp. 447-460.

Hyndman, R. J. y Athanasopoulos, G. (2018). *Forecasting: principles and practice. OTexts*.

Hyndman, R. J. y Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting,* 22, pp. 679-688.

Hyndman, R. J. y Shang, H. L. (2010). Rainbow plots, bagplots and boxplots for functional data. *Journal of Computational & Graphical Statistics,* 19, pp. 29-45.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. y Hunt, T. (2019). *caret: Classification and Regression Training.*

Landry, P.-A. y Lapointe, F.-J. (1997). *Estimation of missing distances in path-length matrices: Problems and solutions,* pp. 209-218.

Landry, P.-A., Lapointe, F.-J. y Kirsch, J. A. W. (1996). Estimating phylogenies from lacunose distance matrices: Additive is superior to ultrametric estimation. *Molecular Biology and Evolution,* 13, pp. 818-823.

Lapointe, F. J. y Kirsch, J. A. W. (1995). Estimating Phylogenies from Lacunose Distance Matrices, with Special Reference to DNA Hybridization Data. *Molecular Biology and Evolution,* 12, pp. 266-266.

Lapointe, F.-J. y Makarenkov, V. (2004). A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics,* 20, pp. 2113-2121.

LeCun, Y. y Cortes, C. (2010). MNIST handwritten digit database.

Paradis, E. y Schliep, K. (2018). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics,* 35, pp. 526-528.

Park, H.-S. y Jun, C.-H. (2009). A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications,* 36, pp. 3336-3341.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Real, R. y Vargas, J. (1996). The Probabilistic Basis of Jaccard's Index of Similarity. *Systematic Biology,* 45, pp. 380-385.

Rudin, W. (1991). *Functional Analysis. International series in pure and applied mathematics.* McGraw-Hill.

Schubert, E. y Rousseeuw, P. J. (2019). Faster k-medoids clustering: Improving the PAM, CLARA, and CLARANS algorithms. En: *Similarity Search and Applications,* (pp. 171-187). Springer International Publishing.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P. y Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS),* 42, pp. 1-21.

Zhang, S., Deng, Z., Cheng, D., Zong, M. y Zhu, X. (2016). Eficient kNN Classification Algorithm for Big Data. *Neurocomputing,* 195, pp. 143-148.

APÉNDICE

Definiciones básicas

Definición 5.1. *Para un conjunto de elementos, $\mathcal{X}$, una distancia o métrica es cualquier función $d(a, b) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ donde $\mathbb{R}$ es el conjunto de los números reales, que verifique las siguientes condiciones (ver en Rudin, 1991):*

- No negatividad: $\forall a,b \in X : d(a,b) \geq 0$.
- Coincidencia: $d(a,b) = 0 \Leftrightarrow a = b$.
- Simetría: $\forall a,b \in X : d(a,b) = d(b,a)$.
- Desigualdad triangular: $\forall a, b, c \in X : d(a,b) \leq d(a,b) + d(c,b)$.

Definición 5.2. *Sea $\boldsymbol{M} = \{m_{ij}\}$ una matriz $n \times m$, se dice no negativa cuando todas sus entradas son no negativas, es decir $m_{ij} \geq 0$ $\forall i,j$.*

Definición 5.3. *Sea $\boldsymbol{M}$ una matriz $n \times n$, se dice que es matriz de disimilitud si es una matriz no negativa, simétrica y con ceros en su diagonal principal (Dhillon, Sra y Tropp, 2003).*

Definición 5.4. *Sea $\boldsymbol{M} = \{m_{ij}\}$ una matriz $n \times n$, se dice que es una matriz de distancias, si es una matriz de disimilitud cuyas entradas satisfacen la desigualdad triangular. Esto es, $\boldsymbol{M}$, es matriz de distancias si y solo si para toda terna de índices (i, j, k) se verifica que (Dhillon, Sra y Tropp, 2005):*

$$m_{ik} \leq m_{ij} + m_{jk}.$$

Definición 5.5. *Una matriz de distancias, $\boldsymbol{M} = \{m_{ij}\}$, se dice aditiva, cuando sus entradas verifican tanto la desigualdad triangular como la cuadrangular (Buneman, 1971):*

$$m_{ij} + m_{kl} \leq \text{máx}[m_{ik} + m_{jl}; m_{il} + m_{jk}] \qquad \forall i, j, k, l.$$

Definición 5.6. *Una matriz de distancias, $\boldsymbol{M} = \{m_{ij}\}$, se dice ultramétrica, cuando es aditiva y sus entradas verifican la desigualdad ultramétrica (Hartigan, 1967):*

$$m_{ij} \leq \text{máx}[m_{ik}; m_{jk}] \qquad \forall i, j, k.$$

Definición 5.7. *Se llama cardinalidad al número de elementos de un conjunto A y se denota por $|A|$ (Dauben, 1990).*

Definición 5.8. *El índice o coeficiente de Jaccard entre dos conjuntos A y B se define como la cardinalidad de la intersección de ambos conjuntos dividida por la cardinalidad de su unión (Real y Vargas, 1996):*

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{M_{11}}{M_{01}+M_{10}+M_{11}}, \qquad [14]$$

donde, M_{11} representa el número total de casos donde A y B tienen un valor de 1, M_{01} el número total de casos donde el valor de A es 0 y el valor de B es 1, M_{10} el número total de casos donde el valor de A es 1 y el valor de B es 0 y M_{00} representa el número total de casos donde A y B tienen un valor de 0.

El índice de Jaccard siempre toma valores entre 0 y 1, correspondiendo este último a la igualdad entre ambos conjuntos. Existe otra medida que se llama el coeficiente de *coincidencia simple* (SMC) (Heltshe, 1988) o *coeficiente de similitud de Rand* que es muy similar al Jaccard pues se define como:

$$SMC=\frac{M_{00}+M_{11}}{M_{01}+M_{10}+M_{11}},$$

pero en este trabajo preferimos usar el índice de Jaccard porque nos centramos en la coincidencia de una de las categorías de los conjuntos *A* y *B*.

Definición 5.9. *Se define el error absoluto medio (MAE), como* (Hyndman y Koehler, 2006):

$$MAE=\frac{\sum_{i=1}^{k}|\widehat{O}_i-O_i|}{k},$$

donde $\widehat{O}$ son los valores estimados y O_i los valores reales u observados, siendo k la cantidad de valores a estimar.

Definición 5.10. *Sea* $\mathbf{O}$ *la matriz original y* $\mathbf{P}$ *la matriz pronosticada, se define la diferencia relativa entre estas matrices como:*

$$\frac{\|\mathbf{O}-\mathbf{P}\|_2}{\|\mathbf{O}\|_2},$$

donde $\|\boldsymbol{A}\|_2=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$.

Definición 5.11. *Sea o el vector original y p el vector pronosticado, se define la diferencia relativa entre estos vectores como:*

$$\frac{\|o-p\|_2}{\|o\|_2},$$

donde $\|o\|_2=\sqrt{\sum_{i=1}^{n}o_i^2}$.

Sobre los autores

Andrés M. Alonso Fernández

Es licenciado en Matemáticas por la Universidad de La Habana (1991), máster en Epidemiología por el Instituto Pedro Kourí (1994) y doctor en Economía por la Universidad Carlos III de Madrid (2001). Ha sido profesor asociado en el Departamento de Matemáticas de la Universidad Autónoma de Madrid e investigador Juan de La Cierva en el Departamento de Estadística de la Universidad Carlos III de Madrid. Actualmente es profesor titular de Estadística y director del Instituto Flores de Lemus en esta universidad. Sus principales intereses de investigación son: análisis de series de tiempo, técnicas de remuestreo, estadística aplicada y econometría. Ha publicado más de 70 artículos de investigación, una monografía y un libro sobre estas temáticas.

Stevenson Bolívar Atuesta

Es profesor asistente del Departamento de Ingeniería Industrial de la Pontificia Universidad Javeriana. Ingeniero Industrial de la Universidad Tecnológica de Pereira (2005), magister en Ingeniería de la Universidad de los Andes (2007) y doctor en Ciencias-Estadística de la Universidad Nacional de Colombia, su área de investigación es series temporales, en particular análisis de series temporales multivariadas y series de tiempo no lineales.

Francisco Corona

Es licenciado en Economía por la Universidad Autónoma de Baja California, maestro en Estadística Aplicada por el Tecnológico de Monterrey y doctor en Economía de la Empresa y Métodos Cuantitativos por la Universidad Carlos III de Madrid. En la actualidad es Investigador del Instituto Nacional de Estadística y Geografía. Sus líneas de investigación son econometría, estadística aplicada y analítica del deporte. Ha publicado en revistas con arbitraje y de circulación internacional, como son *International Journal of Forecasting, Computational Economics, Journal of Applied Statistics, Empirical Economics, Journal of Official Statistics, Latin American Economic Review,* entre otras.

Carlos Cuerpo Caballero

Es licenciado en Economía por la Universidad de Extremadura, máster en Economía por la London School of Economics y doctor en Economía por la Universidad Autónoma de Madrid. Desde 2007 pertenece al Cuerpo Superior de Técnicos Comerciales y Economistas del Estado. Actualmente ejerce como director general de Análisis Macroeconómico en el Ministerio de Asuntos Económicos y Transformación Digital. En etapas anteriores, ha desarrollado su carrera como analista en instituciones como la Comisión Europea o la Autoridad Independiente de Responsabilidad Fiscal (AIReF).

Juan J. Dolado

Es doctor (D Phil) en Economía (Universidad de Oxford, 1988). Catedrático de Fundamentos del Análisis Económico en el Dpto. de Economía de UC3M. Anteriormente fue Lecturer en la Universidad de Oxford (1988-89), Economista-Jefe en la División de Estudios Cuantitativos del Servicio de Estudios del Banco de España (1990-1997), y catedrático de Economía en el European University Institute (EUI) de Florencia (2014-2019). También ha sido presidente de la Spanish Economic Association, miembro del comité ejecutivo de la European Economic Association, y director del programa de Economía Laboral del Centre for Economic Policy Research (CEPR). Sus principales campos de investigación son Econometría, Economía Laboral y Macroeconomía, donde ha publicado/editado 11 volúmenes y cerca de 150 artículos en revistas académicas internacionales y nacionales. Ha sido codirector de las revistas *Econometric Theory, European Economic Review* y *Labour Economics.* Premio Vanguardia de la Ciencia 2011 y Premio Jaime I en Economía 2015 por sus contribuciones a la investigación en economía laboral y econometría. Página web: http://dolado.blogspot.com

Aldo R. Franco Comas

Es licenciado en Matemáticas por la Universidad de La Habana (2014), y máster en Métodos Analíticos para Datos Masivos: Big Data (2017) y en Ingeniería Matemática (2019) por la Universidad Carlos III de Madrid. Ha sido estadístico en la Oficina Nacional de Estadística, Cuba. Actualmente, es científico de datos en Network Centric Software. Sus temas de investigación son las técnicas de aprendizaje automática, redes neuronales y análisis de datos.

José García Montalvo (Universitat Pompeu Fabra).

Doctor en Economía por la Universidad de Harvard y catedrático de Economía en la Universitat Pompeu Fabra (UPF). Es también profesor investigador de la Barcelona Graduate School of Economics y del IVIE. Entre otros galardones ha recibido el Primer Premio Nacional Fin de Carrera, la distinción ICREA Acadèmia, el premio a la Transferencia de Conocimiento del Consejo Social de la UPF, el Premio Cataluña Economía y el Premio Jaime I de Economía. Ha sido consultor de la Organización para la Cooperación y el Desarrollo (OCDE), del Banco Mundial, la Unión Europea y del Banco Interamericano de Desarrollo entre otras instituciones y empresas. Ha publicado diecisiete libros y más de cien artículos en revistas como *American Economic Review, Review of Economics and Statistics, Economic Journal, Journal of Business and Economic Statistics, Journal of Economic Growth* y *Applied Psychology,* entre otras.

Víctor M. Guerrero

Tiene los grados de actuario por la Universidad Nacional Autónoma de México y de maestría en Ciencias y doctor en Estadística por la Universidad de Wisconsin-Madison. Actualmente es profesor emérito, en el Departamento de Estadística del Instituto Tecnológico Autónomo de México. Sus intereses de investigación incluyen el Análisis y Pronóstico de Series de Tiempo, así como la Econometría. Sus artículos han aparecido en revistas con arbitraje y de circulación internacional, como son *Biometrika, Journal of Forecasting, International Journal of Forecasting, International Statistical Review, Journal of Applied Statistics* y *Journal of Official Statistics,* entre otras.

Juan Antonio Mendoza

Es licenciado en Economía por el Instituto Tecnológico de México (ITAM) y maestro en Matemáticas Financieras por la Universidad de Chicago. Actualmente se desempeña como director de Analítica en Grupo Financiero Banorte y sus intereses de investigación se concentran en series de tiempo financieras aplicadas al manejo de portafolios de inversión

Teresa Morales Gómez-Luengo

Es Associate Manager en everis, NTT DATA Company, en proyectos de transformación digital del Sector Público a través de Analítica Avanzada. Técnico comercial y economista del Estado en excedencia, ha desarrollado su carrera profesional en la administración en la Subdirección General de Gestión de la Deuda Pública de la Secretaría General del Tesoro (2014-2016) y en el Gabinete de la Vicepresidencia del Gobierno (2012-2014, 2016-2018). Es licenciada en Economía y Derecho por la Universidad Carlos III de Madrid y máster en Políticas Públicas con especialidad en Data Analytics por la Harris School of Public Policy de la Universidad de Chicago. También ha sido asistente de investigación en la Universidad de Chicago.

Fabio H. Nieto

Es doctor en Ciencias-Estadística de la Universidad Nacional de Colombia, en Bogotá, Colombia. Se desempeñó como profesor de Series Temporales en el Departamento de Estadística de la Universidad Nacional de Colombia, en Bogotá, durante 30 años aproximadamente. Sus cursos trataron temas de Series Temporales Univariadas, Multivariadas y No Lineales. Fue asesor del Banco de La República, el banco central de Colombia, en problemas de predicción *ex post* y *ex ante* de series temporales univariadas y multivariadas no observables y en construcción de índices coincidentes y líderes para la economía colombiana. Ha realizado investigación dentro del contexto de las series temporales, entre otras, en las áreas de estimación de datos faltantes, desagregación de series temporales univariadas y multivariadas, construcción de índices coincidentes, modelos no lineales TAR para series temporales y factores comunes dinámicos estacionales. Actualmente es profesor jubilado de la Universidad Nacional de Colombia.

Daniel Peña

Es profesor emérito de la Universidad Carlos III de Madrid (UC3M) donde ha sido catedrático desde 1990. Antes fue profesor en las Universidades Wisconsin-Madison, Chicago

y Politécnica de Madrid. Autor de 16 libros y más de 250 artículos de investigación sobre Estadística, Econometría y sus aplicaciones. Su investigación ha recibido varios premios, como el Youden Prize al mejor artículo publicado en *Technometrics* en 2006, el Premio a la carrera profesional de la Asociación de Antiguos Alumnos de la ETSII, UPM en 2010, el Premio Jaime I de investigación en Economía en 2011, el Premio Ingeniero del año en 2011, por el Colegio Oficial de Ingenieros Industriales de Madrid, la medalla de honor de la Sociedad Estadística e Investigación Operativa en 2014, la medalla de honor de la UC3M en 2015 y el Premio Nacional de Estadística de España en 2020. Es Fellow de The American Statistical Association, The Institute of Mathematical Statistics, The Royal Statistical Society y miembro electo del International Statistical Institute.

Pilar Poncela

Es catedrática de Fundamentos del Análisis Económico en la Universidad Autónoma de Madrid. Ingeniero Industrial por la Universidad de Valladolid y doctor en Economía por la Universidad Carlos III de Madrid, es actualmente miembro del comité de dirección del International Institute of Forecasters (IIF) y directora del Máster en Análisis Económico Cuantitativo de la Universidad Autónoma de Madrid. Es Fellow del Instituto de Big Data UC3M-Santander. Durante 2016-2019 estuvo trabajando como investigador sénior en la Comisión Europea, Centro Común de Investigación (DG Joint Research Centre, Ispra, Italia) en la unión del mercado de capitales. Con anterioridad realizó estancias de investigación en el Banco de España, la Universidad de Chicago (EEUU), Cass Business School (Reino Unido), ITAM (México) y Universidad Nacional de Colombia, entre otros. Sus principales campos de investigación son la predicción, el análisis de series temporales y la transmisión de "shocks" y ha publicado sus trabajos en revistas tales como *Journal of Business and Economic Statistics, Journal of Econometrics, Journal of Applied Econometrics, International Journal of Forecasting, Applied Energy, Advances in Data Analysis and Classification* y *Signal Processing,* entre otras.

Enrique M. Quilis

Es doctor en Economía por la Universidad Autónoma de Madrid (UAM) y pertenece al cuerpo de Estadísticos Superiores del Estado. Desarrolla su actividad en el Servicio de Estudios y Estadísticas de la Agencia Tributaria. Sus principales tareas se centran en la modelización y análisis de series diarias de origen tributario y en el desarrollo de modelos econométricos orientados a la predicción en tiempo real y al seguimiento de agregados macroeconómicos y tributarios. Con anterioridad ha desempeñado diversos cargos en la Autoridad Independiente de Responsabilidad Fiscal, en la D.G. de Análisis Macroeconómico, en la D.G. del Tesoro y en el Instituto Nacional de Estadística. Ha impartido clases de macroeconomía en el ICADE y en la UAM y tiene experiencia de consultoría nacional e internacional en temas de análisis macroeconómico cuantitativo y series temporales, con publicaciones tanto nacionales como internacionales en estas áreas.

Eva Senra Díaz

Es licenciada en Ciencias Matemáticas por la Universidad Complutense de Madrid y Doctora en Economía por la Universidad Carlos III de Madrid. Es vicerrectora de Economía,

Emprendimiento y Empleabilidad de la Universidad de Alcalá y profesora titular de Métodos Cuantitativos para la Economía y la Empresa del Departamento de Economía. Su trayectoria de investigación está relacionada con la predicción económica, la estadística y la economía aplicada tanto a nivel académico como profesional. En la actualidad es miembro del Consejo Asesor de AIReF, colabora con la Fundación COTEC para la innovación formando parte de la red "Los 100 de COTEC" en calidad de experta en estadística y predicción económica y participa además en el proyecto EURITO, para la creación de nuevos indicadores basados en Big Data, financiado por el programa H2020.

Funcas

Caballero de Gracia, 28
28013 Madrid
Teléfono: 91 596 54 81
Fax: 91 596 57 96

publica@funcas.es
www.funcas.es

ISBN 978-84-17609-54-2