Técnicas estadísticas para analizar la evolución del *e-commerce* en Galicia

Cheyenne Amoroso Sanmiguel

13 de junio de 2022

Este documento incluye el resumen del Trabajo Fin de Máster *Técnicas estadésticas para analizar la evolución del e-commerce en Galicia* dirigido por don Guillermo López Taboada, Profesor Titular de la Universidade da Coruña, doña Teresa Veiga Rodríguez, Especialista de Planificación y Estudios de ABANCA, y don Sergio Díaz Canosa, Especialista de Planificación y Estudios de ABANCA. Por motivos de confidencialidad no es posible la publicación de la memoria completa.

Resumen

El presente proyecto se ha llevado a cabo en el área de Planificación Estratégica y PMO (*Project Management Office*) de ABANCA Corporación Bancaria S.A. que aúna las funciones de planificación estratégica, gobierno del dato y seguimiento de proyectos, asegurando la coordinación entre los mismos y los objetivos estratégicos de la entidad. Lidera la definición de la estrategia del Banco con modelos robustos de previsiones y proyección financiera, y garantiza elevados niveles en su ejecución, apalancados en alineamiento, foco estratégico y control de ejecución de las metas establecidas.

El análisis y seguimiento del entorno macroeconómico es uno de los pilares básicos del proceso de planificación estratégica desarrollado en la entidad. Con carácter permanente, desde la Dirección General de Planificación Estratégica y PMO, se efectúa un estrecho seguimiento de las principales variables de entorno focalizado en España y en Galicia, por ser esta última la principal área de desarrollo de la actividad de ABANCA.

A raíz de la pandemia del COVID-19 surge la necesidad de analizar la situación económica en tiempo casi real para cuantificar impactos y la toma de decisiones. Los indicadores tradicionales de frecuencia mensual o trimestral son insuficientes y es necesario recurrir a datos de alta frecuencia. A lo largo del año 2020 surgen múltiples iniciativas desde distintas entidades

con el objetivo de estudiar la evolución de la economía en un escenario de gran volatilidad e incertidumbre. En este ámbito y con foco en Galicia, se crea el Observatorio ABANCA como un indicador para analizar el consumo y actividad en Galicia a partir de la operatoria con tarjetas de los clientes de la entidad en establecimientos gallegos.

En este contexto extraordinario, los hábitos económicos respecto a los medios de pagos se ven alterados postulando dos nuevas tendencias: el crecimiento del *e-commerce* y la elección de la tarjeta como medio de pago principal. Motivado por el notable aumento del protagonismo de la operatoria online sobre los pagos con tarjeta, surge la intención de generar un estudio específico del *e-commerce* que complemente al Observatorio de evolución de gasto presencial de forma que se pueda relizar la comparativa entre ambos, siendo este el objetivo principal del presente trabajo. La operatoria online no siempre se encuentra bien identificada o clasificada, por ello la primer línea de trabajo de esta memoria consiste en desarrollar un método sencillo y eficiente que permita la clasificación sectorial según la actividad comercial de los *e-commerce* por medio de técnicas clásicas del *Text Mining*.

Debido al gran tamaño del conjunto de datos a tratar de nombres de comercio es necesario llevar a cabo una primera etapa en la que se busca una reducción de la dimensión mediante el uso de técnicas clásicas en el ámbito de la minería de textos. En primer lugar, se debe almacenar el texto considerado de forma que se pueda tratar y acceder a las variables con sencillez. Con este fin se recurre al procedimiento de tokenización de los distintos nombres de comercios que se encuentran en la base de datos. Una vez conseguida la estructura de datos deseada se recurre a la técnica de lematización y eliminación de patrones, *stopwords*, o elementos que no aporten información relevante, consiguiendo así una notable redución de la dimensión. La segunda etapa consiste en la clasificación de los comercios, que pasa por la creación de dos diccionarios, uno de marcas comerciales y otro de palabras clave o *keywords*, que permiten clasificar los comercios de forma sencilla mediante expresiones regulares.

Con el método propuesto se logra clasificar un alto porcentaje de los comercios considerados. La clasificación de los comercios procedentes de la operatoria online permite a la entidad realizar un seguimiento de los niveles de consumo en Galicia sectorialmente para analizar la evolución de los patrones de consumo de los clientes, así como el protagonismo que adquiere cada uno de los sectores. Además de la obtención de la clasificación de los distintos comercios presentes en la base de datos se han establecido nuevas categorías que tienen cada vez más peso en el gasto de los clientes y que previamente no se consideraban, permitiendo así el análisis de las mismas.

El principal objetivo del presente trabajo es el análisis de los pagos diarios con tarjeta a fin de detectar si se han producido cambios en los patrones de compra de la sociedad gallega y que esto sirva de soporte para la toma de decisiones de la entidad. Con el fin de analizar la evolución de una variable temporal y poder detectar cambios es habitual considerar distintas medidas que permitan la comparación de la variable en distintos instantes. Resulta entonces de gran

importancia una correcta y adecuada interpretación de los datos, dado que en caso contrario podría llegarse a conclusiones erróneas.

Los datos de carácter socioeconómico son una gran herramienta para el análisis del ciclo económico. Sin embargo, estos pueden venir influenciados por fluctuaciones estacionales o efectos propios del calendario que pueden enmascarar movimientos a corto y largo plazo e impedir una comprensión clara de los fenómenos subyacentes de una serie temporal. El heladero que festeje haber tenido un incremento de ventas en julio respecto a enero estará sacando conclusiones erróneas, ya que para poder comparar las ventas de enero frente a las de julio es necesario extraer el efecto que tiene el verano sobre las ventas de helados. De igual modo, si se pretende analizar la variación del gasto en comercios, se debe tener en cuenta aquellos períodos o festividades puntuales que produzcan una variación en el gasto. Por ejemplo, carece de sentido comparar el gasto en comercio presencial producido en un día laborable frente a un domingo en el cual el comercio permanece cerrado o el día correspondiente al *Black Friday* frente al mismo día del mes anterior.

Por un lado, se consideran fluctuaciones estacionales aquellos movimientos que ocurren con intensidad similar en un mismo momento del año (mes, trimestre...) y que se espera que sigan ocurriendo a largo plazo. Dichas fluctuaciones se recogen en lo que se denomina componente estacional. A la hora de estudiar el comportamiento de una variable temporal con frecuencia diaria pueden darse tres tipos de estacionalidades: semanal, mensual y anual. Dentro de la estacionalidad semanal se recoge la dependencia del gasto frente al día de la semana que se considere, por ejemplo, el cierre comercial los domingos comentado anteriormente. La componente estacional mensual considera las fluctuaciones debidas al día del mes en el que se realiza la operación, por ejemplo, en el sector de alimentación es habitual realizar un gasto mayor al principio del mes. Por último, en la estacionalidad anual se recogen aquellas variaciones debidas al día del año, en el caso de los el nivel del gasto en agencias de viajes y hoteles en verano estará muy por encima de los restantes meses.

Por otro lado, los efectos de calendario pueden definirse como el impacto que se produce sobre la serie debido a la propia composición del calendario. Al tratar con series de frecuencia diaria gran parte de estos efectos se incluyen en la componente estacional. Sin embargo, existen algunas excepciones como la presencia de festividades móviles. Un ejemplo de esto es la Semana Santa, no es posible comparar la Semana Santa de un año frente a la misma semana el año anterior debido a la presencia de festividades que no ocurren de igual forma en ambos años.

El ajuste estacional y de efectos de calendario consiste en detectar y eliminar esta clase de efectos de la serie a tratar para así lograr entender la evolución de la serie y su comportamiento a largo plazo. Esto permite comprobar con mayor rigurosidad si se han dado cambios estructurales reseñables en el período de tiempo estudiado, así como comparar distintos instantes temporales siendo la comparativa homogénea. Actualmente, no existe ninguna recomendación oficial sobre

qué metodología se debe seguir para el ajuste estacional y de calendario de series temporales con observaciones diarias (Eurostat, 2015). El procedimiento a seguir debe ser suficientemente flexible para lograr estimar todos los patrones estacionales con diferentes periodicidades presentes en la serie. La segunda línea de trabajo consiste en la obtención de las series corregidas de estacionalidades y efectos de calendario. Para ello se exponen dos metodologías que se aplican sobre los datos de interés con el objetivo de compararlas y decantarse por una de ellas. La implementación de cada una de las metodologías se realiza a través del software de programación R Core Team (2013).

La primera de las metodologías tratadas es la metodología DSA (*Daily Seasonal Adjustment*) expuesta en Ollech (2018) que combina una rutina iterativa de ajuste estacional basada en la descomposición STL (Cleveland et al., 1990) con un modelo RegARIMA (Box et al., 2015) para la estimación de efectos de calendario y valores atípicos. Dicha metodología es la que sigue el Banco Federal Alemán para la corrección de distintas series que emplean en la construcción de un índice semanal (Eraslan y Götz, 2021). Entre las series utilizadas se encuentran los pagos diarios con tarjetas de crédito, lo que motiva a emplear esta metodología con los datos provistos por la entidad.

En segundo lugar, se considera la metodología expuesta por la Agencia Tributaria (AEAT) en Cuevas et al. (2021) donde se presenta una alternativa flexible que permite trata el complejo comportamiento de series de alta frecuencia. Para ello se recurre al modelo estructural propuesto por De Livera et al. (2010) llamado TBATS (*Trigonometric seasonality, Box-Cox, ARMA, Trend, Seasonality*) complementado con un tratamiento preliminar de los efectos de calendario a través de un modelo de regresión dinámica.

Los datos empleados en el desarrollo de la memoria son los referentes a los importes totales de la operatoria online y presencial desde enero de 2019 hasta abril de 2022. Con el objetivo de evaluar la capacidad predictiva de los modelos obtenidos con cada una de las metodologías tratadas se dividirá la muestra en dos conjuntos, uno de entrenamiento y otro de test, donde el conjunto de entrenamiento se compone por los datos correspondientes a los años 2019, 2020 y 2021. Debido al efecto de propensión al pago con tarjeta en lugar de efectivo producido a raíz de la crisis del COVID-19, se toman los importes totales referentes a los pagos con tarjeta realizados en comercios presenciales corregidos ajustando la evolución del volumen de transacciones a partir del 14 de marzo de 2020. Tras comparar los modelos obtenidos con cada una de las metodologías tratadas a través de una serie de medidas de error que permiten la selección de una de ellas por medio de la comparación de los ajustes y la capacidad predictiva, se selecciona la metodología DSA para la obtención de las series corregidas. Los resultados expuestos en la memoria se obtienen de las series totales; sin embargo, ambas metodologías se han aplicado sobre las series correspondientes a sectores de interés.

Con el objetivo de analizar la evolución diaria del gasto en *e-commerce* se desarrolla un índice

simple tomando como día base el promedio del gasto diario durante la etapa pre-covid del año 2020 (hasta el 13 de marzo). Al considerar como base este período resulta indispensable tomar la serie corregida de estacionalidad y efectos de calendario dado que, en caso contrario, estaríamos cayendo en el error de comparar dos períodos de tiempo afectados por distintas fluctuaciones. Además, se analiza el peso de algunos sectores de interés en la cesta de la compra online de los clientes junto a la evolución del protagonismo del *e-commerce* sobre algunos sectores de interés mediante el peso de la compra online sobre el gasto total.

Alcanzar los dos objetivos propuestos para el desarrollo de este trabajo permite a la entidad establecer nuevas líneas de trabajo que no habrían sido posibles en caso contrario. En primer lugar, la clasificación del *e-commerce* permite a la entidad optimizar tanto sus procesos internos como su propuesta de valor a los distintos segmentos de clientes, a la vez de poder analizar los perfiles y hábitos de consumo de sus clientes y la situación económica.

En segundo lugar, las dos metodologías expuestas a lo largo de este trabajo para el tratamiento de series diarias que presentan múltiples estacionalidades y efectos de calendario resultan de utilidad no solo para la corrección de las series temporales de importes de pagos con tarjeta, sino también para la corrección de otras series de alta frecuencia que resultan de interés para la entidad. En particular, la entidad está interesada en el desarrollo de un indicador de alta frecuencia que permita rastrear la evolución económica en tiempo casi real. Dado que los principales agregados macroeconómicos acostumbran a trabajarse corregidos de estacionalidad y efectos de calendario, es necesario que el indicador desarrollado se encuentre también corregido de dichos efectos a fin de estudiar la evolución económica de manera similar y que, además, pueda establecerse una relación entre los distintos indicadores.

Además del desarrollo de un indicador para el análisis de la evolución económica nacional, ABANCA también busca disponer de un indicador equivalente para el seguimiento de la actividad económica en Galicia. Para ello resulta de utilidad el uso de las series trabajadas a lo largo de esta memoria como un indicador de consumo en la comunidad.

Finalmente, además de las dos metodologías expuestas en esta memoria, recientemente han sido desarrolladas otras metodologías de interés con el mismo objetivo de tratar series de alta frecuencia. En particular, en trabajos futuros, se pretende aplicar sobre las series consideradas en este trabajo otras dos nuevas metodologías a fin de comparar los resultados obtenidos con los expuestos en esta memoria. La primera de las metodologías que se pretende comparar es la denominada Prophet (Taylor y Letham, 2017), desarrollada por Facebook. En segundo lugar, se propone considerar una versión de la descomposición STL que permite tratar múltiples estacionalidades (Bandara et al., 2021), MSTL.

Referencias

Bandara, K., Hyndman, R. J., y Bergmeir, C. (2021). Mstl: A seasonal-trend decomposition algorithm for time series with multiple seasonal patterns.

Box, G. E. P., Jenkins, Gwilym M., J., Reinsel, G. C., y Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control, 5th Edition*. John Wiley and Sons.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., y Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess (with discussion). *Journal of Official Statistics*, 6:3–73. https://www.bibsonomy.org/bibtex/24bf4893a61f6e30b2dbf7f37884295ed/jwbowers.

Cuevas, A., Ledo, R., y Quilis, E. (2021). Seasonal adjustment of the spanish sales daily data. *SERIEs*, 12. http://dx.doi.org/10.1007/s13209-021-00251-7.

De Livera, A., Hyndman, R., y Snyder, R. (2010). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106:1513–1527. 10.1198/jasa.2011.tm09771.

Eraslan, S. y Götz, T. (2021). An unconventional weekly economic activity index for germany. *Economics Letters*, 204. https://doi.org/10.1016/j.econlet.2021.109881.

Eurostat (2015). *ESS Guidelines on Seasonal Adjustment*. http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-15-001.

Ollech, D. (2018). Seasonal adjustment of daily time series. Bundesbank Discussion Paper 41/2018.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Taylor, S. y Letham, B. (2017). Forecasting at scale. *The American Statistician*, 72. https://doi.org/10.7287/peerj.preprints.3190v2.