«Nowcasting our economy»

¿Es posible predecir el estado actual de nuestra economía a partir de datos de transacciones de pago en terminales punto de venta de comercios?

Daniel Girela @girela_d | *Consultor en el área de Desarrollo Tecnológico de Afi*
Álvaro Jiménez | *Consultor en el área de Desarrollo Tecnológico de Afi*

Nowcasting our economy es una iniciativa que trata de validar la conexión entre indicadores macroeconómicos oficiales y una muestra parcial de la actividad económica real del país, representada con información de medios de pago electrónico de BBVA. Esta información está constituida por transacciones de pago de tarjetas de BBVA en terminales de punto de venta operados por cualquier entidad, y por otro lado, por transacciones de pago de tarjetas de cualquier entidad en terminales de punto de venta de BBVA entre los años 2013 y 2015.

Para tal fin, se ha dispuesto de más de 2.000 millones de transacciones de pago en terminales de punto de venta, para las que se registran las siguientes variables:

- Identificador de cliente y tarjeta.
- Información relativa al cliente: código postal y municipio de residencia; género y edad.
- Información relativa al comercio: ramo; categoría y subcategoría BBVA (grupos de gasto como «transporte», «alimentación», «ropa», etc.); código postal, municipio y coordenadas.
- Información propia de la transacción: tipo de transacción, canal de compra e importe.

La fuente de inspiración principal es el trabajo *Predicting Regional Economic Indices Using Big Data of Individual Bank Card Transactions* (Sobolevsky et al., 2015, disponible aquí), donde emplean un dataset similar (conteniendo únicamente transacciones de tarjeta del año 2011) para ajustar modelos que expliquen la variabilidad espacial de algunos indicadores económicos de nuestro país tales como el PIB, la tasa de paro o la esperanza de vida. En nuestro trabajo, los indicadores que modelizamos han sido los siguientes:

- Demografía empresarial: número de unidades locales (por provincia y año; años 2013 a 2015).
- PIB per cápita (por provincia; año 2013).

- Contribución al PIB per cápita (por provincia; año 2013) de los sectores CNAE del K al N que engloban actividades financieras, de seguros, inmobiliarias, científicas, técnicas y administrativas. Se probaron otros sectores CNAE, pero son estos en los que se consigue un mejor ajuste.
- Renta media anual por hogar (por comunidad autónoma y año; años 2013 a 2015).
- Tasa de paro (por provincia y trimestre; años 2013 a 2015).

El motivo fundamental por el que tiene interés construir modelos para este tipo de indicadores empleando como fuente el *dataset* proporcionado por BBVA está en que los propios indicadores se publican con cierto retardo y a un nivel de desagregación temporal y espacial poco detallado. Disponer de estos modelos podría servir para predecir en tiempo real (*nowcasting*) esos indicadores económicos. Además, podrían emplearse, tomando como *inputs* las mismas variables calculadas a

niveles de desagregación adecuados, para aproximar cuestiones como por ejemplo «el PIB de un barrio», ya que al ser modelos lineales, «desagregaciones a la entrada se convierten en desagregaciones a la salida».

Las variables que se han considerado en la construcción de los modelos se clasifican en los siguientes grupos:

• Gasto en periodos: 30 variables que representan gasto agregado por franjas horarias (mañana, mediodía, tarde, noche, madrugada), épocas del año (navidad, verano, rebajas), fines de semana.

• Variabilidad de la oferta: 22 variables que representan el número de negocios activos por subcategoría BBVA en la zona.

• Gasto categorizado: 19 variables que representan importes gastados por subcategoría BBVA en la zona, así como gastos en comercios lujosos, esto es, en establecimientos donde la transacción media es más alta que la transacción media en su subcategoría BBVA en la zona.

• Gasto por rango de edad: 18 variables que representan volumen y gasto agregado por rango de edad en la zona.

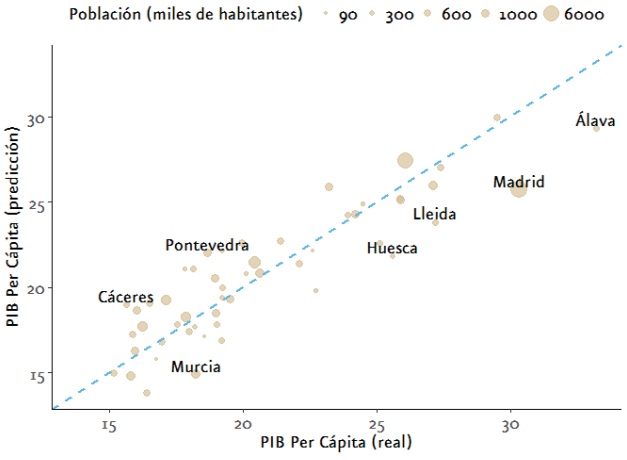
• Movilidad: diez variables que tratan de reflejar patrones de movimiento de los clientes en la zona.

• Densidad de operaciones: tres variables que reflejan volumen, importe total e importe medio de transacciones realizadas en la zona.

• Variabilidad de demanda: dos variables que representan el número de subcategorías BBVA en las que se concentra el 80% del gasto total realizado en la zona, o del gasto total realizado por residentes en la zona.

Un ejemplo de los modelos obtenidos es el que aparece en la siguiente figura, donde para cada provincia está representada nuestra estimación en función del valor real:

Modelo para el PIB per cápita
Valor real vs predicción
(millones de euros)



Fuente: *Afi*.

Una vez construidos los modelos, en los que, naturalmente, no intervienen todas las variables citadas anteriormente sino aquellas que funcionan bien para evitar sobreajustar, observamos que, con nuestro método de muestreo, estos consiguen reflejar la variabilidad espacial de los indicadores en cuestión (llegando a conseguir coeficientes de determinación de en torno al 80% para el PIB per cápita, por ejemplo) con mayor éxito que la correlación con las fluctuaciones de los cálculos trimestrales o anuales de dichos indicadores. Esto sugiere que una posible línea de mejora sería combinar el uso de estas técnicas con modelos econométricos de mayor complejidad e, incluso, alguna fuente de información de naturaleza macro. Además de esto, cabe reseñar que algunas de las variables que aparecen, de manera recurrente, en los modelos de los distintos indicadores, son:

• Volumen de transacciones realizadas por residentes en la provincia entre las 18:00 y las 22:00 horas.

• Volumen de oferta relativa a servicios de transporte en el área.

• Volumen de gasto en alimentación sobre gasto total realizado en el área.

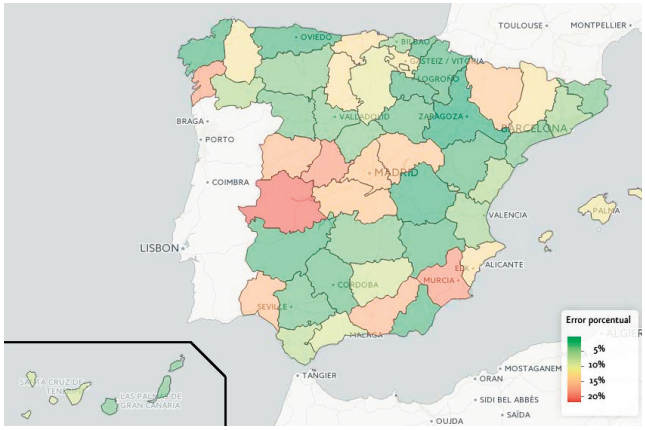
• Volumen de transacciones realizadas por extranjeros en el área.

• Densidad de negocios activos en el área.

• Número de viajes interprovinciales realizados por residentes en el área.

Para finalizar, podemos ver a continuación un mapa de calor del error de las estimaciones obtenidas con el modelo del PIB per cápita de la primera figura ::

Modelo para el PIB per cápita
Mapa de calor de calidad de las predicciones



Fuente: *Afi*.

Nota técnica del proyecto: Se ha utilizado un cluster Hadoop gestionado con Cloudera Manager. Todos los datos han sido almacenados en formato Parquet en una base datos administrada con Hive e Impala. Tras realizar las agrupaciones y cálculos necesarios, los posteriores análisis, modelizaciones y representaciones gráficas de resultados se han llevado a cabo en lenguaje R, usando como entorno de desarrollo RStudio.

Equipo de Afi que ha participado en el proyecto: Miguel Ángel Corella, Borja Foncillas, Daniel Girela, Álvaro Jiménez, Elena Montesinos, Esteban Moro, José Manuel Rodríguez, María Romero y Diego Vizcaíno.