El uso del Big Data para la producción estadística. La experiencia de BBVA Research

Fernando Bolívar García
Sirenia Vázquez Báez
BBVA Research

Introducción

La era digital y la innovación tecnológica están revolucionando el paradigma industrial, incluyendo el sector financiero. El flujo de información procedente de la actividad cotidiana de los individuos genera un manantial de datos que ofrece la oportunidad de ampliar nuestra comprensión sobre el comportamiento de los individuos, la economía y la sociedad. Cuando leemos las noticias en internet, buscamos algún término o palabra en algún buscador, participamos en redes sociales, hacemos una compra o transacción financiera de manera electrónica, añadimos información a este universo de datos. Una vez anonimizados y procesados, estos datos masivos plantean nuevos retos y ofrecen soluciones de valor e innovadoras a preguntas relevantes que no podían ser respondidas por datos tradicionales. Esto es lo que denominamos en BBVA Research cono "Economía en Tiempo Real y Alta Definición".

Volumen, Velocidad y Variedad

Las características principales de **volumen, velocidad y variedad**, son claves en la definición y en el entendimiento del Big Data. El volumen hace referencia a la cantidad de información recopilada no solo por las organizaciones, sino toda la generada por redes sociales, búsquedas online, datos geolocalizados, interacciones con los clientes, datos recogidos durante los procesos de negocio, etc. En definitiva, un tamaño inmenso de datos estructurado y no estructurado cuya gestión y análisis es todo un reto y brinda la oportunidad de enriquecer las bases de datos tradicionales con datos de alta dimensión.

La **velocidad**, se entiende como la rapidez con que los datos son generados, almacenados y procesados. Cada vez más, el tiempo es una variable fundamental para la comprensión y anticipación de eventos en el entorno económico y financiero. La alta frecuencia se convierte en una importante herramienta de diagnóstico y alerta temprana para detectar rápidamente cambios de comportamiento, especialmente en el cambiante panorama actual.

Por último, y no menos importante, el concepto de **variedad** recoge la amplia gama de información disponible en nuestro entorno. Los datos clásicos en formato fila-columna han dejado de ser los protagonistas, actualmente los datos desestructurados como el texto, imágenes, videos o audios están ganando terreno. Esta diversidad supone un incremento cuantitativo y cualitativo de la información disponible tanto para el análisis social, económico o financiero.

Estos tres pilares básicos del Big Data requieren además de ciertos atributos que trasladen su potencial en utilidad para el analista. Entre ellos, podemos añadir conceptos como los de **veracidad, valor y visualización**. A medida que los datos crecen, es necesario controlar la integridad y fiabilidad de la información recogida ya que de esto dependerá principalmente el acierto de las decisiones. Asimismo, la abundancia de datos puede difuminar el propósito de cualquier estrategia Big Data, por lo que es primordial mantener el foco de convertir la información en conocimiento como factor diferencial de manera rentable y a la vez eficiente. Finalmente, uno de los puntos clave y definitivos es el modo en el que la información obtenida es presentada y transmitida ya que de

esto dependerá mayoritariamente el alcance y el valor agregado generado.

Ventajas y desafíos

El crecimiento explosivo de datos recogidos y el avance tecnológico imparable brindan un sinfín de oportunidades que conllevan una serie de ventajas competitivas, pero no están ajenos a inconvenientes.

A nivel agregado, dicha información puede transformarse en indicadores que midan sentimiento, comportamiento o actividad económica. Dichos indicadores permiten conocer en tiempo real lo que está ocurriendo, lo que proporciona un margen extraordinario a las autoridades, empresas e individuos para la toma de decisiones . Además, nos indican con elevado detalle los grupos de individuos que necesitan una respuesta individualizada (Políticas Inteligentes). La crisis sanitaria del Covid-19 es un buen ejemplo de ello, pues la información generada a gran escala fue de vital importancia para, entre otras cosas, observar en tiempo real el impacto de las medidas de contención de la pandemia sobre hogares, empresas y la economía en su conjunto. De hecho, ha sido durante esta crisis del Covid que se ha desencadenado una mayor demanda de información en tiempo real y alta granularidad, ya sea para contar con indicadores alternativos o generar modelos de alta frecuencia.

Además de la inmediatez, a un coste menor que otros procedimientos se puede obtener una imagen con alta granularidad, en alta definición, en términos geográficos, sectoriales o temporales. Esto abre la posibilidad de plantearse preguntas que hasta ahora era imposible responder empíricamente. Un ejemplo de ello es la información generada a través de las transacciones financieras, la cual ha abierto nuevos caminos para entender mejor el comportamiento de hogares, empresas e individuos. Como mencionan Baker y Kueng (2021), "la granularidad de los datos y el hecho de que a menudo están emparejados con otras cuentas financieras de las que es titular el individuo, permiten formular preguntas sobre la totalidad del balance o los flujos financieros del hogar. Los investigadores pueden controlar mejor por posibles variables de confusión y rechazar explicaciones alternativas para una pregunta determinada"[1].

Por otro lado, la infraestructura tecnológica actual está democratizando el dato permitiendo la colaboración entre distintas disciplinas, la combinación de fuentes y de formatos consiguiendo un conocimiento más profundo del mercado. Estas ventajas de aplicación empiezan a experimentarse no solo entre organizaciones privadas sino por organismos públicos como institutos nacionales de estadística, bancos centrales y centros académicos de investigación pues complementan y enriquecen la información que recaban con métodos más "tradicionales".

El crecimiento explosivo de datos recogidos y el avance tecnológico imparable brindan un sinfín de oportunidades que conllevan una serie de ventajas competitivas, pero no están ajenos a inconvenientes

Sin embargo, como toda tecnología, el Big Data conlleva una serie de desafíos que deben ser tenidos en cuenta. Uno de los principales retos es la calidad o la escasez de los datos, así como el muestreo y la representatividad de los mismos. Otros factores relevantes a tener en cuenta, además del problema de la dimensionalidad, son la multicolinealidad y el sesgo intrínseco, ya que los grandes volúmenes de datos no eluden directamente los problemas de inferencia habituales de estadística. Por último, trabajar con datos diarios hace que el problema de desestacionalización de los datos sea más complejo. Existen también desafíos "no técnicos" como garantizar en todo momento la privacidad y la confidencialidad de los datos que

1 Baker, S. R., & Kueng, L. (2021). *Household Financial Transaction Data* (No. w29027). National Bureau of Economic Research.

juegan un papel fundamental en la seguridad jurídica de los individuos. Finalmente, existen desafíos de seguridad, dada la exposición de los datos a potenciales ciberataques.

Con todas estas y otras ventajas y desventajas, ¿cómo podemos los economistas beneficiarnos de la revolución del Big Data?, ¿abre el Big Data nuevos temas de investigación para los economistas financieros? o ¿nos permiten responder preguntas tradicionales de forma novedosa y más reveladora?

Aplicaciones BBVA Research

En BBVA Research llevamos varios años haciendo uso de la información y herramientas disponibles para ilustrar la utilidad del Big Data dentro del campo social, económico y financiero. El potencial es grande, pero las dificultades no son menores, pues para hacerlos útiles, los datos deben pasar por distintas etapas, con retos en cada una de ellas y siempre preservando la anonimidad de los individuos que realizan estas operaciones.

Nada ha cambiado aquí, el principal reto al que se enfrenta el investigador es el de plantearse la pregunta adecuada

El primer reto al que se enfrenta BBVA Research al trabajar con Big Data es asegurar una buena materia prima. En esencia, esto supone mantener una infraestructura adecuada para ingestar, gestionar, procesar y anonimizar la información. Este proceso es complejo y costoso aunque una vez garantizado los resultados superan con creces el coste. Una vez que la información está disponible, el siguiente reto es identificar exactamente el dato que se está buscando, procesarlo y extraerlo, ya sea de fuentes internas o externas. Nada ha cambiado aquí, el principal reto al que se enfrenta el investigador es el de plantearse la pregunta adecuada.

Una vez que la pregunta y los datos están disponibles, comenzamos la fase de procesamiento, tratamiento y validación. La duración de esta etapa dependerá de la calidad de la información (existencia de nulos, dispersión o baja calidad del dato). Hay que tener en cuenta que en muchos casos no existe un horizonte temporal lo suficientemente largo, ya sea para aplicar técnicas estadísticas convencionales a una serie de tiempo generada con Big Data o para incorporar estos datos a los modelos de pronósticos que suelen utilizar horizontes de tiempo amplios. Así, para lidiar con ello es necesario aplicar no solo los métodos estadísticos tradicionales sino también nuevas metodologías, como por ejemplo análisis de lenguaje natural, regularización, reducción de la dimensionalidad, modelos de aprendizaje automático, entre otros. La última etapa del proceso corresponde a la narrativa y visualización que son fundamentales para explicar lo que hacemos y por qué lo hacemos.

El resultado del proceso mencionado son algunos de los indicadores en tiempo real y alta definición como los de consumo e inversión en tiempo real. En el primer caso, utilizando las transacciones financieras estudiamos el comportamiento agregado de los consumidores gracias a la información agregada y anonimizada de los pagos con tarjeta BBVA y tarjetas ajenas en terminales de punto de venta españoles. Concretamente, dicha información nos ha permitido analizar el impacto del gasto en España durante la actual pandemia de la Covid-19, así como las políticas aplicadas para controlarlo a nivel diario. Los metadatos de las transacciones también nos permiten estudiar la variación de los comportamientos en función de la geografía, los sectores y el modo de venta (por ejemplo, *online/offline*). Llegamos a la conclusión de que estos datos son capaces de captar las tendencias relevantes de gasto y, lo que es más importante, lo hacen en tiempo real.

En el segundo caso, podemos utilizar las transacciones entre empresas para desarrollar un índice para medir la inversión en tiempo real y alta definición. Este indicador es innovador pues utili-

za datos agregados de transferencias financieras empresa-empresa e individuo-empresa. En particular, se utilizan aquellas transferencias en las cuales las empresas vendedoras son productoras de bienes y servicios de inversión. Dado que la información está desagregada por sector de actividad económica y está geolocalizada, podemos conocer con detalle el tipo de inversión (Construcción, Maquinaria y Equipo o Bienes Intangibles) y la región donde esta se produce.

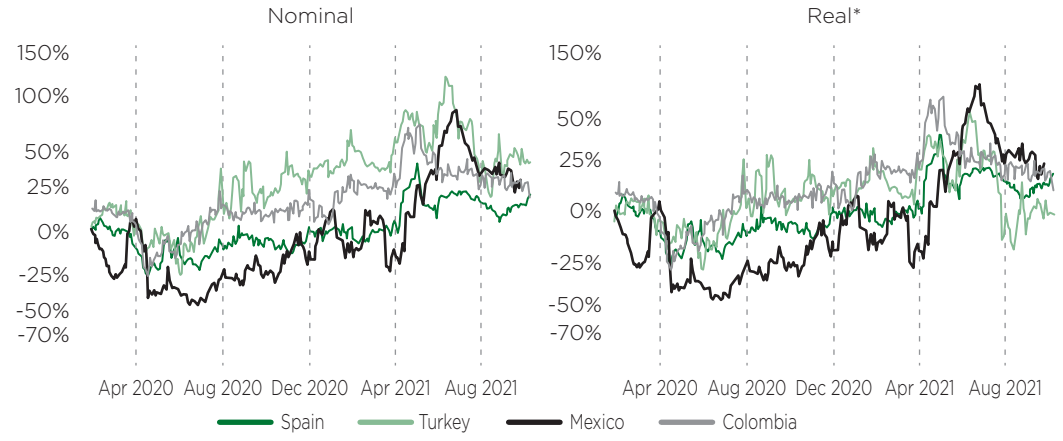
Estos indicadores tienen una alta correlación (de 0.60 o superior) con los componentes de la inversión de cuentas nacionales y, dada la alta frecuencia de los datos utilizados, es posible conocer la evolución de la inversión con anticipación. Gracias a ello, se pudo observar el impacto que tuvo sobre la inversión la interrupción de las cadenas globales de valor generada por la pandemia. Del mismo modo, conforme se ha normalizado la situación, se observa una recuperación.

Una aplicación especial de estos indicadores es el caso de la inversión en bienes intangibles, para los cuales existe escasa información pública para medir su composición y comportamiento. Este componente puede tener, para el caso de España, un gran potencial para monitorear el efecto de políticas de sostenibilidad y de asignación de recursos para la inversión, como es el caso de los Fondos de Recuperación para Europa.

Pero también hemos desarrollado estudios temáticos o focalizados en un tipo de consumo. Como aplicación derivada del gasto de consumo, podemos estimar el gasto turístico nacional en España mediante el gasto realizado con tarjetas BBVA fuera de la provincia habitual de residencia. Esto nos permite monitorizar los flujos de gasto turístico dentro de España a nivel agregado, por comunidades autónomas y por provincias, de forma diaria. De nuevo, la alta frecuencia y granularidad de los datos de tarjeta permiten avanzar y complementar las cifras oficiales de turismo nacional, que como se está demostrando, está siendo clave en el proceso de recuperación. Del mismo modo, conseguimos detectar los efectos permanentes y temporales de la crisis en las preferencias de los turistas españoles durante las limitaciones de movilidad, así como la llegada de turistas extranjeros.

Además de los datos propios del banco, utilizamos un conjunto amplio de información no estructurada que nos permite entender mejor el entorno mediante la aplicación de técnicas

FIGURA 1. INDICADORES DE INVERSIÓN BIG DATA: INVERSIÓN TOTAL.
(% interanual acumulado 28 días)



Indicadores de inversión estimados a partir de las transacciones monetarias entre empresas y hogares a las empresas nacionales que producen los activos fijos definidos por el código NACE. "Series reales deflactadas por los precios de producción, excepto Perú, cuyo deflactor es el índice de precios mayoristas y España, que utiliza el deflactor implícito del componente de la inversión del PIB por el lado de la demanda. Forthcoming (Carvalho et al. 2021).

Fuente: BBVA Research.

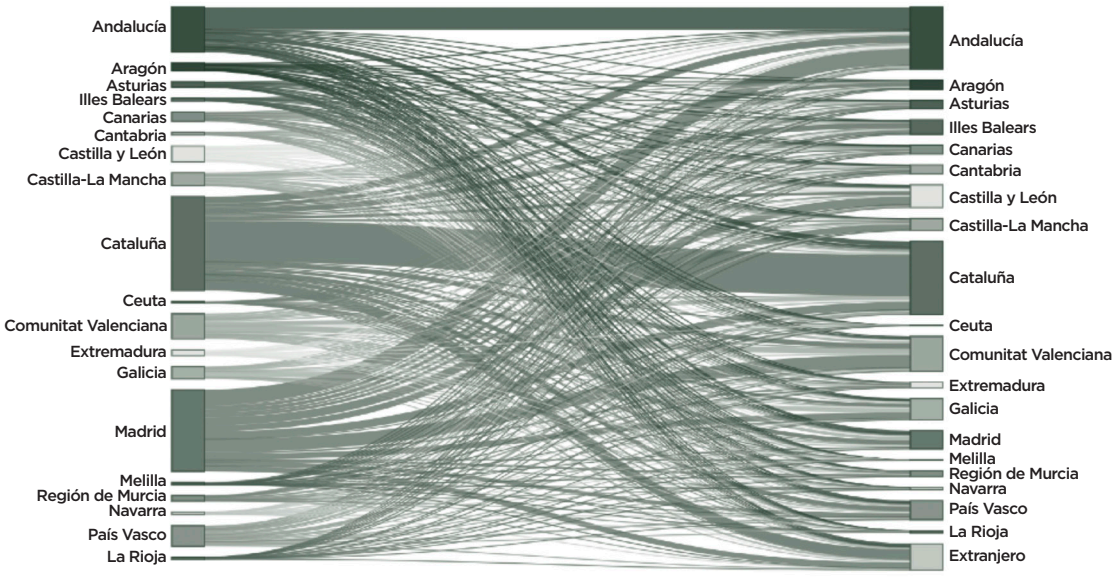
Como aplicación derivada del gasto de consumo, podemos estimar el gasto turístico nacional en España mediante el gasto realizado con tarjetas BBVA fuera de la provincia habitual de residencia

de analítica avanzada. Para entender la evolución y percepción de algunos eventos sociales o geopolíticos empleamos información procedente de redes sociales (Twitter), medios de comunicación a escala mundial (GDELT) o el volumen de búsquedas realizadas en internet (Google Trends). Estos datos generados por la interacción social nos han permitido estudiar temas económicos de gran calado como el estancamiento del comercio mundial y el retroceso de la globalización, así como entender el marco de la sostenibilidad o el universo Fintech. Del mismo modo, hacemos uso de la información publicada en la web, por ejemplo, el análisis de la narrativa de los bancos centrales en sus comunicados de prensa y actas de política monetaria nos ayuda a entender las estrategias de comunicación, y a anticipar sus decisiones.

Reflexión

Los indicadores mencionados son solo un ejemplo del uso de datos masivos de transacciones financieras. BBVA Research ya ha trabajado en otras aplicaciones, como la medición de impacto de la política monetaria en tiempo real, el desarrollo de indicadores para medir la evolución del sector exterior y de los distintos sectores de actividad económica, o la construcción de tablas insumo producto en alta frecuencia. Asimismo, desde BBVA Research entendemos estas fuentes como un activo clave para tener respuestas rápidas a eventos de interés dada su alta frecuencia, así como para comprender la percepción social de los mismos, puesto que recogen la opinión generada por la sociedad, actores particulares, empresas e instituciones. ●

FIGURA 2. FLUJOS DE GASTO PRESENCIAL CON TARJETA REALIZADO FUERA DE LA PROVINCIA HABITUAL DE RESIDENCIA DE COMUNIDAD ORIGEN A DESTINO (Porcentaje de gasto sobre el total, Julio-Agosto 2021)



Fuente: BBVA Research a partir de BBVA.