Big Data y nuevas geografías: la huella digital de las actividades humanas*

Javier Gutiérrez Puebla
Universidad Complutense de Madrid. Departamento de Geografía Humana
javiergutierrez@ghis.ucm.es

Recibido: enero de 2018
Aceptado: enero de 2018

Resumen

El término *Big Data* se ha popularizado en los últimos años y hace referencia a la producción de cantidades ingentes de datos. La actividad humana es captada a través de múltiples redes de sensores y dispositivos, dejando por tanto una huella digital. El análisis de esta huella digital tiene un gran potencial para la investigación geográfica del comportamiento humano. En este artículo se describen las principales características del Big Data y se destaca la importancia de los datos masivos para la ciencia y particularmente para la Geografía, centrando la atención en el estudio de los patrones espacio-temporales de la actividad humana.

Palabras clave: Big Data; datos geolocalizados; comportamiento humano

Resum. *Big Data i noves geografies: l'empremta digital de les activitats humanes*

El terme *Big Data* s'ha popularitzat en els últims anys i fa referència a la producció de quantitats ingents de dades. L'activitat humana és captada a través de múltiples xarxes de sensors i dispositius i, per tant, deixa una empremta digital. L'anàlisi d'aquesta empremta digital té un gran potencial per a la investigació geogràfica del comportament humà. En aquest article es descriuen les principals característiques del Big Data i es destaca la importància de les dades massives per a la ciència i, particularment, per a la Geografia, centrant l'atenció en l'estudi dels patrons espaciotemporals de l'activitat humana.

Paraules clau: Big Data; dades geolocalitzades; comportament humà

* Este artículo se basa parcialmente en la conferencia «Big Data y nuevas geografías», pronunciada en la Universidad Autónoma de Barcelona en 2017, y en la ponencia «El uso del Big Data en la investigación de la ciudad, la movilidad y el turismo», presentada en el XXV Congreso de la Asociación de Geógrafos Españoles.

https://doi.org/10.5565/rev/dag.526 ISSN 0212-1573 (imprès), ISSN 2014-4512 (en línia)

Résumé. *Big Data et nouvelles géographies : l'empreinte digitale des activités humaines*

Le terme *Big Data* est devenu populaire ces dernières années et il se réfère à la production d'énormes quantités de données. L'activité humaine est captée par une multitude de réseaux de capteurs et de dispositifs, laissant ainsi une empreinte digitale. L'analyse de cette empreinte présente un grand potentiel pour l'étude géographique du comportement humain. Cet article décrit les principales caractéristiques du *Big Data* et il souligne l'importance des données massives pour la science et en particulier pour la géographie, en se concentrant sur l'étude des modèles spatio-temporels de l'activité humaine.

Mots-clés: Big Data; données géolocalisées; comportement humain

Abstract. *Big data and new geographies: The digital footprint of human activity*

The term 'big data' has become popular in recent years and refers to the production of huge amounts of data. Human activity is captured through multiple networks of sensors and devices, thus leaving a digital footprint. The analysis of this digital footprint has a great potential for geographical research on human behavior. This article describes the main characteristics of big data and highlights the importance of massive data for science and particularly for the field of geography, focusing on the study of spatio-temporal patterns of human activity.

Keywords: big data; geo-located data; human behavior

Sumario

1. Introducción

Big Data es un término que se ha popularizado en los últimos años y hace referencia a la producción de cantidades ingentes de datos tanto por medio de múltiples redes de sensores y dispositivos, como a partir de la actividad de los usuarios en Internet. Estos datos aportan información sobre multitud de procesos, desde el crecimiento de las ciudades (sensores remotos) y la contaminación del aire (sensores localizados en las ciudades) hasta la movilidad de la población (teléfonos móviles) o el consumo (tarjetas de crédito). Tienen en común que se trata de datos de carácter masivo, de distinta naturaleza a los convencionales, y requieren herramientas específicas para su tratamiento, las denominadas tecnologías Big Data.

Gran parte de estos datos ofrecen información sobre las actividades humanas. Los humanos dejamos un rastro digital, de forma voluntaria o involunta-

ria, cuando realizamos actividades. Estamos rodeados de dispositivos y sensores que permiten la monitorización de nuestra actividad. Así, por ejemplo, dejamos nuestra huella digital cuando utilizamos nuestro teléfono móvil, pagamos con tarjeta de crédito, utilizamos el transporte público con nuestra tarjeta de transporte o cuando participamos en las redes sociales. Estos datos constituyen una valiosísima materia prima para el estudio del comportamiento humano. Permiten analizar pautas espaciales y procesos que no podían ser estudiados (al menos de la misma forma) con las estadísticas oficiales o con encuestas. En definitiva, las nuevas fuentes de datos hacen posible que los investigadores puedan arrojar luz sobre fenómenos geográficos que antes quedaban ocultos. En este sentido, se puede afirmar que el Big Data permite explorar «nuevas geografías» de las actividades humanas.

Las nuevas fuentes de datos usadas para analizar la actividad humana y el ritmo de las ciudades son muy variadas. Así mismo, las temáticas que se han abordado con estas fuentes de datos son muy diversas. Este artículo tiene como objetivo revisar el estado de la cuestión en torno a la utilización de las nuevas fuentes de datos masivos en nuestra disciplina, centrando la atención en el estudio de los patrones espacio-temporales de la actividad humana a partir de datos geolocalizados.

El artículo se estructura de la siguiente forma. Tras esta breve introducción, el segundo apartado se dedica a caracterizar los datos masivos, situándolos en contexto, y destacando su valor para el mundo de la empresa. El tercer apartado se centra en la importancia de los datos masivos para la ciencia y particularmente para la Geografía, recurriendo al denominado «cuarto paradigma» de la investigación científica. El cuarto apartado aborda cuestiones relativas a la resolución espacial y temporal de distintos tipos de datos masivos, y su importancia para identificar patrones espaciales y analizar procesos espacio-temporales. El artículo se cierra con unas consideraciones finales sobre algunas de las barreras existentes para la utilización del Big Data en la investigación geográfica y sobre la relación entre las encuestas y la estadística oficial con el Big Data.

2. La revolución del Big Data: características de los datos masivos

El término *Big Data* hace referencia a la producción de datos masivos por medio de sensores y dispositivos, a un ritmo anteriormente desconocido. Estamos viviendo una auténtica revolución de los datos, en la que los datos adquieren un valor creciente para las empresas y para la sociedad en su conjunto. Esta revolución hay que situarla en el contexto de las *Smart Cities*, ciudades sensorizadas que se conciben como «constelaciones de instrumentos conectadas a través de múltiples redes que proporcionan datos continuos sobre los movimientos de personas y mercancías y sobre el estado de estructuras y sistemas» (Batty et al., 2012); de la web 2.0, en la que los usuarios no son meros receptores de información, sino que Internet es concebida como una plataforma en la que sus usuarios generan una enorme cantidad de contenidos (Goodchild, 2007); del Internet de las cosas, donde diversos dispositivos están conectados entre sí

intercambiando datos y desencadenando acciones o procesos en función de la información recibida (Xia et al., 2013); y de la denominada cuarta revolución industrial (Bloem et al., 2014; Schwab, 2017), que estaría caracterizada por la transición hacia nuevos sistemas que están construidos sobre la infraestructura de la revolución digital: sistemas ciberfísicos basados en la automatización, la robotización, la inteligencia artificial y el propio Internet de las cosas.

Los datos masivos tienen un gran valor para las empresas. Como señala Batty (2013), la mayor parte de la información que ahora llamamos Big Data se produce de forma automática, rutinaria, y por diversas formas de sensores. Casi todos estos datos son capturados y almacenados para llevar a cabo procesos de control y gestión en las empresas (por ejemplo, gestión de cargos en las tarjetas de crédito), pero después han sido utilizados para usos distintos a los que fueron concebidos, como el análisis del comportamiento de los consumidores para diseñar estrategias de marketing, predecir las tendencias del mercado o controlar el fraude. Así, por ejemplo, los datos de gastos con tarjetas bancarias fueron concebidos para gestionar los pagos de los usuarios, pero se pueden utilizar también para identificar las áreas con mayor gasto de la ciudad o el impacto económico de un evento turístico. De igual forma, los datos de las compañías telefónicas sirven para efectuar el correspondiente cargo al cliente por uso de los servicios telefónicos, pero también son utilizados para realizar estudios de movilidad o de geomarketing.

Además, hay que considerar que los usuarios de Internet disponen de multitud de servicios gratuitos. El usuario no paga por utilizar el buscador de Google, calcular una ruta entre dos puntos con Google Maps, publicar sus fotografías en Instagram o utilizar Twitter y Facebook. Sin embargo, compañías tecnológicas como Google o Facebook se encuentran entre las primeras en el ranking mundial por capitalización bursátil. Estas compañías ofrecen servicios al usuario y este aporta sus datos cuando los utiliza. Y estos datos tienen un gran valor económico. Así, por ejemplo, según Statista (<https://es.statista.com/grafico/7905/>), considerando los ingresos de Facebook y el número de usuarios por grandes áreas mundiales, se puede estimar que el valor de los datos de un usuario medio en 2016 fue de 16 dólares. Los datos que obtiene Facebook de sus usuarios (datos personales, páginas visitadas, «likes», etc.) se utilizan para elaborar perfiles y a partir de ellos enviar publicidad a los usuarios que tienen las características de consumo adecuadas. Numerosas compañías (bancos, operadores de telefonía, redes sociales, etc.) venden los datos de sus clientes una vez anonimizados o, lo que es más frecuente, venden estudios basados en esos datos. Los datos han adquirido tal valor que desde el sector empresarial se afirma que son el petróleo de la economía del siglo XXI.

Los datos masivos se pueden clasificar básicamente en dos grandes tipos, atendiendo al modo en que son generados: los producidos por máquinas y los generados por los usuarios en Internet. Al primer grupo corresponden los generados por múltiples redes de sensores y dispositivos que registran la huella digital de la actividad humana, generalmente de forma pasiva, es decir, sin que el usuario tenga voluntad de generar esa huella. Esto ocurre, por ejemplo,

con los registros (tracks) de los GPS de los teléfonos móviles que portamos, la actividad de nuestro teléfono móvil (llamadas, mensajes, sesiones de datos) captada por las antenas de las operadoras de telefonía, los pagos que realizamos con tarjetas bancarias, el uso que hacemos del transporte público o sistemas de bicicletas públicas a través de las correspondientes tarjetas, la captación de imágenes de peatones o vehículos por parte de cámaras instaladas en espacios públicos o carreteras, los registros de consumo de agua y electricidad de los hogares, etc.

Por otro lado, están los datos generados por los usuarios en Internet. La red se ha convertido en un gran repositorio que contiene datos de lo que la gente dice, compra y busca, y de cómo unos conectan con otros (Miller, 2010). En el contexto de la web 2.0, el usuario de Internet ha pasado de tener una actitud pasiva, como mero receptor de información (que fluía de arriba abajo, es decir, desde las empresas e instituciones hasta los ciudadanos), a adoptar una posición activa, como generador de contenidos (en un flujo predominante de abajo a arriba). Los usuarios generan enormes volúmenes de datos cuando participan en redes sociales generalistas (como Facebook o Twitter) o especializadas (por ejemplo redes para compartir fotografías, como Instagram o Panoramio), cuando hacen búsquedas en Internet, envían mensajes electrónicos, suben o descargan vídeos, etc. También cuando participan en grandes proyectos colaborativos, como por ejemplo OpenStreetMaps o Wikipedia. Para tener una idea de los volúmenes de datos generados en Internet basta decir que cada minuto en todo el mundo se envían 156 millones de mensajes electrónicos y 452.000 tweets, y se hacen 950.000 accesos a Facebook (<http://www.eleconomista.es/tecnologia/noticias/8236038/03/17/>).

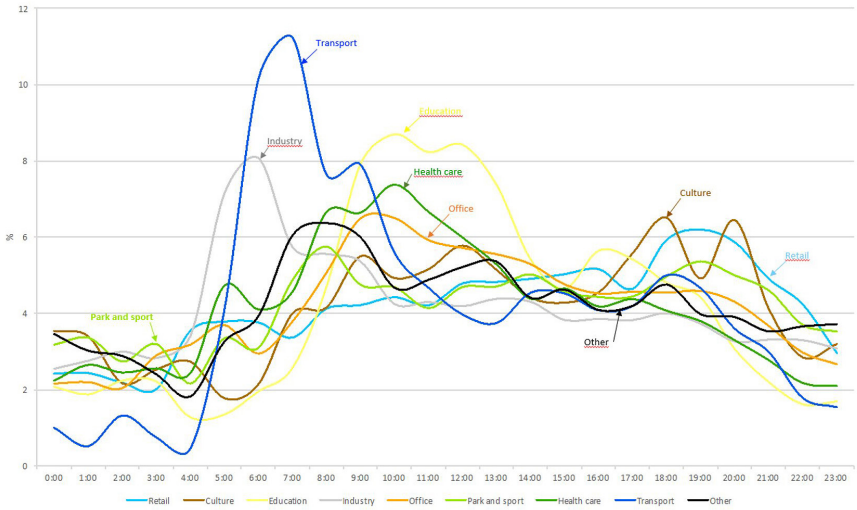
Las nuevas fuentes de datos que reciben genéricamente el nombre de Big Data se caracterizan no solo por su volumen, sino también por otros rasgos distintivos frente a las fuentes de datos tradicionales. Para caracterizar el Big Data generalmente se recurre al esquema de las 3 vs (volumen, velocidad y variedad), si bien a veces se añaden algunas características adicionales para alcanzar las 5 vs (veracidad y valor) o incluso las 7 vs (variabilidad y visualización). Aquí se hace referencia a las cinco de las principales características del Big Data, conocidas como las 5 vs (ver, por ejemplo, Kaisler et al., 2013; Kitchin, 2013; Chen et al., 2014):

— *Volumen:* el Big Data se caracteriza por su volumen, su carácter masivo. Según la compañía tecnológica Oracle, los datos producidos en todo el mundo crecen a un ritmo del 40% anual. Las unidades de medida para estas ingentes cantidades de datos son los terabytes (un billón de bytes), petabytes (mil billones de bytes), exabytes (un millón de billones de bytes)… Las operadoras de telefonía móvil, por ejemplo, recogen miles de millones de registros al día, lo que al cabo de varios años supone un ingente volumen de datos. A diferencia de lo que ocurría en el pasado, los datos históricos se mantienen (no se borran) porque sirven de apoyo para analizar procesos y realizar predicciones.

— *Velocidad:* relacionado con el volumen está la velocidad a la que los datos son generados y procesados. Se generan de forma continua, de manera que es posible seguir procesos en *streaming* y hacer análisis en tiempo casi real. La API de Twitter, por ejemplo, permite descargar tweets en tiempo real e identificar los «trending topics», los temas candentes en un determinado momento y lugar. Frente a las estadísticas oficiales, que ofrecen una instantánea (una foto fija) de un proceso en un momento dado, la velocidad asociada al Big Data aporta datos con alta resolución temporal, lo que permite seguir procesos sociales de forma continua (una película).
— *Variedad:* la variedad hace referencia a la diversidad de tipos, formatos y fuentes de datos, desde datos estructurados (que pueden ser presentados en forma de tabla y manejados con sistemas gestores de bases de datos SQL) a datos semiestructurados (como los ficheros HTML o JSON) y datos no estructurados (como archivos de texto, correos electrónicos, imágenes o vídeos, que pueden ser tratados más adecuadamente con sistemas NoSQL).
— *Veracidad:* la veracidad hace referencia a la fiabilidad de los datos. El hecho de que los datos sean masivos, exhaustivos o casi exhaustivos (se puede trabajar con poblaciones en lugar de con muestras) y de que deriven de acciones de la gente (no se registra lo que la gente dice que hace, como ocurre con las encuestas, sino lo que la gente hace) son elementos a favor de la veracidad de estos datos; pero esto no implica que sean completamente fiables.
— *Valor:* en la actual sociedad tecnológica la disponibilidad de datos cobra un valor creciente, hasta el punto de que se dice que los datos son el petróleo de la cuarta revolución industrial. Pero el dato en sí mismo no tiene valor. Lo que le da valor económico es convertirlo en información para generar conocimiento útil para la acción, para la toma de decisiones.

Otra característica del Big Data es la geolocalización, el hecho de que en su mayor parte se trate de datos geolocalizados. Se estima que el 80% de los datos masivos son espaciales, ya sea porque se dispone de las coordenadas o la dirección postal del lugar o porque el propio contenido de los datos hace referencia explícita a lugares concretos (Leszczynski y Crampton, 2016). Por lo tanto, son datos cartografiables. Pero el trabajo del investigador de ciencias sociales debe ir más allá del dato geolocalizado y de la generación de mapas simples (por ejemplo, mapas de tweets), para pasar a analizar patrones y procesos espaciales mediante el empleo del análisis espacial y estadístico, situando los datos en su contexto social, económico y político. El Big Data nos aporta datos para estudiar fenómenos que antes conocíamos pero que no podíamos medir. Ahora es posible medir y por tanto también analizar, modelizar y predecir. Es el caso, por ejemplo, de las actividades de consumo. En los años noventa emergió la denominada Geografía del consumo, ligada al florecimiento de los centros comerciales, pero las investigaciones de entonces apenas disponían de datos de consumo más allá de los obtenidos en algunas encuestas en lugares específicos. La disponibilidad de datos de tarjetas bancarias ofrece la posibilidad

Figura 1. Relación entre usos del suelo y uso de Twitter como indicador del nivel de actividad de la ciudad según horas del día



Fuente: García-Palomares et al. (2018).

de analizar los patrones espaciales del consumo, segmentar grupos de consumidores, contrastar modelos, establecer correlaciones, formular predicciones, etc.

La geolocalización nos permite enriquecer los datos mediante procesos relativamente simples. En general, las fuentes de datos masivos ofrecen un número reducido de campos (usuario, coordenadas, momento temporal y algunos atributos específicos); pero a partir de la localización y el momento temporal se pueden detectar patrones e inferir nuevos atributos (por ejemplo, los lugares de trabajo y de residencia). Además, estos datos pueden ser cruzados espacialmente con otras fuentes de datos Big Data y con los datos de las estadísticas oficiales. Podemos relacionar, por ejemplo, los niveles de contaminación de la ciudad (obtenidos por medio de sensores) con la presencia de población (estimada a partir de datos de telefonía móvil) para conocer la cantidad de población expuesta a la contaminación en cada momento del día y cada área de la ciudad (Dewulf et al., 2016; Castell et al., 2017). O relacionar los datos de movilidad de la población obtenidos a partir de Twitter con las características del lugar de residencia según los datos censales, para conocer el uso del espacio de la ciudad según grupos sociales o raciales (Netto et al., 2015; Shelton et al., 2015). Así mismo, se puede indagar sobre el ritmo diario de la ciudad analizando el número de usuarios activos en Twitter según usos del suelo y formulando modelos explicativos (García-Palomares et al., 2018) (ver figura 1).

Estas características del Big Data suponen la aparición de nuevos problemas para su tratamiento. Para el proceso de datos masivos se utiliza la computación paralela, que se basa en el principio de que los problemas grandes se pueden dividir

en partes más pequeñas que pueden resolverse de forma simultánea, es decir, en paralelo, realizando el mismo cálculo en distintos grupos de datos, ya sea en ordenadores multinúcleo o en varios ordenadores conectados entre sí (computación distribuida). Los sistemas distribuidos son escalables, de forma que puede adaptar su capacidad de computación en función de las necesidades, añadiendo nuevos ordenadores. Los procesos se pueden llevar a cabo en sistemas propios o mediante alquiler de servicios de computación en la nube. También emergen problemas derivados de la variedad de estos datos, no necesariamente estructurados, que puede aconsejar utilizar sistemas NoSQL, como CouchDB o MongoDB, más flexibles y rápidos que los sistemas gestores de bases de datos relacionales.

3. El Big Data en la investigación: la emergencia del cuarto paradigma en la ciencia y en la Geografía

La evolución de la ciencia se puede describir a través de una sucesión de paradigmas. Jim Gray, desde el mundo de la Física, distingue cuatro paradigmas (Gray y Szalay, 2007; Hey et al., 2009): el empírico, orientado hacia una descripción de los fenómenos naturales, como por ejemplo el movimiento de los cuerpos celestes; el teórico, basado en la formulación de modelos y generalizaciones, como la ley de la gravedad; el computacional, que sitúa en primer plano las simulaciones de fenómenos complejos gracias a la creciente capacidad de cálculo de los ordenadores; y, finalmente, el cuarto paradigma, el de la ciencia de la exploración de datos (ciencia orientada a los datos). Este cuarto paradigma significaría un retorno al empirismo, ya que se basa en el análisis de grandes volúmenes de datos para obtener respuestas a preguntas científicas sin que sea necesario que esa investigación esté orientada por una teoría previa. Se trataría de «dejar que los datos hablen por sí mismos» sin formular hipótesis o modelos previos (Kitchin, 2014). La extracción de conocimiento de bases de datos parte de la creencia de que las técnicas estadísticas tradicionales no son capaces de descubrir información oculta en bases de datos masivas, heterogéneas y de carácter no-científico. El reconocimiento de patrones, los algoritmos de clusterización, el *machine learning*, la búsqueda numérica, la inteligencia semántica y la visualización científica se acomodarían mejor al Big Data, al no requerir los supuestos estrictos de la estadística clásica (como la independencia, la estacionaridad y la normalidad) (Miller, 2010).

Este esquema evolutivo puede resultar atractivo, pero es cuestionable en algunos puntos: en primer lugar, porque describe la evolución de solo una parte de la ciencia (ciencias experimentales) y, en segundo lugar, porque plantea una visión lineal de la evolución de la ciencia, cuando en realidad se producen superposiciones entre distintos estadios. De hecho, el análisis de datos masivos no sería posible sin los avances computacionales registrados en las décadas anteriores y, además, en contra de lo que algunos sugieren, se puede apoyar en teorías y modelos. Pero en cualquier caso la diferenciación de estos cuatro paradigmas tiene la virtud de situar el foco de atención en los aspectos que destacan más en cada una de ellos.

Los cuatro paradigmas de Gray pueden ser trasladados, con ciertas cautelas, al campo de la Geografía, desde una perspectiva básicamente cuantitativa. La fase descriptiva correspondería a los primeros estadios de la Geografía, desde los repertorios descriptivos de la Antigüedad hasta la Geografía regional francesa. El segundo paradigma tendría su correlato con la Geografía teorética y cuantitativa, orientada a construir modelos y establecer generalizaciones. La tercera fase se identificaría con el desarrollo de la tecnología de los Sistemas de Información Geográfica, que permiten realizar análisis y simulaciones de sistemas espaciales complejos. Por último, el cuarto paradigma sería el de la utilización de datos geográficos masivos (Big Data geolocalizado) para investigar de una forma nueva fenómenos que van desde el cambio climático hasta el pulso de la ciudad (Batty, 2010).

El Big Data abre nuevas oportunidades de investigación a los geógrafos. Al ofrecer datos detallados, actuales y a bajo coste hace posible un entendimiento mucho más sofisticado, de grano fino, de la sociedad y el mundo en que vivimos. Permite pasar de estudios con escasez de datos a estudios con abundancia de datos; de instantáneas a análisis dinámicos; de datos agregados de grano grueso a datos de alta resolución; de hipótesis y modelos relativamente simples a simulaciones y teorías más complejas y sofisticadas. En fin, para los científicos sociales positivistas se abriría una nueva era de ciencia computacional orientada a los datos (Kitchin, 2013). El Big Data permite explorar cuestiones que antes no podían ser investigadas con encuestas o con estadísticas oficiales. Y ese es su principal valor, hacer posible el desarrollo de investigaciones que arrojan luz sobre fenómenos que antes no habían sido estudiados o lo habían sido solo de forma parcial y local. Además, la era del Big Data también ofrece oportunidades para los científicos post-positivistas, dada la disponibilidad de ingentes cantidades de datos no estructurados, como datos de redes sociales o documentos digitales de todo tipo (libros, periódicos, documentos, fotografías, vídeos).

Kitchin (2013) no solo ve oportunidades en el uso del Big Data en la investigación, también detecta riesgos. Estos, básicamente, se pueden resumir en la pretensión de exhaustividad (trabajar con todos los datos y no con muestras) y en la negación de la necesidad de utilizar teorías, modelos o hipótesis (dejar hablar a los datos por sí solos). En cuanto a la pretensión de exhaustividad, hay que tener en cuenta que, aunque los *datasets* del Big Data pueden ser exhaustivos, capturando un dominio completo de datos y proporcionando una resolución también completa (continua), son a la vez una representación y una muestra, conformada por la tecnología y plataforma utilizada, la ontología de datos empleada y el entorno regulador, y está sujeta a sesgo de muestreo (Kitchin, 2013). Se puede trabajar con todos los datos de una fuente de datos (por ejemplo, todos los datos en una red social de fotografías compartidas tomadas en un lugar); pero esos datos finalmente pueden ser solo una parte de un conjunto mayor (todas las fotografías tomadas en ese lugar), sin que conozcamos el margen de error derivado de trabajar solo con esa fuente de datos. De la misma forma, el Big Data no registra todas las acciones de la gente, sino solo una parte de ellas.

Por lo tanto, el sesgo es inherente al Big Data; pero muy variable, dependiendo de la fuente de datos utilizada. Datos provenientes de tecnologías usadas por la mayor parte de la población, como los registros de la actividad de los teléfonos móviles (con una penetración de casi el 100% en los países desarrollados), tendrán un nivel de sesgo bajo si se dispone de *datasets* provenientes de compañías que tienen una alta penetración en el país objeto de estudio y si no hay razones para pensar que esa operadora esté orientada hacia un determinado segmento del mercado. En cambio, las redes sociales, con un nivel de penetración menor y más utilizadas por grupos específicos, tenderán a tener un nivel de sesgo mayor. En España, la red social con mayor tasa de penetración en 2016 es Facebook (46%), seguida de Twitter (25%) e Instagram (15%) (We Are Social: <https://wearesocial.com/sg/blog/2017/01/digital-in-2017-global-overview>). Estos usuarios no representan al conjunto de la población del país. Así, por ejemplo, entre los usuarios de Twitter en España están infrarrepresentadas las mujeres (46%), los mayores de 54 años (6%) y los que no tienen estudios universitarios (59%) (<https://www.masquenegocio.com/wp-content/uploads/2016/01/Twitter-en-España.pdf>). Además, una parte importante de los usuarios de las redes sociales hacen uso de las mismas solo de forma esporádica, con lo que muchos de ellos pueden no quedar registrados en el conjunto de datos con el que trabajemos, reduciendo el nivel de representatividad de la fuente de datos. Finalmente, cuando se hace un filtrado para seleccionar aquellos datos que son más aptos para un determinado estudio se incrementa el sesgo de los datos. Es lo que ocurre, por ejemplo, cuando se seleccionan los usuarios con más actividad en sus teléfonos móviles para reproducir sus trayectorias espacio-temporales en estudios de movilidad o cuando se filtran los tweets para descargar solo los geolocalizados, que representan entre un 1% y un 3% del total de los tweets.

En cuanto al segundo riesgo señalado por Kitchin, el cuarto paradigma no tiene por qué ser una vuelta al empirismo (dejar hablar a los datos), sino que plantea una serie de cuestiones teóricas, que potencialmente pueden abrir nuevas puertas para las investigaciones geográficas a lo largo de múltiples frentes (DeLyser y Sui, 2014). Como señala Batty (2013), el Big Data necesita apoyarse en la teoría; los datos sin teoría carecen de sentido. Dejar hablar a los datos por sí solos, por ejemplo estableciendo relaciones, puede llevar a correlaciones espurias, sin ningún valor explicativo. El análisis de correlación debe estar guiado por una teoría previa que sirva para justificar la lógica de la relación entre variables. Por lo tanto, es necesario un conocimiento de la disciplina específica, no solo conocimientos informáticos. Se requieren equipos interdisciplinares que combinen conocimientos técnicos y temáticos.

4. El análisis de los patrones espacio-temporales de la actividad humana a partir del Big Data

La Geografía siempre se ha ocupado de la distribución de los fenómenos sobre el espacio y de la evolución temporal de esos patrones espaciales. La mayor parte

de los datos masivos son datos geolocalizados y tienen un registro del momento temporal en que fueron producidos. En general, son datos de grano fino, con considerable resolución espacial y temporal, que permiten la monitorización de procesos espacio-temporales de forma antes impensable, prácticamente en tiempo real. Sin embargo, tanto la resolución espacial y temporal de los datos como su contenido es muy variable según las distintas fuentes de datos.

4.1. Resolución espacial

La geolocalización de una gran parte de los datos masivos proviene de los dispositivos GPS alojados en los teléfonos móviles, que almacenan las coordenadas *x y* con un error posicional de unos pocos metros (centímetros si se utiliza GPS diferencial). Esto ocurre, por ejemplo, con los datos de aplicaciones de redes sociales generalistas (como Twitter), de fotografías compartidas (Flickr, Panoramio, Instragram) o de rutas (Strava, Wikiloc).

En el caso de los datos de telefonía móvil, la geolocalización es menos precisa. La actividad del teléfono móvil (llamadas, mensajes, sesiones de datos) es captada por la antena más próxima, con lo que se conoce simplemente que ese dispositivo móvil se encuentra en el área de cobertura de esa antena. Las áreas de cobertura se delimitan generando polígonos de Voronoi a partir de las coordenadas de las antenas, de forma que finalmente se conocerá que ese teléfono móvil en un momento determinado se encontraba dentro del área de cobertura (polígono de Voronoi) de una determinada antena. Dado que la densidad de antenas es mucho mayor en áreas urbanas que en ámbitos rurales, la exactitud posicional oscila en España entre algunos centenares de metros en las primeras y varios kilómetros en las segundas (Picornel et al., 2015). Esa exactitud puede ser mejorada utilizando técnicas de triangulación a partir de las antenas más próximas a la localización del teléfono móvil, de forma que es posible estimar las coordenadas *x y* con un error que en áreas urbanas puede situarse en torno a los 50 o 100 metros.

En otros casos, la geolocalización de los datos viene dada por la del dispositivo fijo que los registra. Esto ocurre, por ejemplo, con los sensores que captan los niveles de contaminación en las ciudades, las TPV (terminales de puntos de venta) que registran los datos relativos a las transacciones que se realizan con tarjetas bancarias, los dispositivos de las bases para el uso de bicicletas públicas de alquiler o las cámaras localizadas en las autopistas que, mediante el reconocimiento de matrículas, permiten hacer conteos de vehículos y realizar los correspondientes cargos por peajes a los usuarios.

4.2. Resolución temporal

También la resolución temporal de los datos es muy variable según el tipo de fuente y el usuario. Los tracks almacenados con el GPS de los teléfonos móviles tienen una resolución temporal muy alta, normalmente de un registro cada pocos segundos. La resolución temporal de los registros activos de los

teléfonos móviles se puede calificar de media o media-alta. Estudios empíricos informan de un registro cada 8,2 horas (González et al., 2008), aunque otros rebajan ese intervalo a solo 260 minutos (Calabrese et al., 2011). Esto son solo medias, muy influidas por valores extremos (algunos usuarios tienen una muy baja actividad en sus teléfonos móviles). Por eso una práctica habitual en los estudios de movilidad realizados a partir de registros de teléfonos móviles es seleccionar aquellos usuarios que tienen mayor resolución temporal, de forma que sea posible conocer con detalle su posición en el espacio y en el tiempo.

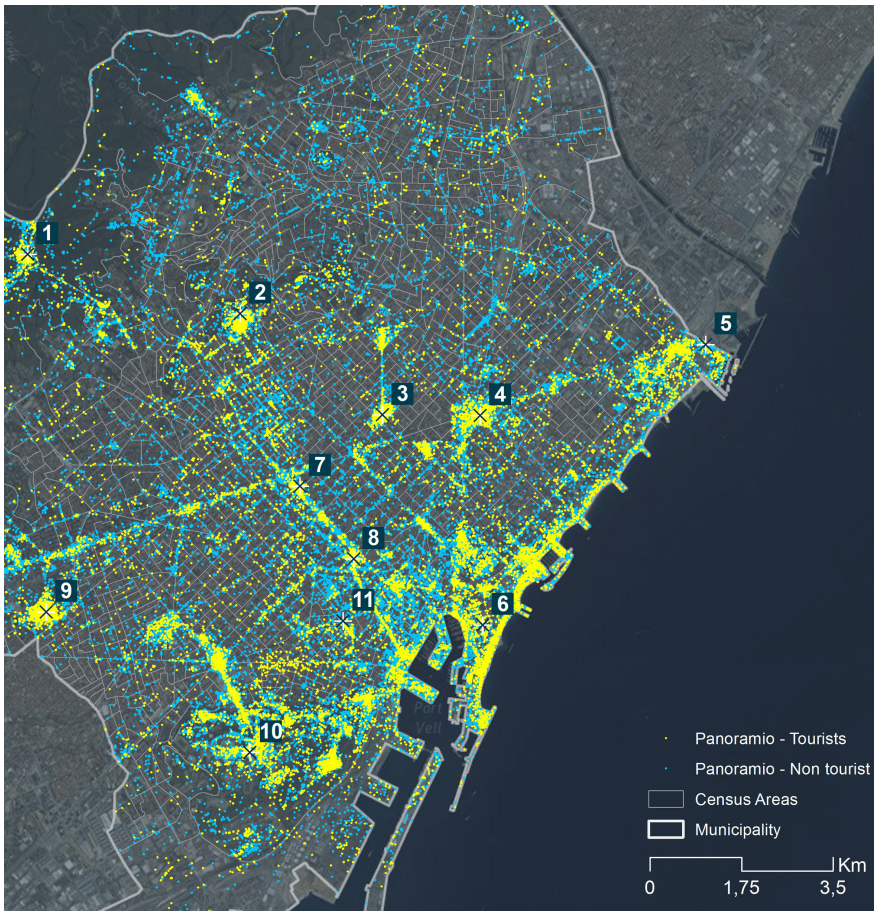
Por término medio, el número de registros diarios que dejan los usuarios de la telefonía móvil es mucho mayor que en el caso de las redes sociales; pero existe una gran variabilidad entre los usuarios de estas redes. Algunos utilizan las redes sociales de forma esporádica, pero otros lo hacen de forma compulsiva, dejando un número mucho mayor de registros. También existen disparidades importantes entre redes sociales. El uso de Twitter o Facebook puede ser diario en muchos casos, pero los datos de fotografías geolocalizadas suelen estar asociados a periodos en los que los usuarios tienen una mayor disponibilidad de tiempo (vacaciones y fines de semana). Los registros con tarjeta de crédito tienen también una resolución temporal relativamente baja, si bien esta resolución aumenta en periodos vacacionales, en los que cambian los estilos de vida y las pautas de consumo. Por su parte, los registros de las tarjetas inteligentes de transporte son diarios en el caso de los usuarios que realizan movimientos recurrentes, pero esporádicos en los demás casos, y solo aportan la localización de los individuos en momentos muy puntuales del día.

La resolución temporal de los datos condiciona el tipo de estudio que se quiera realizar. Si simplemente se quieren conocer las pautas de localización de ciertos fenómenos humanos en el conjunto de varios meses o años, basta con una resolución temporal baja. Cuando se quiere analizar la localización de la población en los distintos momentos del día la resolución temporal de los datos debe ser media-alta o alta, ya que si fuera baja o estuviera temporalmente sesgada nos daría resultados poco fiables. Con los datos de Twitter, por ejemplo, es posible identificar los lugares de residencia y de trabajo de gran parte de sus usuarios, así como otros lugares donde realizan actividades. Esto no sería posible, en cambio, con Flickr, una red social de fotografías geolocalizadas con una granularidad temporal mucho más baja. La resolución temporal debe ser alta cuando se quiere analizar la movilidad diaria de la población, ya que entonces deben identificarse los tiempos de permanencia de cada individuo en una localización concreta y las horas en las que realiza desplazamientos. La fuente más utilizada para este fin son los registros de telefonía móvil, ya que los teléfonos móviles tienen un nivel de penetración muy alto y la resolución temporal de sus registros es media-alta (ver, por ejemplo, Alexander et al., 2015; Tool et al., 2015).

4.3. Pautas espaciales y procesos espacio-temporales

Las fuentes de datos que contienen información sobre las coordenadas *x y* de cada registro permiten elaborar mapas de puntos con facilidad utilizando un

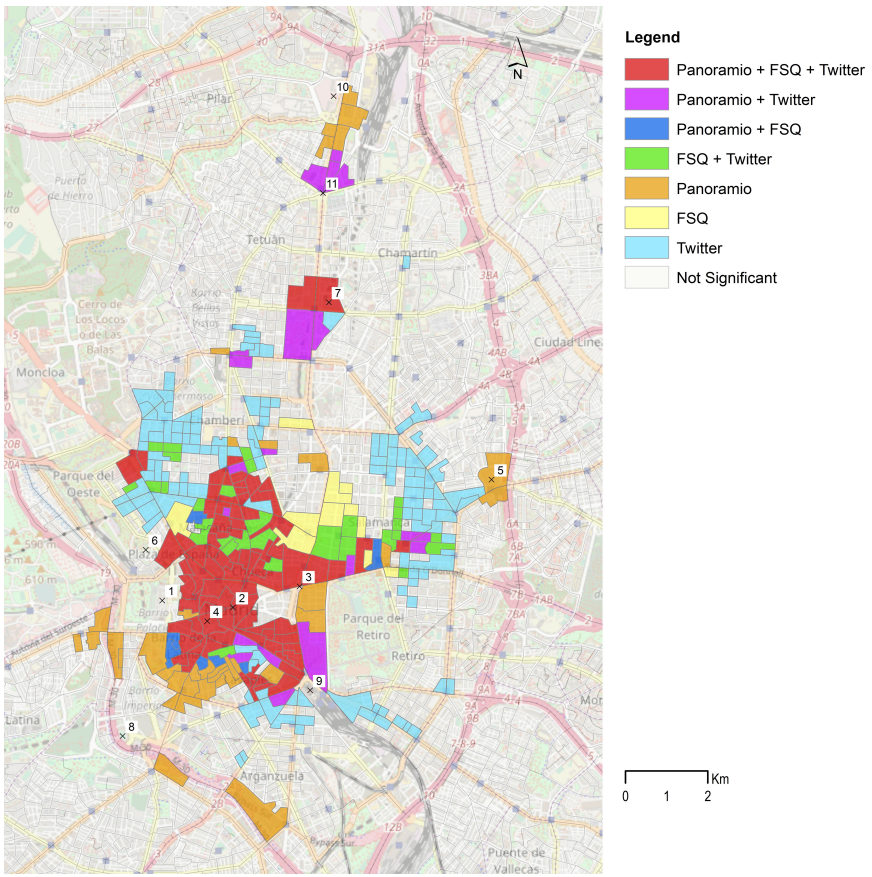
Figura 2. Distribución espacial de las fotografías geolocalizadas tomadas por turistas y residentes en Barcelona en la red social Panoramio. 1) Tibidabo; 2) Park Güell; 3) Sagrada Familia; 4) Torre Agbar; 5) Fòrum; 6) Barceloneta; 7) Casa Milà; 8) Plaza de Catalunya; 9) Camp Nou; 10) Montjuïc; 11) El Raval



Fuente: Gutiérrez et al. (2017).

Sistema de Información Geográfica. Los mapas de tweets (ver, por ejemplo, <http://mappinglondon.co.uk/2014/all-the-tweets>) reproducen la estructura de las ciudades y de las vías de comunicación. De la misma forma, los mapas de fotografías geolocalizadas reflejan la localización de los turistas en los principales puntos de atracción de las ciudades (Gutiérrez et al., 2017) (figura 2). A partir de los mapas de puntos se pueden construir mapas de densidades, que permiten analizar pautas espaciales con mayor facilidad. Así, por ejemplo, la densidad de llamadas telefónicas a distintas horas del día es una buena proxy de la localización de la población, que se concentra en áreas de actividad durante

Figura 3. Clústeres espaciales de turistas en Madrid según Panoramio, Foursquare y Twitter. 1) Palacio Real; 2) Puerta del Sol; 3) Plaza de Cibeles; 4) Plaza Mayor; 5) Plaza de Toros de las Ventas; 6) Templo de Debod; 7) Estadio del Real Madrid; 8) Estadio del Atlético de Madrid; 9) Estación de Atocha - Museo Reina Sofía; 10) Cuatro Torres; 11) Torres Kio



Fuente: Salas Olmedo et al. (2018).

el día y en áreas residenciales durante la noche (ver, por ejemplo, <http://www.envplan.com/misc/b32047>).

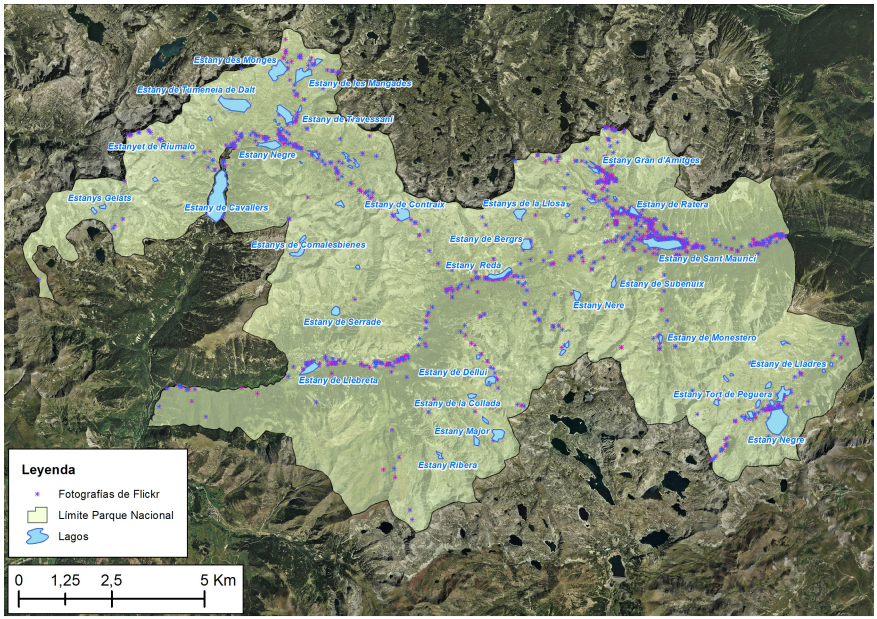
El principal inconveniente de trabajar con mapas de puntos o con mapas de densidades obtenidas directamente a partir de los puntos es que se sobreestima el peso de los usuarios más activos, particularmente los «convulsivos», que dejan muchas más huellas digitales por día que los usuarios medios. Un usuario de Twitter puede enviar varias de decenas de tweets en una hora, mientras que otro puede haber enviado solo uno. Lo que nos interesa no es el número de tweets enviados, sino que el envío de tweets nos indica la presencia de ese usuario en ese lugar y en esa hora del día. Para evitar este

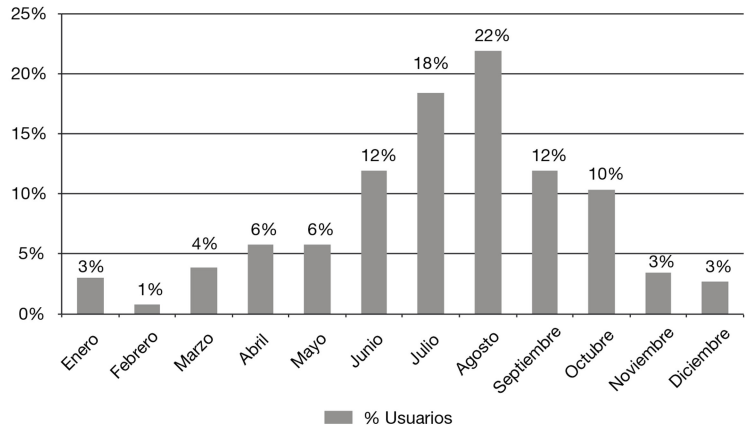
efecto de sobreestimación se puede hacer una agregación espacial y temporal de los datos de forma que tengamos el número de usuarios (y no de huellas digitales) en cada área de la ciudad y franja horaria. Dado que en los registros de estos datos suele aparecer un identificador de usuario, la operación para convertir huellas digitales en usuarios es muy sencilla en un Sistema de Información Geográfica. Una vez que se dispone del número de usuarios únicos en cada zona y franja horaria es posible analizar su evolución temporal (García-Palomares, 2018) y también efectuar comparaciones entre fuentes de datos distintas. Así, por ejemplo, se puede comparar la localización de los turistas normalizando los datos de usuarios únicos obtenidos a partir de distintas fuentes de datos, expresivas de distintos tipos de actividades: las fotografías localizadas son expresivas de los lugares con atractivo visual visitados; los datos de tarjetas bancarias o, en su defecto, de la red social Foursquare, reflejan actividades de consumo; los de Twitter están relacionados con la presencia en los alojamientos, ya que los turistas suelen enviar tweets a última hora de la tarde cuando se conectan a la red Wi-Fi de su hotel o apartamento, etc. (Salas-Olmedo, 2018). El análisis de clústeres espaciales basado en la autocorrelación local permite identificar las áreas de la ciudad especializadas en una o varias de estas actividades (figura 3).

Dado que los registros de las distintas fuentes contienen información del momento en que se dejó la huella digital, es posible analizar procesos espacio-temporales una vez realizada la agregación espacio-temporal de los datos. Así, por ejemplo, la agregación de los registros de fotografías geolocalizadas en un parque nacional por meses permite hacer un análisis de la estacionalidad del fenómeno turístico (figura 4). Así mismo, se pueden seguir procesos de grano fino, como es la variación espacio-temporal de la localización de la población en la ciudad. Agregando los datos por periodos temporales (por ejemplo, cada cuarto de hora) se pueden generar vídeos que muestran la variabilidad de los puntos de calor de la ciudad en función de la distribución cambiante de la población (García-Palomares et al., 2018).

Hasta aquí se ha destacado la información relativa a la localización espacial y temporal de los registros de estas nuevas fuentes de datos. Pero estos registros también contienen otros campos de gran utilidad, que pueden ser utilizados para caracterizar a los usuarios. Algunas fuentes de datos contienen información sobre el domicilio del usuario (telefonía móvil, tarjetas bancarias, tarjeta inteligente de transporte, Flickr). En el caso de que no se disponga de esta información de forma explícita (por ejemplo, Twitter), en muchos casos puede ser estimada a partir del historial del usuario (por ejemplo, el lugar desde donde más tuitea en las horas nocturnas revela la localización del domicilio del usuario). Algunas fuentes también aportan información sociodemográfica del usuario (edad y sexo). Y los registros de transacciones con tarjetas bancarias incluyen el establecimiento en el que se ha hecho el gasto, así como la cuantía del mismo. En general, a medida que aumenta la resolución temporal de las fuentes de datos tiende a decrecer el valor de la información que aportan los atributos de los campos. Los registros de

Figura 4. La huella digital de los visitantes al Parque Nacional de Aigüestortes i Estany de Sant Maurici según Flickr. Arriba: localización de las fotografías; abajo: distribución temporal de los usuarios





Fuente: Carolina Barros, tesis doctoral en elaboración.

vidad de los teléfonos móviles tienen una resolución temporal media-alta pero pocos atributos; los registros de las redes sociales son menos frecuentes, pero contienen campos con información más valiosa (por ejemplo, el texto de un tweet); los registros de transacciones con tarjetas bancarias están mucho más espaciados en el tiempo, pero contienen una información muy rica

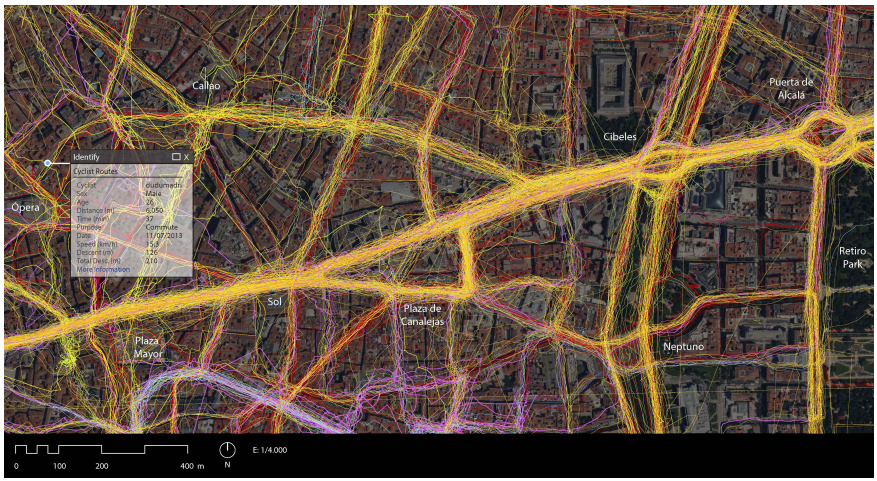
sobre el valor de la transacción, el tipo de establecimiento, características del cliente, etc. Cruzando estos datos y agregándolos geográficamente se puede obtener información de gran interés desde la perspectiva del turismo, por ejemplo gasto general según procedencia, localización de los lugares de consumo y pautas temporales del gasto (ver, por ejemplo, Sobolevsky et al., 2016). Así mismo, es posible evaluar el impacto económico de los eventos turísticos con alto nivel de desagregación espacial, comparando el gasto durante el evento con un periodo anterior normal. Así, por ejemplo, con datos del BBVA se pudo comprobar que la semana del World Pride 2017 supuso una importante inyección económica para la ciudad de Madrid, con un gasto total directo de 115 millones de euros, un 15% más respecto a un periodo normal (<https://www.hosteltur.com/123005_world-pride-tuvo-impacto-115-m-madrid.html>).

4.4. Análisis de movilidad

Las nuevas fuentes de datos también ofrecen información muy útil en los estudios de movilidad. Conociendo la localización cambiante de cada persona a partir de su huella digital se pueden analizar sus pautas generales de movilidad. Obviamente, si la resolución temporal de los datos es alta o media-alta (por ejemplo, registros de actividad de teléfonos móviles) se pueden realizar análisis detallados de movilidad diaria. Si lo que se pretende es analizar movimientos turísticos, fuentes de datos con menor resolución temporal como Twitter o los registros de transacciones con tarjetas bancarias pueden ser suficientes (Hawelka et al., 2014; Sobolevsky et al., 2014; Bassolas et al., 2016).

Si la resolución temporal de la fuente es media-alta (actividad de los teléfonos móviles) es posible identificar los periodos en los que el individuo permanece en un lugar y los viajes entre lugares, en la línea de las trayectorias espacio-temporales de la Geografía del tiempo de la Escuela de Lund. Incluso es posible identificar la convergencia espacio-temporal de individuos con datos de telefonía móvil: cuando dos personas que intercambian llamadas telefónicas se encuentran al mismo tiempo en el mismo lugar se puede asumir que se han encontrado (Picornell, 2015). Los resultados se pueden obtener de forma mucho más rápida y con menor coste que si se realizan encuestas, si bien la información obtenida es menos rica y la localización de los orígenes y destinos resulta no demasiado precisa (basada simplemente en polígonos de Voronoi en torno a las antenas de telefonía). En cualquier caso, una vez obtenidas las trayectorias espacio-temporales de los usuarios es posible analizar las pautas de movilidad diaria y elaborar modelos predictivos (De Domenico et al., 2013) y generar matrices origen-destino que cuantifican el volumen de movimientos según franjas horarias (Cáceres et al., 2007; Alexander et al., 2015; Tool et al., 2015). Así mismo, se han utilizado este tipo de datos para estimar velocidades de desplazamiento y tiempos de viaje (Bar-Gera, 2007) y calcular el tráfico de la red viaria en carreteras que no disponen de estaciones de aforo convencionales (Cáceres, 2012).

Figura 5. Tracks de ciclistas urbanos en Madrid



Fuente: Gustavo Romanillos, <http://www.huellaciclistademadrid.es/>.

También es posible analizar la movilidad de la población a partir de fuentes de datos específicas, como las tarjetas inteligentes de transporte, que registran los viajes realizados por sus abonados en transporte público. A partir de los datos de la tarjeta inteligente de transporte se pueden generar matrices origen-destino (Munizaga et al., 2010) y explorar las dinámicas espacio-temporales y los flujos en transporte público (Tao et al., 2014). Otra fuente muy útil son los registros GPS de los movimientos de los usuarios (tracks), a partir de los cuales se puede identificar el origen y destino de cada viaje, la ruta seguida, las velocidades, tiempos de viaje, pendientes, etc. (Romanillos et al., 2016). La información suministrada por medio de estos registros GPS es muy rica y precisa, pero generalmente solo se dispone de ella para colectivos poco numerosos. Un ejemplo de las rutas registradas con aplicaciones para ciclistas urbanos se puede ver en la figura 5.

4.5. Análisis de accesibilidad

Las nuevas fuentes de datos aportan información muy precisa, según horas del día, de tiempos de viaje entre orígenes y destinos, lo que permite abordar análisis de accesibilidad de forma dinámica, entendiendo por accesibilidad la facilidad para alcanzar los destinos deseados. Así, por ejemplo, la API de Google Maps permite calcular rutas de mínimo tiempo de viaje entre dos puntos considerando el modo de transporte, el día de la semana y el momento del día. En el caso del transporte privado Google Maps utiliza datos históricos (y en su caso también datos obtenidos en tiempo real) generados por millones de teléfonos móviles que cuentan con la aplicación Google Maps. En el caso del

transporte público, los cálculos se realizan a partir de ficheros GTFS (General Transit Feed Specification) facilitados por las autoridades de transporte de distintas ciudades del mundo. A partir de la API de Google Maps se pueden obtener matrices de tiempos de viaje especificando los orígenes, los destinos y la hora del mismo, de forma que se tenga en cuenta el efecto de la congestión en el transporte privado o de la variación de las frecuencias en el transporte público. Varios trabajos han calculado la variación temporal en la accesibilidad en transporte público a los servicios utilizando la API de Google Maps o directamente los ficheros GTFS (Farber et al., 2014; Stępniak y Goliszek, 2017). Modificando los ficheros GTFS se pueden realizar simulaciones sobre cambios en los tiempos de viaje en transporte público producidos por supresión de líneas o modificación de las frecuencias de servicio (Farber y Fu, 2017). También es posible analizar dinámicamente los tiempos de viaje en transporte privado. Además de la API de Google Maps existe un producto de una conocida compañía de navegadores (TomTom Speed Profiles) que ofrece la velocidad de circulación en cada tramo de la red cada cinco minutos, lo que permite realizar análisis comparativos entre ciudades con una altísima resolución temporal (Moya-Gómez y García-Palomares, 2015).

La accesibilidad varía a lo largo del día no solo en función de la variación de los tiempos de viaje en coche o en transporte público, sino también en función de los cambios en la capacidad de atracción de los destinos. La variación temporal de este segundo componente ha sido ignorada en la mayor parte de los estudios dinámicos de accesibilidad. La consideración conjunta de las dos fuentes de variación temporal, ya se trate de transporte público (Boisjoly y El-Geneidy, 2016) o privado (Moya-Gómez et al., 2017), ofrece resultados más realistas y permite analizar el impacto de cada uno de los dos componentes (tiempos de viaje y capacidad de atracción de los destinos) por medio de la construcción de escenarios de evaluación.

5. Consideraciones finales

Las nuevas fuentes de datos permiten abordar viejos problemas de forma nueva y analizar temas que antes quedaban ocultos para el investigador. En los últimos años se ha publicado un gran número de trabajos que utilizan Big Data desde distintas perspectivas y dejan claras las oportunidades que estos datos ofrecen para la Geografía. Los datos masivos suelen tener una alta resolución espacial y temporal, lo que permite la monitorización y el análisis de procesos geográficos. Además, en muchos casos tienen una muy amplia cobertura territorial, lo que facilita la comparación entre diferentes espacios.

Sin embargo, existen barreras que dificultan el uso de estas nuevas fuentes de datos en Ciencias Sociales. Por una parte, existen barreras tecnológicas, ya que es necesario utilizar tecnologías Big Data o, en su caso, realizar operaciones de filtrado en las API para descargar *datasets* que ya sean procesables con tecnologías convencionales. Por otra parte, existen barreras en cuanto a la disponibilidad de los datos, ya que muchos de estos son producidos por

empresas que ven en ellos una oportunidad de negocio y por lo tanto no suelen compartirlos con los investigadores de forma libre. Algunas empresas venden sus datos (una vez anonimizados) y otras venden los estudios que hacen con sus datos. Pero también existen datos que se pueden descargar libremente sin coste alguno, ya se trate de muestras o de *datasets* completos, si bien para ello es necesario que el usuario programe en la API correspondiente para especificar los datos que quiere descargar. También las administraciones públicas se han convertido en productoras de datos masivos (por ejemplo, datos de las tarjetas inteligentes de transporte o del uso de sistemas de bicicletas públicas), si bien suelen mostrarse reticentes a compartir los microdatos con los investigadores por problemas de privacidad.

El creciente uso del Big Data no significa que en el futuro pueda sustituir a las encuestas. Las encuestas seguirán siendo necesarias porque suministran una información que no es aportada por el Big Data. Motivaciones específicas sobre acciones pasadas o futuras pueden ser investigadas por medio de encuestas y difícilmente estas encuestas pueden ser sustituidas por nuevas fuentes de datos. En Internet existen multitud de ficheros de texto, generados por los usuarios, de los que se puede obtener relevante información cualitativa, particularmente mediante técnicas de inteligencia semántica. Es posible medir el grado de satisfacción de los consumidores a través de sus comentarios en la red (por ejemplo, comentarios de los clientes de hoteles en Tripadvisor) o investigar sobre los sentimientos de la población a través de sus comentarios en redes sociales como Twitter (utilizando software de inteligencia semántica). Por lo tanto, antes de preguntar (encuestar) habría que escuchar (analizar ficheros de texto en la red) y una vez descartada la posibilidad de que las nuevas fuentes de datos den respuesta a nuestras preguntas de investigación pasar a encuestar. Analizar información disponible en la red es mucho más rápido y tiene un menor coste que realizar encuestas. Las nuevas fuentes de datos pueden hacer innecesarias algunas encuestas (por ejemplo, la procedencia de los asistentes a una gran manifestación o evento puede ser investigada recurriendo a datos de Twitter o de telefonía móvil), pero otras seguirán siendo necesarias.

También la estadística oficial puede verse afectada por el Big Data. En España, el censo de 2021 no se va a elaborar a partir de una operación censal clásica, sino a partir de registros administrativos (datos del padrón, de Hacienda, de la Seguridad Social, etc.). Estos registros deberán ser cruzados para obtener una información semejante a la facilitada en los censos tradicionales. Sin embargo, parte de la información habitualmente suministrada en el censo, como la relativa a la movilidad diaria o a la población vinculada no puede ser obtenida de los registros administrativos. El INE prevé facilitar esta información a partir del procesamiento de datos de telefonía móvil. Así mismo, la diferenciación entre vivienda principal, secundaria y vacía se puede conseguir recurriendo a los datos de consumo de electricidad de los hogares. Por lo tanto, en el futuro las nuevas fuentes de datos del tipo Big Data pueden jugar un papel relevante no solo en la investigación y en la gestión empresarial, sino también en la producción de estadísticas oficiales y particularmente en la elaboración de los censos.

Referencias bibliográficas

Alexander, L.; Jiang, S.; Murga, M. y González, M. C. (2015). «Origin-destination trips by purpose and time of day inferred from mobile phone data», *Transportation Research Part C: Emerging Technologies*, 58, 240-250.

Bar-Gera, H. (2007). «Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: a case study from Israel», *Transportation Research Part C: Emerging Technologies*, 15 (6), 380-391. <http://dx.doi.org/10.1016/j.trc.2007.06.003>

Bassolas, A., Lenormand, M., Tugores, A., Gonçalves, B. y Ramasco, J. J. (2016). «Touristic site attractiveness seen through Twitter». *EPJ Data Science*, 5 (1), 12.

Bates, J. (2012). «'This is what modern deregulation looks like': co-optation and contestation in the shaping of the UK's Open Government Data Initiative». *The Journal of Community Informatics*, 8 (2).

Batty, M. (2010). «The pulse of the city». *Environment and Planning B: Planning and Design*, 37, 575-577.

— (2013). «Big Data, smart cities and city planning». *Dialogues in Human Geography*, 3 (3), 274-279.

Bloem, J.; Van Doorn, M.; Duivestein, S.; Excoffier, D.; Maas, R. y Van Ommeren, E. (2014). *The Fourth Industrial Revolution. Things Tighten*.

Boisjoly, G. y El-Geneidy, A. (2016). «Daily fluctuations in transit and job availability: A comparative assessment of time-sensitive accessibility measures». *Journal of Transport Geography*, 52, 73-81.

Cáceres, N. (2012). «Traffic Flow Estimation Models Using Cellular Phone Data», *IEEE Transactions on Intelligent Transportation Systems*, 1-12. <http://dx.doi.org/10.1109/TITS.2012.2189006>

Cáceres, N., Wideberg, J. P. y Benítez, F. G. (2007). «Deriving origin-destination data from a mobile phone network», *IET Intelligent Transport Systems*, 1, 15-26. <http://dx.doi.org/10.1049/iet-its:20060020>

Calabrese, F.; Lorenzo, G. D.; Pereira, F. C.; Liu, L. y Ratti, C. (2010). *Analyzing Cell-phone Mobility and Social Events. NetMob-Analysis of Mobile Phone Networks*. Cambridge, MA.

Castell, N.; Dauge, F. R.; Schneider, P.; Vogt, M.; Lerner, U.; Fishbain, B.; Broday, D. y Bartonova, A. (2017). «Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?». *Environment International*, 99, 293-302.

Chen, M.; Mao, S. y Liu, Y. (2014). «Big Data: A survey». *Mobile Networks and Applications*, 19 (2), 171-209.

Chen, C.; Ma, J.; Susilo, Y.; Liu, Y. y Wang, M. (2016). «The promises of Big Data and small data for travel behavior (aka human mobility) analysis». *Transportation Research Part C: Emerging Technologies*, 68, 285-299.

Delyser, D. y Sui, D. (2014). «Crossing the qualitative-quantitative chasm III: Enduring methods, open geography, participatory research, and the fourth paradigm». *Progress in Human Geography*, 38 (2), 294-307.

De Domenico, M.; Lima, A. y Musolesi, M. (2013). «Interdependence and predictability of human mobility and social interactions», *Pervasive and Mobile Computing*, 9 (6), 798-807. <http://dx.doi.org/10.1016/j.pmcj.2013.07.008>

Dewulf, B.; Neutens, T.; Lefebvre, W.; Seynaeve, G.; Vanpoucke, C.; Beckx, C. y Van de Weghe, N. (2016). «Dynamic assessment of exposure to air pollution using mobile phone data». *International Journal of Health Geographic*, 15 (1), 14.

Farber, S.; Morang, M. Z. y Widener, M. J. (2014). «Temporal variability in transit-based accessibility to supermarkets». *Applied Geography*, 53, 149-159.

Farber, S. y Fu, L. (2017). «Dynamic public transit accessibility using travel time cubes: Comparing the effects of infrastructure (dis) investments over time». *Computers, Environment and Urban Systems*, 62, 30-40.

García-Palomares, J. C.; Salas-Olmedo, M. H.; Moya-Gómez, B.; Condeço-Melhorado, A. M. y Gutiérrez, J. (2018). «City dynamics through Twitter: relationships between land use and spatiotemporal demographics». *Cities*, 72, 310-319.

Gonzalez, M. C.; Hidalgo, C. A. y Barabasi, A.-L. (2008). «Understanding individual human mobility patterns». *Nature*, 453 (5), 779-782.

Goodchild, M. F. (2007). «In the World of Web 2.0». *International Journal*, 2, 24-32.

Gray, J. y Szalay, E. (2007). *eScience - a transformed scientific method. Presentation made to the NRC-CSTB*. <http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt>

Gutiérrez, J.; García-Palomares, J. C., Romanillos, G. y Salas-Olmedo, M. H. (2017). «The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona». *Tourism Management*, 62, 278-291.

Hawelka, B.; Sitko, I.; Beinat, E.; Sobolevsky, S.; Kazakopoulos, P. y Ratti, C. (2014). «Geo-located Twitter as proxy for global mobility patterns». *Cartography and Geographic Information Science*, 41 (3), 260-271.

Hey, T.; Tansley, S.; y Tolle, K. I. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

Kaisler, S.; Armour, F.; Espinosa, J. A. y Money, W. (2013, enero). «Big Data: Issues and challenges moving forward». En*: System Sciences (HICSS)*, 2013. 46th Hawaii International Conference. IEEE. 995-1004.

Kitchin, R. (2013). «Big Data and human geography: opportunities, challenges and risks». *Dialogues in Human Geography*, 3 (3), 262-267.

— (2014). «Big Data, new epistemologies and paradigm shifts». *Big Data & Society*, 1 (1), 1-12.

Leszczynski, A. y Crampton, J. (2016). «Introduction: spatial big data and everyday life». *Big Data & Society*, 3 (2), 1-6.

Miller, H. J. (2010). «The data avalanche is here. Shouldn't we be digging?». *Journal of Regional Science*, 50 (1), 181-201.

Moya-Gómez, B. y García-Palomares, J. C. (2015). «Working with the daily variation in infrastructure performance. The cases of Madrid and Barcelona». *European Transport Research Review*, 7 (2), 1-13

Moya-Gómez, B.; Salas-Olmedo, M. H.; García-Palomares, J. C. y Gutiérrez, J. (2017). «Dynamic accessibility using Big Data: The role of the changing conditions of network congestion and destination attractiveness». *Networks and Spatial Economics*.
<https://doi.org/10.1007/s11067-017-9348-z>

Munizaga, M.; Palma, C. y Mora, P. (2010). «Public transport OD matrix estimation from smart card payment system data». En: *Proceedings from 12th World Conference on Transport Research*. Lisboa, Paper, 2988.

Netto, V. M.; Pinheiro, M.; Meirelles, J. V. y Leite, H. (2015). «Digital footprints in the cityscape: finding networks of segregation through Big Data». En: International Conference on Location-Based Social Media Data. Atenas, 1-15.

Picornell, M.; Ruiz, T.; Lenormand, M.; Ramasco, J. J.; Dubernet, T. y Frías-Martínez, E. (2015). «Exploring the potential of phone call data to characterize the relationship between social network and travel behavior». *Transportation*, 42, 647-668.

Romanillos, G. y Zaltz Austwick, M. (2016). «Madrid cycle track: visualizing the cyclable city». *Journal of Maps*, 12 (5), 1218-1226.

Salas-Olmedo, M. H.; Moya-Gómez, B.; García-Palomares, J. C. y Gutiérrez, J. (2018). «Tourists' digital footprint in cities: Comparing Big Data sources». *Tourism Management*, 66, 13-25.

Shelton, T.; Poorthuis, A. y Zook, M. (2015). «Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information», *Landscape and Urban Planning*, 142, 198-211.

Schwab, K. (2017). *The fourth industrial revolution*. Crown Business.

Sobolevsky, S.; Sitko, I.; des Combes, R. T.; Hawelka, B.; Arias, J. M. y Ratti, C. (2016). «Cities through the prism of people's spending behavior». *PloS one*, 11 (2), 1-19.

Stępniak, M. y Goliszek, S. (2017). «Spatio-temporal variation of accessibility by public transport-the equity perspective». En: *The Rise of Big Spatial Data*. *Springer International Publishing*, 241-261.

Tao, S.; Rohde, D. y Corcoran, J. (2014). «Examining the spatial-temporal dynamics of bus passenger travel behavior using smart card data and the flow-comap», *Journal of Transport Geography*, 41, 21-36. <http://dx.doi.org/10.1016/j.jtrangeo.2014.08.006>

Toole, J. L.; Colak, S.; Sturt, B.; Alexander, L. P.; Evsukoff, A. y González, M. C. (2015). «The path most traveled: Travel demand estimation using big data resources». *Transportation Research Part C: Emerging Technologies*, 58, 162-177.

Xia, F.; Yang, L. T.; Wang, L. y Vinel, A. (2012). «Internet of things». *International Journal of Communication Systems*, 25 (9), 1101-1102.