Boletín de la Asociación de Geógrafos Españoles, 79
ISSN: 0212-9426 eISSN: 2605-3322

Cómo citar este trabajo: Zornoza Gallego, C., & Salom Carrasco, J. (2018). *Tweets* geolocalizados para la movilidad diaria: metodología de análisis y detección de residencias en el área urbana de Valencia. *Boletín de la Asociación de Geógrafos Españoles, 79,* 2464, 1–28. http://dx.doi.org/10.21138/bage.2464

Tweets geolocalizados para la movilidad diaria: metodología de análisis y detección de residencias en el área urbana de Valencia

Geolocalized Tweets for assessing daily mobility: methodology to analyse and detect homelocation in the urban area of Valencia

Carmen Zornoza Gallego

carmen.zornoza@uv.es

Julia Salom Carrasco

julia.salom@uv.es

Instituto de Desarrollo Local, Departamento de Geografía
Universidad de Valencia (España)

Resumen

Se analizan los datos geolocalizados de la red social Twitter con el fin de conocer las posibilidades que ofrecen para el estudio de las pautas de movilidad diarias, aplicado al caso del área urbana de Valencia, España. Basado en dicho análisis, se propone una metodología para el tratamiento y explotación de los datos, focalizada en la detección del lugar de residencia de sus usuarios, información básica en el análisis de la movilidad. El buen ajuste de los resultados con distintas fuentes de comprobación ratifica la adecuación de la metodología y las amplias posibilidades de la fuente analizada.

Palabras clave: nuevas fuentes de datos; Twitter; Movilidad diaria; Sistemas de Información Geográfica (SIG).

Recepción: 24.07.2017 **Aceptación:** 30.12.2017 **Publicación:** 15.12.2018

Abstract

Geolocalized data from social network Twitter is analyzed with the aim of studying its possible use in a daily mobility pattern investigation. The area for the practical application is Valencia's urban area, Spain. Based on the previous analysis, a methodological proposal is created to the use of data, focused on the detection of the user's home location, a core information in a mobility study. The proper adjustment of the results with the sources of evidences validates the methodology and shows that the possibilities of this information are vast.

Key words: new sources of data; Twitter; Daily Mobility; Geographical Information Systems.

1 Introducción: nuevas fuentes de datos para el análisis de la movilidad

El acceso a nuevas fuentes de datos o fuentes de datos masivos provenientes del uso de la tecnología, hasta la fecha inexistentes, abre nuevas e importantes posibilidades en el mundo de la investigación.

La generación de esta enorme cantidad de datos, en los países desarrollados, se relaciona con la penetración masiva de la tecnología en las actividades cotidianas de sus habitantes. Cada vez que se realiza una llamada telefónica, se conecta el navegador del vehículo, se busca información en Internet, se utiliza una red inalámbrica, o se envía un correo electrónico, se crea un rastro digital. Así, millones de usuarios dejan rastros de sus actividades cotidianas, que contienen información sobre diversas esferas de su vida. De esta forma se nutre la base de datos más importante que jamás se haya creado sobre el comportamiento humano, planteando un nuevo escenario a las ciencias que lo estudian.

La geografía se ocupa de dichos comportamientos con la particularidad de incluir la componente espacial, poniendo de manifiesto el valor del "dónde" como elemento explicativo básico. Las nuevas fuentes de datos responden a esta necesidad de localización, ya que gran parte de las mismas, o recogen el lugar en el que fueron creados, o es posible inferir su localización con técnicas postproceso. Según apuntan Li et al. (2016), el cambio en la disponibilidad de datos geoespaciales en los últimos años es enorme, ya que éstos hasta hace poco se producían bajo demanda y con un alto coste para las administraciones o empresas, mientras que actualmente la mayor parte de los mismos están siendo creados a través del uso del Smartphone por personas individuales. Sin embargo, este aumento en la disponibilidad de información espacial supone a la vez una oportunidad y un reto para los geógrafos, ya que la espectacularidad de la muestra no debe eclipsar la necesidad de conocer profundamente el tipo de registros que maneja y su validez. Extraer su máxima potencialidad requiere combinar su gestión con nuevos marcos teóricos, estrategias metodológicas y mecanismos de control de resultados.

Uno de los ámbitos de estudio en donde estas fuentes ofrecen mayores posibilidades es en el estudio de la movilidad diaria, un campo de gran interés puesto que permite conocer cómo se estructura la actividad humana dentro un área y diseñar en consecuencia los servicios acordes a las necesidades existentes. Estos estudios han dependido tradicionalmente de datos procedentes de aforos o encuestas, que resultan complejos y costosos para los proyectos, por lo que la posibilidad de recabar información de millones de usuarios en tiempo real y sin coste específico para la investigación resulta, como mínimo, a tener en cuenta.

En los últimos años, diversos investigadores han abordado el estudio de la movilidad diaria utilizando fuentes de datos masivos. Estas fuentes permiten, bajo determinadas circunstancias, realizar un seguimiento de las trayectorias diarias de movilidad de la población a partir de la geolocalización de su actividad en las redes de comunicación (llamadas telefónicas, mensajes en redes sociales, etc.) a lo largo del día. Sus resultados han permitido avanzar ampliamente en el modelado de los patrones de movilidad, formulando nuevos retos y forjando las bases de nuevos avances. A continuación, se presentan algunas de las investigaciones de referencia en el área, que constituyen la base de la propuesta metodológica desarrollada en este artículo.

El grado de regularidad espacial y temporal en las trayectorias humanas fue estudiado por González et al. (2008) a partir del seguimiento de las llamadas telefónicas de 100.000 usuarios durante seis meses. Observaron cómo cada individuo tiene un número reducido de lugares altamente frecuentados. La integración de los patrones de todos los usuarios, a pesar de la aparente diversidad en los recorridos, dio como resultado una única distribución, lo cual indica que los humanos siguen patrones simples reproducibles en el tiempo y en el espacio. En este sentido trabajaron también Song et al. (2010), quienes cuantificaron el límite en la predicción de la movilidad humana en un 93%. En su estudio, fueron capaces de reconocer la poca variación en la predictibilidad, que además era independiente de las distancias que cada usuario cubre regularmente.

Las investigaciones de Kwan (1999) y Huang et al. (2015) demostraron que se pueden obtener pautas generales de movilidad a partir de las individuales, hecho fundamental para plantear un estudio de la movilidad diaria a partir del seguimiento de una muestra de ciudadanos. Uno de los estudios de referencia para la presente investigación es el de Huang et al. (2015), donde se analizan las pautas de movilidad humana empleando una perspectiva individual espacio-temporal a partir de registros Twitter. En ese trabajo, la recopilación de datos durante un amplio periodo de tiempo les permitió obtener la evolución temporal de las posiciones individuales, diferenciando patrones regulares e irregulares de desplazamiento.

Uno de los puntos críticos de la investigación de la movilidad a partir de datos masivos es la detección del lugar de residencia de los usuarios. La identificación del lugar de residencia del

usuario es el punto de partida necesario para realizar el seguimiento de su trayectoria individual de movilidad diaria, e identificar y caracterizar los distintos desplazamientos que realiza, vinculándolos a las diferentes actividades (trabajo, ocio, compras, etc.) y los diversos espacios geográficos.

En este sentido, los resultados de Jurdak et al (2015) concluyen la existencia de correlación entre el lugar desde donde más tweets realiza un usuario, denominado posición dominante, y su residencia. Estos autores emplean diversos procedimientos de filtrado de datos orientados a conocer y eliminar sesgos provenientes de la información. En primer lugar, a partir de distintas funciones, definen una distancia de 100 m. como umbral máximo para determinar movimientos dentro lo que consideran un mismo lugar. Por otra parte, partiendo de la idea de que los 140 caracteres que contiene un tweet puedan ser escasos para que un usuario se exprese, consideran la posibilidad de que algunos de ellos generen tweets consecutivos desde una misma ubicación, lo que sesgará el cálculo de posiciones dominantes. Esta posibilidad se incrementa ante ciertos eventos que aumentan la utilización de Twitter en un momento determinado (deportivos, políticos, sociales...).

Por otra parte, la necesidad de diferenciar el tipo de actividades realizadas en las distintas ubicaciones, y, con ello, el motivo del desplazamiento, se recoge en los trabajos de Hasan (2013) y Gabrielli et al (2014). Estos autores combinan datos de Twitter y Foursquare para caracterizar los desplazamientos urbanos en función de la finalidad de los mismos, lo cual les permite subrayar la importancia de diferenciar las tipologías de las posiciones observadas para ajustar los resultados.

En España, la utilización de nuevas fuentes aplicadas a la geografía aún no se encuentra muy extendida. Pese a ello, comienzan a surgir buenos ejemplos de su potencial. Uno de ellos es el estudio de la movilidad humana desde la perspectiva socioeconómica de las diferentes regiones españolas realizado por Llorente et al (2015). En el trabajo citado, los autores relacionan los ritmos diurnos observados, los patrones de movilidad y los tipos de comunicaciones con la tasa de desempleo. Los resultados muestran que esta tasa impacta en los demás factores, y que las horas de uso de la plataforma en empleados y desempleados varían, permitiendo obtener información asociada.

Por otra parte, Bejar et al (2016) analizaron los patrones espaciotemporales que siguen los usuarios de Twitter e Instagram en las ciudades de Barcelona y Milán. A través de diversos algoritmos de frecuencia detectaron las rutas más comunes, diferenciando entre turistas y residentes. El uso de clústeres relacionaba la frecuencia de visita a los lugares con la cantidad y tipo de usuario. Sus resultados recogen puntos turísticos, puntos asociados al ocio de residentes, puntos asociados al transporte público, o lugares asociados a eventos puntuales.

Una buena parte de los trabajos han sido desarrollados en el marco de la geografía del turismo. García-Palomares et al (2015) tomaron como fuente fotografías geolocalizadas de Panoramio para identificar las mayores atracciones turísticas en diversas ciudades europeas. Los resultados,

diferenciados según se trate de residentes o turistas, muestran distintos patrones de concentración que ayudan a identificar áreas de saturación o a establecer controles para no sobrepasar la capacidad de carga. También en el ámbito del turismo han trabajado Serrano-Estrada et al. (2014), empleando datos de Twitter y Foursquare. Con el fin de analizar los lugares preferentes en ciudades turísticas, analizan el caso de Benidorm, comparando los registros en temporada alta y baja, de forma que es posible reconocer el modelo turístico implantado.

Otra aplicación frecuente es la delimitación de usos del suelo, especialmente en espacios urbanos. En este sentido, Frias-Martinez et al (2014) delimitaron usos del suelo a partir de tweets geolocalizados. La combinación de la componente espacial y temporal de los tweets permite determinar usos del suelo que varían en función del momento temporal, información novedosa y útil para caracterizar de forma más precisa las dinámicas urbanas.

Sin embargo, tal y como se ha dicho, el uso de estos datos para resolver problemas geográficos exige un análisis cuidadoso, así como el desarrollo de nuevos marcos teóricos, estrategias metodológicas y mecanismos de control de resultados. Todas estas fases se desarrollan aquí aplicadas al estudio de la movilidad en el área urbana de Valencia a partir de los tweets geolocalizados. Específicamente, se propone una metodología para la detección de residencias, empleando como herramienta fundamental los Sistemas de Información Geográfica (SIG). Una vez identificado el lugar de residencia del usuario de Twitter, será posible diferenciar entre los mensajes emitidos por personas residentes y no residentes (evaluando así la ocupación "temporal" de ese espacio), así como realizar un seguimiento de los desplazamientos de los residentes fuera del domicilio, obteniendo distintos parámetros que permitirán conocer sus pautas de movilidad.

El trabajo se expone comenzando con un apartado general que engloba las consideraciones previas para abordar un estudio geográfico basado en nuevas fuentes, en particular: los tipos de fuentes existentes, su posibilidad de acceso, y su adecuación al estudio a realizar. A continuación, se desarrolla la parte metodológica, dividida en dos apartados: el análisis exploratorio de la fuente (determinante para la construcción del proceso de depuración y tratamiento de los datos), y el desarrollo metodológico propiamente dicho, que consta de dos etapas: el filtrado de la información, que permite la detección del lugar de residencia de los usuarios, y la fase de validación, realizada tanto a partir de datos cuantitativos como cualitativos. El proceso concluye que la fuente y la metodología son adecuadas para el tipo de estudio diseñado.

2 El uso de las nuevas fuentes de información en una investigación geográfica: consideraciones previas

Este primer apartado recoge aspectos generales a tener en cuenta sobre cómo abordar una investigación geográfica basada en las nuevas fuentes de información, cuyas características particulares y novedosas difieren susceptiblemente de las tradicionales.

Un primer paso fundamental es el conocimiento y comprensión del origen de los datos, diferenciando entre las fuentes disponibles en función de quién (personas o sensores), y cómo se genera la información, y teniendo en cuenta la intencionalidad del proceso. Aunque mayoritariamente las fuentes a utilizar son las que se derivan de la actividad humana, es necesario tener en cuenta la existencia de sensores, ya que en muchos casos estas tipologías no son estancas, y se debe discernir entre unos y otros comportamientos.

La creación de datos por parte de sensores se enmarca en la tecnología denominada Machine-to-Machine, en la que los dispositivos capturan eventos (velocidad, temperatura, presión, salinidad, etc.) que se transmiten a otras aplicaciones, generando de este modo información significativa. También forman parte de este grupo la información biométrica (huellas digitales, escaneo de la retina, genética, etc.) o la monitorización de especies.

Por su parte, los datos creados por personas a través de la tecnología se engloban en cuatro tipos principales de fuentes: Redes sociales (Facebook, Twitter, LinkedIn, blogs…), comunicaciones móviles (llamadas, mensajes, localizaciones GPS, conexiones wifi…), transacciones (compras por internet, uso de tarjetas, actividades bancarias…), y páginas web colaborativas (Open Street Map, Wikipedia…). Una de las características que más difiere de las fuentes tradicionales es que en la mayor parte de los casos el fin con el que se generan no es el propio dato, sino que atiende a otras necesidades. Así, el uso de redes sociales, dispositivos móviles o transacciones financieras produce datos con el objetivo de comunicarse, comprar, etc., que no tienen que ver con su uso analítico potencial. A este aspecto se refería Goodchild (2007) cuando denominó este proceso como "Citizens as sensors" (Ciudadanos como sensores), apuntando a la idea de la falta de intencionalidad de los usuarios en generar información. De los cuatro tipos señalados, sólo los datos de las páginas web colaborativas se han producido con el objetivo mismo de ofrecer dicha información. Por ejemplo, la plataforma Open Street Map se nutre de voluntarios que digitalizan redes viarias a nivel mundial con el fin de que estén disponibles de forma libre. Este caso encaja dentro de lo que el mismo Goodchild (2007) denominó Volunteered Geographic Information (VGI).

Finalmente, una tercera posibilidad, empleada frecuentemente en proyectos de investigación, es la creación de receptores propios de información, como podrían ser la instalación de redes wifi o la implantación de tarjetas de transporte específicas.

Una excelente revisión de fuentes y trabajos de investigación a partir de datos masivos geolocalizados se encuentra en Gutiérrez-Puebla et al (2016). En este trabajo se muestran diversos ejemplos de aplicación en la investigación, diferenciándolos según el tipo de fuente: registros de llamadas de teléfonos móviles, redes sociales, comunidades de fotografías geolocalizadas, registros de transacciones con tarjetas de crédito, tarjetas inteligentes de transporte, navegadores, etc.

Una vez se han considerado las distintas fuentes con las que se podría abordar un proyecto de investigación, la gran cuestión a tratar es el acceso a las mismas. Éste resulta ser generalmente el factor inicial clave, reduciendo el espectro de posibilidades considerablemente. Se está generando la mayor fuente de datos de la historia, pero son distintas empresas quienes los almacenan y hacen el uso que consideren oportuno, dentro de la legalidad pertinente. Por lo tanto, si se pretende trabajar con datos generados o almacenados por una empresa externa, se deberá establecer un marco de colaboración con la misma. Es decir, que para trabajar con registros de tecnología móvil o de tarjetas de crédito se necesitará la cooperación de empresas de dichos sectores. Por el contrario, en el caso de redes sociales es más habitual que las plataformas pongan cierta cantidad de información accesible a través de descarga.

Una consideración adicional, y determinante, es la adecuación de la fuente al objetivo de estudio. Ante el diseño de cualquier investigación, la adecuación entre los datos y los objetivos de la misma es esencial. Es decir, los datos con los que se cuenta deben cubrir las necesidades y permitir la obtención de resultados óptimos. En este sentido, las expectativas creadas en torno a las nuevas fuentes son enormes, pero también son muchas las dudas que se plantean ante su utilización en un proyecto de investigación. Se hace imprescindible conocer bien la adecuación de la muestra al trabajo a realizar porque éste será el factor definitorio de la calidad de los resultados.

En primer lugar, ya que los datos han sido producidos con un objetivo distinto al de la propia investigación, es necesario analizar si la fuente evaluada permite lograr los resultados de buscados. Por ejemplo, los datos de la red social Instagram se crean con el fin de compartir fotografías con cierto grupo de personas, por lo que hay que considerar cuidadosamente si se puede, cómo, y con qué restricciones, conseguir información de movilidad, turismo o consumo a través de ellos.

En segundo lugar, en un estudio con contenido territorial, es importante conocer la precisión espacial real de la muestra, ya que, pese a que se detallan unas coordenadas determinadas, no suelen reflejarse las especificaciones técnicas de su creación.

Finalmente, en las investigaciones referidas a comportamientos humanos una cuestión esencial es la posibilidad de caracterización de las personas que integran la muestra. En el caso de las fuentes

masivas, aspectos como la distinción de género, edad, estudios, nivel socioeconómico, etnia…, no se recogen de forma directa. Aunque es posible hacer análisis secundarios o combinaciones con otros datos para deducir algunos de ellos, se trataría siempre de información indirecta. La caracterización por edad es especialmente significativa si se tiene en cuenta que el manejo de las tecnologías puede eliminar de la muestra a sectores de la población especialmente vulnerables, como la población anciana o infantil.

3 Una aplicación práctica: la detección de residencias a partir de Twitter en el área urbana de valencia

Con el fin de ejemplificar las exigencias de una investigación basada en las nuevas fuentes, en las páginas siguientes se desarrolla una metodología dirigida a conocer la movilidad de la población a partir de *tweets* georreferenciados, en el área urbana de Valencia. En primer lugar, se profundiza en las características de los datos, tanto en general (punto 3.1) como en el área específica de estudio (punto 3.2), para, a continuación, diseñar una metodología que tiene como objetivo detectar el lugar de residencia de los usuarios clasificados como residentes, eliminando los turistas. El filtrado de datos (apartado 3.3), basado en un análisis combinado de posiciones espaciales y temporales (información directa) con datos sobre el tipo de usuario (información derivada), es fundamental para eliminar aquello que no es útil para el estudio. Se emplean como herramienta los SIG, ya que permiten combinar el tratamiento de datos alfanuméricos con su localización espacial, dotando al estudio de su componente territorial básica. Tras la detección de residencias se efectúan dos validaciones distintas, una cuantitativa a partir del padrón continuo (apartado 3.4) y otra cualitativa a partir del propio contenido de los tweets (3.5). El proceso permite concluir que la fuente y la metodología son adecuadas para el tipo de estudio diseñado.

3.1 Análisis exploratorio de la fuente: La Red Social Twitter

Al abordar el estudio de la movilidad con datos masivos, el primer problema que se plantea es el posible acceso a las fuentes. En este sentido, la opción más factible es el uso de redes sociales, ya que en muchas de ellas se pueden conseguir los datos gratuitamente utilizando distintos métodos de descarga. La necesidad de contar con la colaboración de empresas de telefonía móvil o entidades bancarias para tener acceso a datos de llamadas o transacciones desechó esta posibilidad, al menos por el momento.

A este respecto, Twitter presenta algunas ventajas, ya que permite descargar en tiempo real un porcentaje de la información que se genera en su plataforma. Para ello, se requiere la programación de la API de Twitter que ofrece en tiempo real la información georreferenciada que se produce en el área seleccionada. Esta red resulta además óptima por el amplio número de

usuarios en España, y por la posibilidad de añadir una localización específica a cada uno de sus tweets.

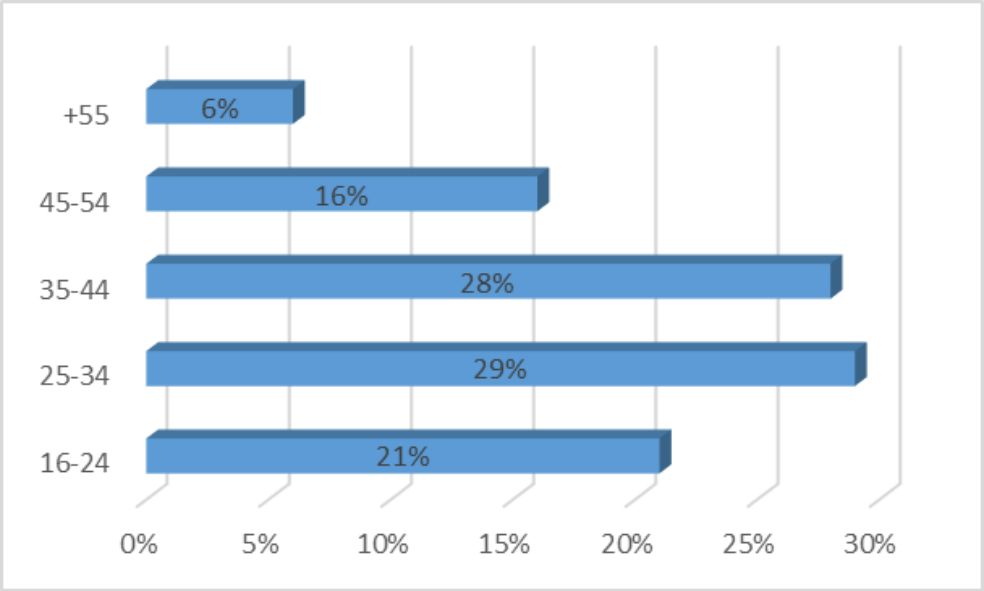
El siguiente paso es ver la adecuación que presenta la información proporcionada por los tweets geolocalizados al objetivo de revelar patrones asociados a la movilidad del área. Para reconocer dicha adecuación, se parte de la información básica que existe para cada tweet georreferenciado, siendo la siguiente:

- Latitud, longitud
- Fecha y hora de producción del tweet.
- Identificador único del usuario.
- Nombre de usuario.
- Localización: parámetro seleccionado por el usuario que no tiene porqué corresponderse con una localización real.
- Idioma: seleccionado por el usuario; no tiene porqué corresponderse al utilizado en la generación de tweets.
- Fuente del tweet: Twitter, Instagram, Foursquare, Endomondo, etc. Permite conocer el tipo de datos que comparte y constituye la base para averiguar cómo se generan las posiciones espaciales, y su precisión asociada.
- Identificador único del tweet.
- Texto del tweet.

Las coordenadas geográficas, junto con la fecha y la hora en que se produce cada tweet, constituyen la información básica para modelar un patrón espacio-temporal. Además, el hecho de que cada usuario tenga un identificador único permite también hacer un seguimiento individualizado de los trayectos, fundamental para conocer sus pautas de movilidad.

Una consideración importante es el grado de representatividad de la muestra. A este respecto, y como se ha comentado anteriormente, conocer las características demográficas de la población analizada resulta crucial para cualquier estudio de ámbito social. Las estadísticas de Twitter publicadas en el informe "Twitter users in Spain" de enero 2016 permiten abordar esta cuestión. El estudio demográfico muestra que los usuarios masculinos son mayoría, con un 54 %, frente a los femeninos, con un 46 %, lo cual indica un pequeño sesgo de género. La distribución poblacional de la Figura 1 muestra que no existe información para menores de 16 años y que hay poca representatividad en mayores de 55 años.

Figura 1. Porcentaje de usuarios Twitter por grupos de edad



Fuente: Informe "Twitter users in Spain" (Twitter, 2016)

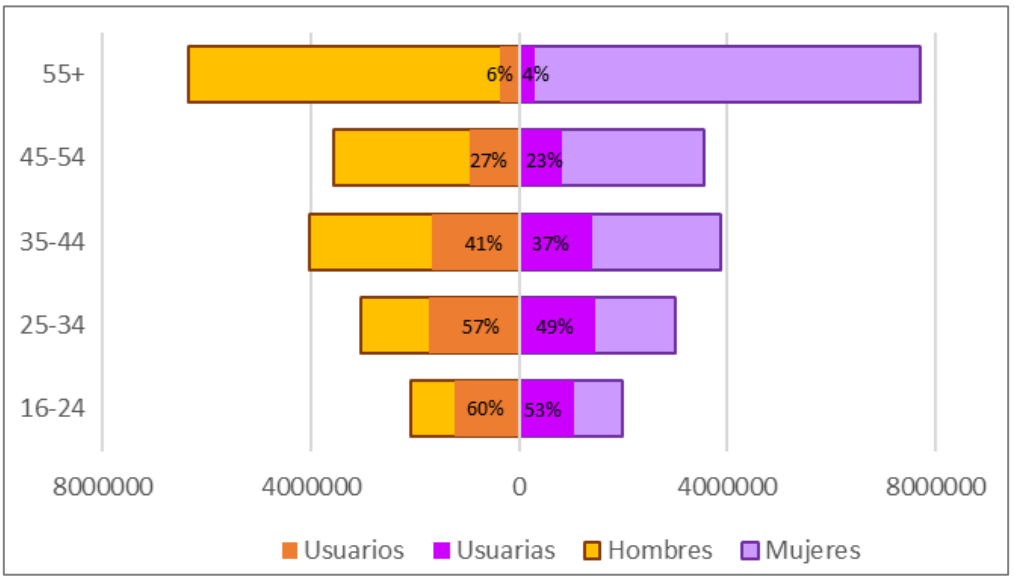
Al combinar la distribución por sexo y edad del número de cuentas Twitter en España (cifrado en 11 millones según la propia plataforma en 2016) con la estructura por edad y sexo de la población española de acuerdo con el Padrón continuo de población a 1 de Enero de 2015, se estima la ratio de penetración de Twitter por edades y género que se muestra en la Figura 2.[1] Los valores mostrados en este gráfico deben tomarse como valores máximos, ya que existen factores que no han podido ser evaluados por falta de información, como la existencia de cuentas de empresas, cuentas que no se han dado de baja pero que se encuentra inactivas, o duplicidad de cuentas.

En cualquier caso, los resultados permiten conocer mejor la población sobre la cual se trabaja. Como se ha comentado, el sesgo tecnológico se hace patente en los mayores de 55 años, entre los que solo un 6 % de los hombres y un 4 % de las mujeres tiene un perfil abierto en la plataforma. El caso de la población infantil (menores de 16) es distinto, ya que, aunque las redes sociales y las tecnologías forman parte de su vida, las políticas de uso y la legislación vigente no permiten ofrecer información al respecto. El grupo de edad con mayor ratio es el comprendido entre los 16 y 24 años, donde el 60 % de los hombres y el 53 % de las mujeres tiene un perfil abierto. Los datos obtenidos para el grupo de edad de 25 a 34 son ligeramente inferiores a los anteriores, mientras que en el grupo de 35–44, el porcentaje se reduce a un 41% de los hombres y un 37 % de las mujeres. Finalmente, para el grupo de edad entre 45 y 54 la ratio es de 27 % y 23 %, respectivamente.

1 Ratio de penetración por edad y sexo= (total usuarios Twitter* % grupo de edad*% sexo)/(total población del grupo de edad y sexo según el padrón)*100

Una lectura prospectiva de estos resultados permite suponer que el alto grado de penetración tecnológica en la sociedad hará que en un plazo temporal medio los rangos de edad sigan ampliándose y este sesgo se reduzca e incluso se elimine. Esto podría ser una de las mayores potencialidades de los datos provenientes de la tecnología.

Figura 2. Ratio de penetración Twitter por edades y sexo



Fuente: elaboración propia a partir de datos de Twitter (2016) e INE (2015)

Un último aspecto a tratar, de gran relevancia en el caso, es la precisión espacial de la fuente. Las coordenadas geográficas de un tweet georreferenciado tienen 6 decimales, lo cual equivale a menos de 10 cm., precisión que en principio resulta difícil de aceptar, ya que mediciones de tal exactitud requieren equipos de tecnología superior a la de un teléfono móvil. Sin embargo, como veremos a continuación, no es la tecnología del dispositivo lo que finalmente determina la precisión de la georreferenciación del mensaje.

A este respecto, hay que tener en cuenta que la información georreferenciada descargada de la plataforma de Twitter no proviene únicamente de esta red social, sino que es utilizada de enlace con otras plataformas. Es decir, que un usuario que genere un contenido en cualquier otra plataforma (Instagram, Foursquare…) puede volcarlo en su cuenta Twitter simultáneamente. Así, en el caso de estudio, las fuentes de los datos se distribuyen inicialmente según la Tabla 1.

Tabla 1. Fuente inicial de los tweets[2]

Fuente	Tweets	Porcentaje
Instagram	80 254	45,39 %
Twitter	65 368	36,97 %
Foursquare	7725	4,37 %
Tráfico	6341	3,59 %
Tendencias	6089	3,44 %
Tiempo	5391	3,05 %
Otros	5635	3,19 %

Fuente: elaboración propia a partir de datos de la API Twitter (2015–2016)

Como puede verse, existen diversas fuentes que vuelcan información en Twitter; de ellas, las tres principales, Instagram, Twitter y Foursquare, suponen el 86,73 % del total. De esto se deriva que el 63,03 % de los tweets provienen inicialmente de plataformas distintas a Twitter. Esto es importante, ya que cada una de ellas georreferencia bajo sus propios parámetros, por lo que es necesario conocer su funcionamiento básico.

Así, en Twitter hay dos maneras de usar servicios de ubicación:

- Pulsando el marcador de "localización" al redactar el tweet y seleccionando de forma manual a partir del listado que ofrecen la localización que se desea etiquetar, o
- Activando el botón específico de "Compartir localización exacta", de forma que la aplicación accede al GPS y comparte la latitud y longitud con la precisión del dispositivo GPS del móvil

La API no incluye información del procedimiento por el que se ha geolocalizado el tweet, existen datos de distinta precisión que provienen de la misma fuente.

En Instagram, plataforma que produce más tweets geolocalizados que el propio Twitter, los servicios de ubicación funcionan como en la primera opción de Twitter; es decir, que es el usuario quien selecciona su posición a partir de un listado específico. Este listado no contiene una gran base de datos con el callejero, sino que principalmente contiene municipios, barrios, comercios o lugares emblemáticos. Esto hace que las posiciones reveladas desde Instagram sean más generales, ya que no permite usar la localización GPS ni contiene un buen callejero.

El proceso de georreferenciación de Foursquare es similar al de Instagram, pero a partir de listados de direcciones propios. Es decir, se seleccionan lugares específicos, pero cuyas coordenadas no coinciden con las de Instagram ni con las de Twitter.

2 Bajo la referencia "Tráfico" se han agrupado las plataformas de la Dirección General de Tráfico y Tuimobility; "Tendencias" engloba Tendencias Valencia y Trensmap; y "Tiempo" incluye SandaySoft y Rain-alarm.

Por tanto, la precisión de los datos es distinta según su procedencia y, sobre todo, según la forma de interactuar de cada usuario. Esto supone que la precisión del GPS de los dispositivos no sea crucial, pero sí lo sea reconocer el tipo de información que cada usuario decide mostrar de sí mismo. Es decir, un usuario que crea un tweet geolocalizado y selecciona como localización del listado "Comunidad Valenciana" o "Valencia", no proporciona una información geográfica válida para el estudio, ya que la referencia espacial no es significativa. Se concluye que no existe una misma precisión espacial para toda la fuente, sino que cada uno de los registros contiene la suya propia, y ésta varía en función de la decisión del usuario que lo genera.

En consecuencia, aunque la revisión de las características de la fuente indica una buena adecuación general de la información proporcionada a los objetivos del trabajo, existen aspectos que deben ser objeto de tratamiento específico en el proceso de depuración y análisis de los datos para su uso en estudios de movilidad.

3.2 Descripción de la muestra y patrones temporales de uso

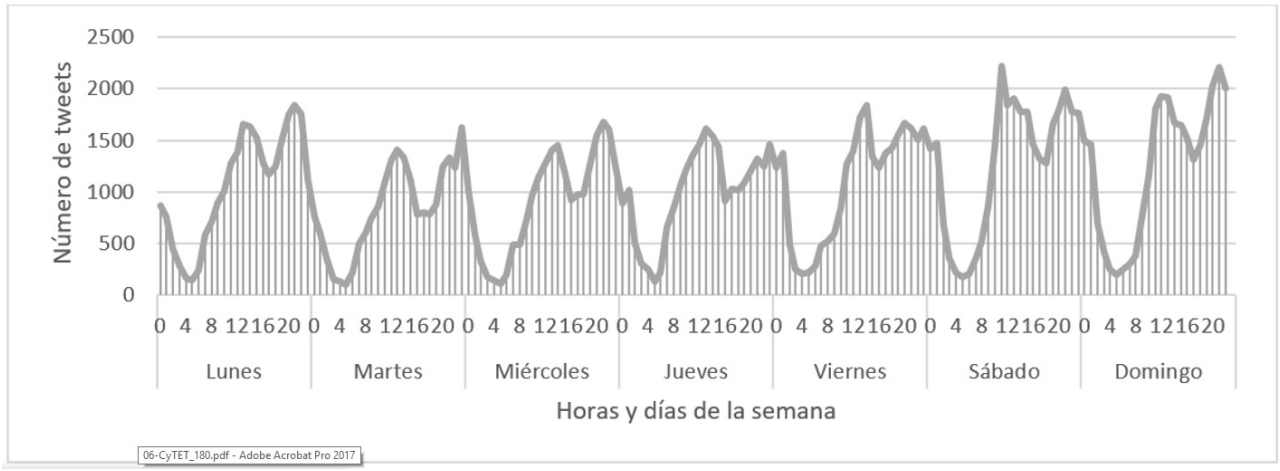
En este apartado se presentan los resultados de un primer análisis centrado en los patrones temporales de uso, una información relevante para la detección del lugar de residencia del usuario y, por ende, para el estudio de sus patrones de movilidad.

Aunque la investigación general se enmarca en un estudio de la movilidad en el Área Metropolitana de Valencia, en una primera aproximación se ha ampliado esta área para aumentar la cantidad de municipios testados y ver qué ocurre en las zonas periféricas con menor población. Además, este procedimiento deja abierta la posibilidad de trabajar con distintas delimitaciones del espacio metropolitano, y observar si los patrones de movilidad marcados por Twitter responden mejor a unas u otras.

El área de estudio cubre una superficie de 10.763 km^2 que incluye 2.544.264 habitantes, una dimensión que condiciona los resultados, ya que su tamaño es mediano en relación a los ámbitos tratados en estudios de otros autores. Esto supone que la generación de datos sea menor a la de otros entornos analizados desde esta perspectiva. La recopilación de información se realizó en el periodo comprendido entre las fechas 5/06/2015 y 21/02/2016, habiendo sido posible recabar información de 146 días, 22 641 usuarios y un total de 177 675 tweets. Esto supone una media de 1217 tweets diarios, y un promedio de 7'84 tweets por usuario a lo largo del periodo analizado.

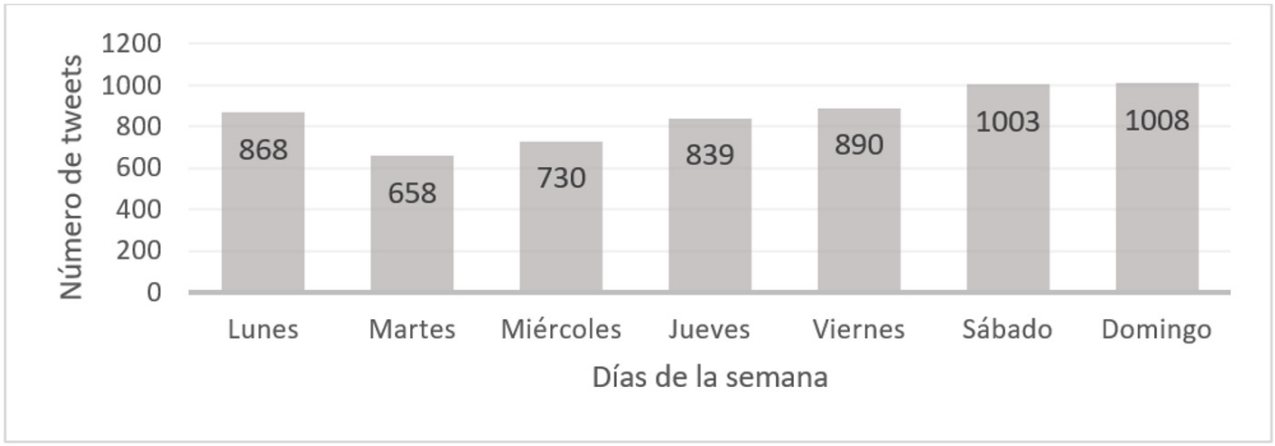
La Figura 3 representa una primera aproximación al patrón temporal de la información recogida que presenta fuertes variaciones entre horas y días de la semana.

Figura 3. Patrón temporal uso Twitter



Fuente: elaboración propia a partir de datos de la API (2015–2016)

Figura 4. Media tweets por hora y día de la semana



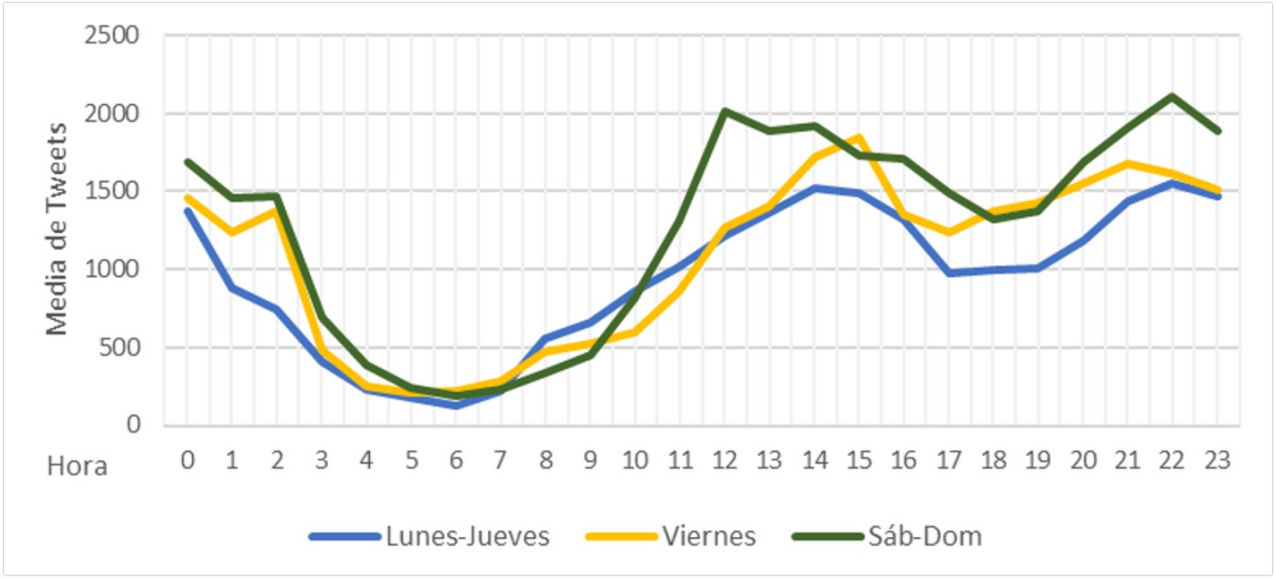
Fuente: elaboración propia a partir de datos de la API Twitter (2015–2016)

Aunque en general los patrones de actividad horaria diarios son similares, existen ciertas diferencias entre los días entre semana y los fines de semana. La Figura 4 permite comparar las diferencias de comportamiento entre los distintos días, y muestra que el uso medio diario es mayor durante el fin de semana que entre semana, y que la actividad de martes y miércoles es especialmente baja.

Por su parte, la Figura 5 muestra el patrón horario de emisión de los tweets agrupando por un lado los emitidos en el periodo que va de lunes a jueves, los de sábado y domingo por otro lado, y un tercer grupo constituido por los tweets de los viernes.

En general, las horas de menor actividad se dan de madrugada, desde las 3 a las 8 de la mañana y hasta las 9 los fines de semana. La actividad matutina semanal aumenta progresivamente hasta que alcanza su máximo entre las 14 y las 15 horas, momento en el cual disminuye hasta las 17 horas, donde mantiene una baja actividad hasta las 19 horas. A partir de este momento vuelve a aumentar, registrando su máximo diario a las 22 horas.

Figura 5. Media Tweets por hora y grupo de días

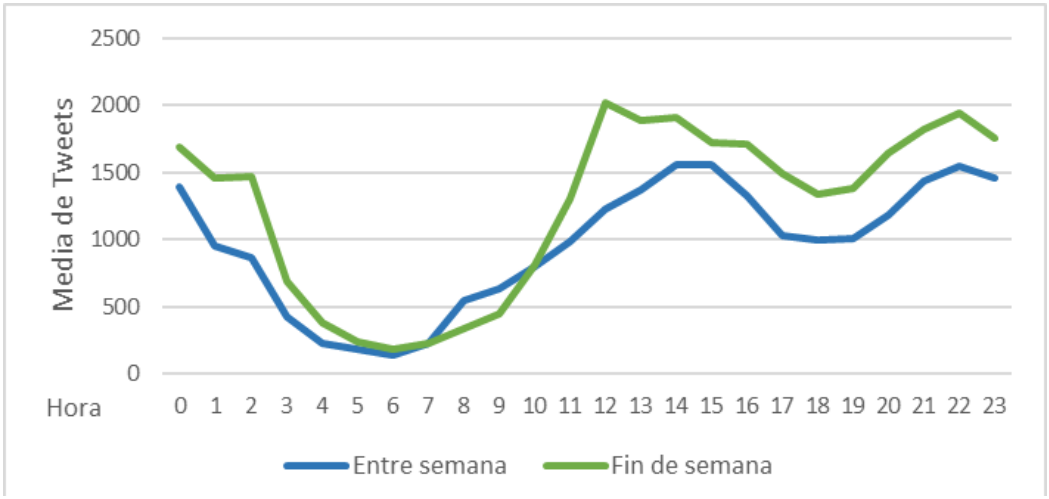


Fuente: elaboración propia a partir de datos de la API Twitter (2015–2016)

La actividad del sábado y domingo es inferior a la de los días de entre semana hasta las 10 horas, momento a partir del cual se invierte la tendencia y los valores del fin de semana se mantienen superiores, siguiendo una tendencia similar entre horas. El cese durante la noche se produce más lentamente, manteniendo una actividad considerable hasta las 3 de la madrugada.

Finalmente, los viernes se observa un patrón ligeramente distinto, alcanzando un máximo a las 15 horas y manteniendo una actividad más alta y más constante que entre semana. Por sus características, se considerará periodo entre semana hasta las 18 horas, y de fin de semana desde entonces. Resultado de esto se obtiene la Figura 6.

Figura 6. Media tweets por hora y grupo de días



Fuente: elaboración propia a partir de datos de la API Twitter (2015–2016)

2.3 Propuesta metodológica para la detección de residencias

Una vez presentadas las características de la fuente y de los datos obtenidos, se expone la propuesta metodológica de análisis, que en este caso se centra en la detección de las residencias de los usuarios, primera fase y dato fundamental para el cálculo de los indicadores de movilidad. Siguiendo a Jurdak et al (2015), el lugar de residencia del usuario se identifica con el lugar desde el que más frecuentemente se twittea, lo que se denomina "localización dominante". Según la clasificación que realizan Bojic et al (2015), existen cinco métodos usualmente utilizados para la detección de residencias:

1. A partir del lugar en donde se detecta un máximo número de registros
2. Empleando el lugar ocupado durante un mayor número de días activos, considerando como día activo aquél en el que se detecta al menos un registro
3. Calculando el lugar en donde transcurre el máximo periodo de tiempo entre el primer y el último registro
4. Añadiendo al lugar con el máximo número de registros la variable temporal de 7 pm a 7 am, cuando se supone que los usuarios se encuentran en casa
5. Añadiendo al lugar con un mayor número de días activos la variable temporal de 7 pm a 7 am

En el caso aquí planteado se ha decidido emplear únicamente la variable espacial, con la excepción de las restricciones por comportamiento (explicada en el apartado 3.3.1.b) que añaden en cierto modo la variable temporal. No se utilizará para la detección de residencias una restricción por horas durante la noche, como en los métodos 4 y 5, para evitar influir en el análisis horario, y de esta forma observar sin condicionantes el uso de la plataforma en relación a la localización de los usuarios.

Como se ha comentado, las nuevas fuentes de datos son una oportunidad magnífica para mejorar en el conocimiento de los comportamientos humanos, pero sus características requieren un proceso de filtrado y validación de resultados que permitan dotar al estudio de la consistencia necesaria. El proceso consta de las siguientes etapas:

1. Filtrado de los datos para obtener información de movilidad
    a. Filtrado de usuarios por fuente
    b. Filtrado de tweets por comportamiento
    c. Filtrado de usuarios según movilidad reflejada
2. Detección del lugar de residencia
3. Validación: análisis cuantitativo (padrón continuo) y cualitativo (encuesta interna)

3.3.1 Filtrado de los datos para obtener patrones de movilidad

a) Filtrado por tipo de usuario

Twitter es una plataforma que se utiliza para múltiples propósitos, entre ellos: ofrecer información del estado del tráfico, meteorología, tendencias, foros, etc. Las posiciones de este tipo de datos, la mayoría provenientes de sensores, no encierran información sobre la movilidad de los ciudadanos, por lo que deben ser eliminadas. Generalmente estos sensores tienen asociada una "fuente del tweet" propia, es decir, una plataforma desde la que se lanza el mensaje, por lo que para su detección se calcula la ratio de tweets y de usuarios de cada fuente. Si una misma fuente tiene un número bajo de usuarios que generan gran cantidad de tweets, son considerados sensores. Esta información es objeto del primer filtrado, cuyo resultado se recoge en la Tabla 2. Aquí se ve como solo 14 usuarios han lanzado 19. 456 tweets, eliminados del análisis al considerarse no válidos.

b) Filtrado por comportamiento

El comportamiento de un usuario dentro de la red social varía en función de ciertos eventos de su vida, que pueden corresponderse o no con eventos generales. En consecuencia, si en un momento dado un usuario crea muchos tweets desde una misma posición en respuesta a un acontecimiento determinado, este momento puntual puede afectar al cálculo de la posición dominante. Para eliminar este efecto, y siguiendo a Jurdak et al (2015), se filtraron en un principio los tweets generados en un intervalo menor a tres horas y en una distancia menor de 100 m., dejando sólo el primero de ellos. No obstante, los resultados obtenidos seguían estando afectados por este comportamiento, observando que hay usuarios que alargan estos momentos más de lo esperado. Por esto, se amplía el intervalo temporal hasta un día completo. El número de tweets se reduce en un 35 %, mientras el número de usuarios, como era de esperar, se mantiene (ver Tabla 2).

c) Filtrado por movilidad reflejada

Este es un paso necesario en el caso específico del análisis de movilidad. En él se excluyen los usuarios con movilidad poco representativa, es decir, que no aportan información relevante. Para ello se consideran, por un lado, la cantidad de ubicaciones geográficas diferentes que ocupa cada usuario, y, por otro, el periodo de tiempo durante el que han hecho uso de la plataforma dentro del área de estudio. El primer criterio permite eliminar los usuarios que muestran escasa movilidad, incluyendo negocios o instituciones, ya que estos lugares suelen tener su propia dirección en el listado y se localizan siempre en el mismo lugar. Por otra parte, el periodo de tiempo durante el que han hecho uso de la plataforma dentro del área de estudio permite diferenciar entre población residente y turistas. Dado que interesa conocer la movilidad de la población residente, se eliminarán los usuarios que puedan corresponder a turistas utilizando las fechas de sus tweets. En la Tabla 2 se recogen los resultados de los dos filtrados anteriores.

Tabla 2. Resumen del proceso de filtrado

Filtrado	Tweets retenidos tras filtrado	Usuarios retenidos tras filtrado	Tweets sobre muestra inicial (%)	Usuarios sobre muestra inicial (%)
0. Ninguno	177 675	22 641	100,00	100,00
1. No humanos por fuente	158 219	22 627	89,05	99,94
2. Comportamiento	103 518	22 627	58,26	99,94
3. Sin movilidad o poco representativa	56 378	3007	31,73	13,28

Fuente: elaboración propia a partir de datos de la API Twitter (2015–2016)

La Tabla 3 muestra las diferentes opciones posibles de este tercer filtrado, con el fin de observar cómo responden los datos a las restricciones de posiciones y tiempo aplicadas.

Tabla 3. Proceso de filtrado por movilidad

Filtrado por movilidad	Tweets	Usuarios	Tweets sobre muestra inicial (%)	Usuarios sobre muestra inicial (%)
3.1. Sólo 1 posición	13 105	11 423	7,38	50,45
3.2. Entre 2 y 4 posiciones	21 824	6905	12,28	30,50
3.3. Más de 5 posiciones	68 589	4299	38,60	18,99
3.3.1. Diferencia >= 15 días	66 722	3882	37,55	17,15
3.3.2. Diferencia >= 60 días	56 378	3007	31,73	13,28

Fuente: elaboración propia a partir de datos de la API Twitter (2015–2016)

El desglose de los resultados para realizar el tercer filtrado muestra características muy interesantes de la fuente, como, por ejemplo, que los usuarios que sólo tienen una localización registrada suponen el 50,45 % del total (fila 3.1). Aquí se agrupan principalmente los usuarios esporádicos de la plataforma y aquellos que twittean desde un negocio o institución. En el otro extremo, se considera que un usuario que revela un mínimo de 5 ubicaciones diferentes es un usuario avanzado que ofrece un buen seguimiento de sus rutinas y sobre el que se puede localizar con mayor fiabilidad su lugar de residencia, por lo que éste es el criterio que se adopta como primera restricción. A su vez, dentro de los usuarios que cumplen esta condición, y con el fin de eliminar aquéllos con una estancia puntual en el área, se aplica la restricción de tener tweets separados en el tiempo por un mínimo de 60 días. Por tanto, al final del proceso se cuenta con 3007 usuarios adecuados para calcular su lugar de residencia.

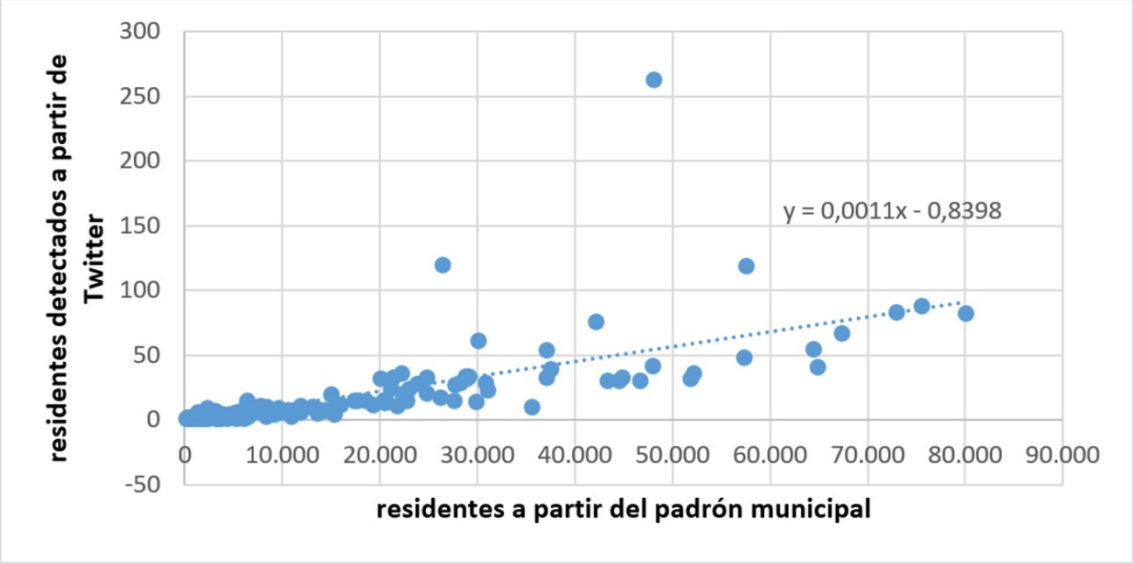
3.3.2 Detección de residencias y validación de resultados

Como ya se ha mencionado, para conocer el lugar de residencia de cada usuario se parte del supuesto de que éste coincide con el lugar desde el que se más se twittea, por lo que el siguiente

paso es calcular la localización dominante de cada uno de ellos. Este cálculo se hace empleando el radio de 100 metros propuesto por Jurdak el al. (2015) y ya empleado anteriormente para el filtrado de tweets compulsivos. Se selecciona la posición más repetida por usuario. Existe el caso de usuarios que, cumpliendo todos los requisitos de filtrados anteriores, no repiten nunca ubicación, por lo que no es posible obtener una posición dominante. En el caso práctico tratado son 362 usuarios de los 3007 totales quienes no permiten hacer esta asignación. Se retienen pues un total de 2645 usuarios, denominados "trazables", a los que se les ha asignado un lugar de residencia.

Para validar esta información se realiza un test de correlación de Pearson que compara la cantidad de usuarios localizados en cada municipio a través de Twitter y el número de habitantes según los datos del padrón municipal 2015. La hipótesis que se plantea es que, la distribución espacial de los lugares de residencia de los usuarios se ajustará, en términos generales, a la distribución espacial de la población en el área de estudio. Se utiliza la escala municipal salvo en el caso del municipio de Valencia, que, debido a su tamaño demográfico, se divide en distritos. El resultado inicial ofrece un coeficiente de correlación R= 0.72, lo cual quiere decir que el padrón y los residentes detectados por Twitter se encuentran bien correlacionados. No obstante, el análisis del diagrama de dispersión (Figura 7) muestra tres puntos que no siguen el patrón marcado por los demás. Estos tres puntos son los distritos centrales de Ciutat Vella y Extramurs y el distrito de Poblats Marítims, donde se ubica el puerto de la ciudad.

Figura 7. Resultado primera detección de residencias
y comparación con el padrón 2015

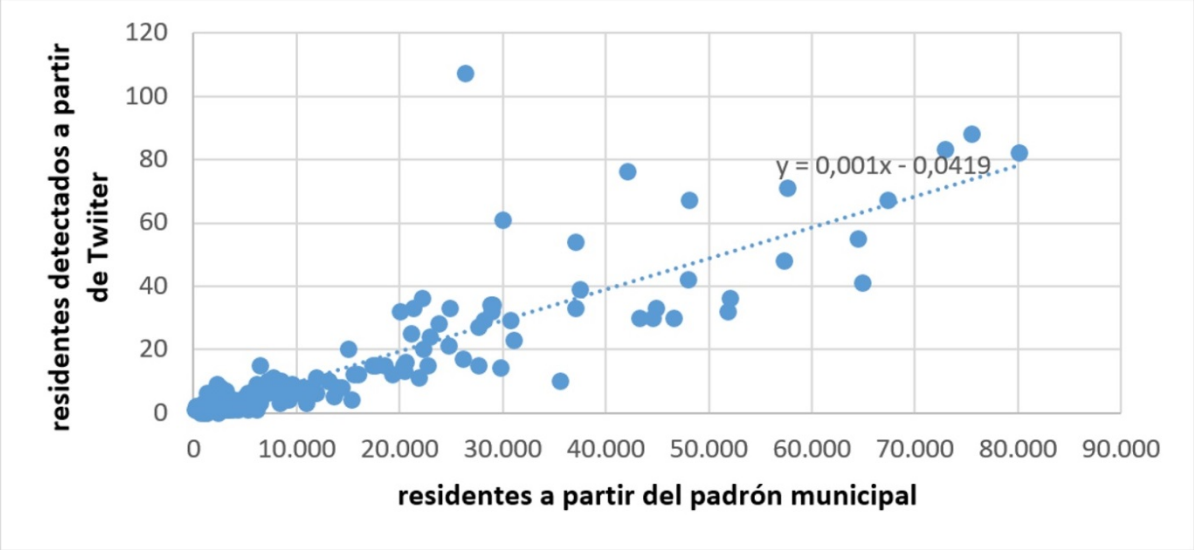


Fuente: elaboración propia a partir de datos
del padrón continuo y análisis Twitter (2015–2016)

La observación de detalle de estos tres distritos atípicos revela el impacto de las localizaciones no significativas que se describían en el apartado del análisis exploratorio de la fuente. En el caso de los distritos del centro, se detecta que Instagram, Twitter y Foursquare colocan aquí a los usuarios que marcan "Valencia" en su listado de lugares, una localización general que no tiene la precisión suficiente para hacer una detección del lugar de residencia o permitir un análisis de movilidad, lo cual induce a error. En el caso de la zona del puerto, la explicación parece estar en el fuerte impacto que tiene la localización de lugares de ocio nocturno y la realización de eventos en la emisión de tweets.

Por tanto, estos resultados muestran la necesidad de filtrar las posiciones sin precisión espacial suficiente para un análisis de movilidad metropolitano o provincial. La eliminación de estos puntos permite mejorar el resultado en el Test de Pearson hasta un coeficiente de correlación R= 0,88 (ver Figura 8).

Figura 8. Resultado detección de residencias
y comparación con el padrón 2015

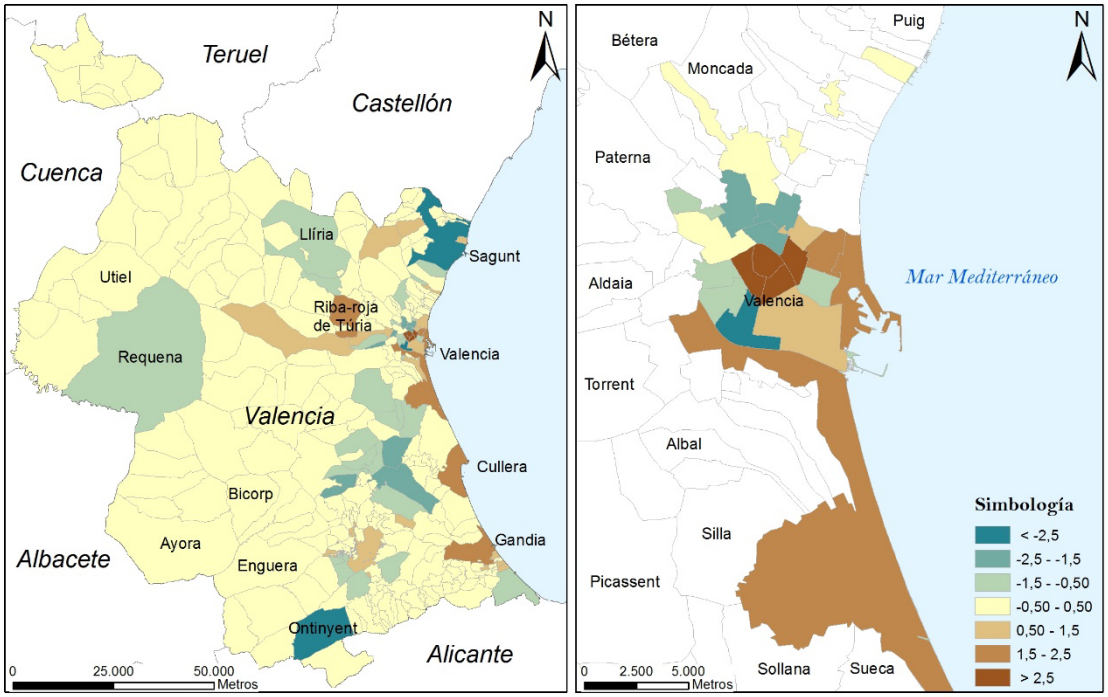


Fuente: elaboración propia a partir de datos
del padrón continuo y análisis Twitter (2015–2016)

Se puede pues concluir que la tendencia de los usuarios a usar Twitter desde su lugar de residencia es alta, aunque no todos siguen este patrón, y en algunos casos la posición dominante se detecta en lugares que se corresponden con espacios de ocio. Los efectos se observan en la Figura 9, donde los mayores valores de residuales positivos (mayor número de tweets emitidos de lo que correspondería por su número de habitantes) corresponden a los espacios urbanos con importantes actividades de ocio, tales como Gandia o Cullera. Dentro del municipio de Valencia, es el distrito de Ciutat Vella, que aloja numerosos espacios de ocio, el que presenta el valor más elevado. Y

esto pese a que, tal y como se explica en el punto 3.3.1, apartado c), se han eliminado los usuarios que realizan tweets esporádicos, indicativos de una estancia puntual en el área, es decir, a los turistas. Por tanto, lo más probable es que se trate de población residente en el área que frecuenta estos espacios de forma regular por motivos de ocio. El aumento del tamaño de la muestra debe desembocar en la disminución de este impacto.

Figura 9. Residuales estandarizados de la regresión
entre número de habitantes en 2015 y número de residencias detectadas



Fuente: elaboración propia a partir de datos
del padrón continuo y análisis Twitter (2015–2016)

La última fase de validación está basada en un análisis cualitativo, a partir del contenido del mensaje. Para ello se realiza un muestreo aleatorio de 100 usuarios, cuyos tweets se revisan para comprobar si su lugar de residencia coincide con el lugar identificado como tal. Para ello, se toman como referencia los mensajes que, de forma específica, mencionan encontrarse en su casa, y se revisan los demás tweets emitidos para reconocer las temáticas más comunes asociadas al lugar de residencia. Las más frecuentes han resultado ser: Proceso de dormir/despertarse/desayunar, celebraciones familiares, hacer deberes/estudiar, opiniones sobre política/programas de televisión. Se procede a leer los tweets de los usuarios seleccionados para buscar las temáticas señaladas.

Los resultados clasifican las residencias según hayan sido bien detectadas a nivel de punto, bien detectadas a nivel municipal, residencias no seguras, y residencias mal detectadas.

- Las "residencias bien detectadas a nivel de punto" son aquellas que tienen buena precisión en su localización, es decir, que se considera que la residencia real no se encuentra a más de 100 m. del lugar identificado. En la muestra, un 41 % de las residencias han sido correctamente detectadas a nivel de punto.

- Las "residencias bien detectadas a nivel municipal" son aquéllas en las que se observa que el usuario reside en el municipio, pero los textos de sus tweets no hacen mención específica a las actividades señaladas en el caso anterior. Se clasifica de esta forma cuando: el municipio coincide con la localización por defecto que existe en sus tweets, cuando se hace referencia a actividades de proximidad (colegio, gimnasio), o cuando la mayoría de los tweets se localizan en el municipio. El 28% de las residencias entran dentro de esta categoría.

- Las "residencias no seguras" son aquellos puntos para los que el contenido de sus mensajes no permite reconocer si se trata o no del lugar de residencia del usuario. Un 20 % del total de las residencias detectadas entran en esta categoría.

- - Finalmente, se considera una "residencia mal detectada" cuando los textos de los tweets revelan que su residencia se encuentra en otro lugar, o que hacen referencia específica a encontrarse en el centro de trabajo o en un lugar de ocio. El 11% de las residencias detectadas se clasifican aquí.

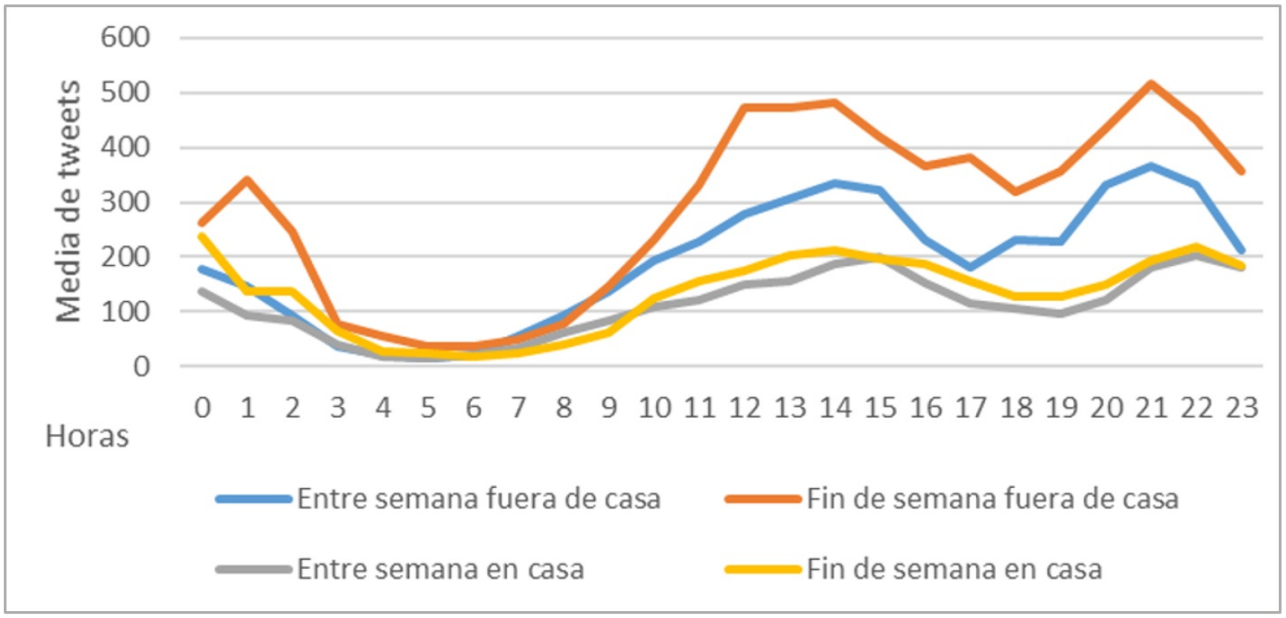
En conclusión, los resultados obtenidos son buenos, ya que el 69 % de las residencias han sido correctamente detectadas, al menos a nivel municipal. Se puede afirmar que el proceso de validación ha resultado exitoso y que la metodología empleada para la detección de residencias de usuarios Twitter es oportuna.

3.3.3 Análisis de resultados

La detección de las residencias, una vez validada, permite realizar un análisis más preciso de las pautas de comportamiento espacio-temporal de la población analizada.

En primer lugar, es posible identificar las diferencias de comportamiento existentes en el uso de Twitter en función de si los usuarios se encuentran dentro o fuera de casa. La Figura 10 hace un seguimiento por horas del número de mensajes enviados diferenciando entre el lugar de emisión (residencia o fuera de la residencia) y periodo semanal (entre semana y fin de semana, de acuerdo con la distinción establecida en el punto 3.2).

Figura 10. Media Tweets diarios de residentes por lugar y periodo semanal



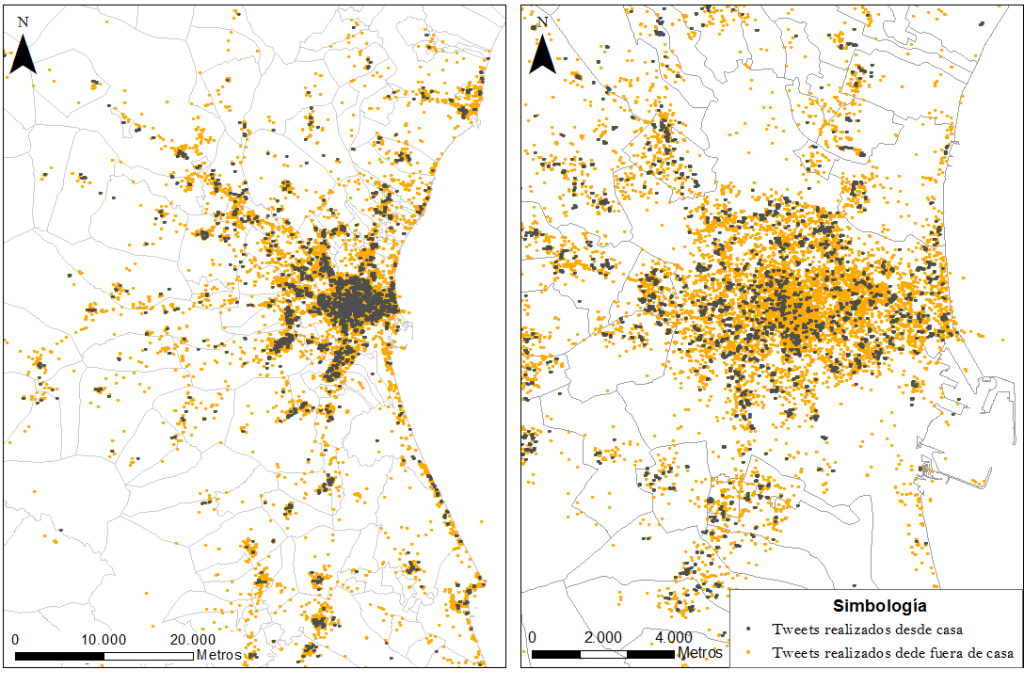
Fuente: elaboración propia

Si se compara esta gráfica con las presentadas en el apartado 3.2, se observa la disminución en el número de tweets, ya que ahora sólo se analizan los de los usuarios sobre los que se ha localizado su lugar de residencia. Esto no supone una mengua en la información, sino un aumento, ya que al añadir el lugar de residencia es posible caracterizar el empleo de la plataforma en distintos espacios. Aunque los patrones horarios observados en gráficas anteriores se repiten, existen diferencias interesantes al incluir la residencia. En primer lugar, la plataforma se utiliza a horas muy regulares desde casa, independientemente del día de la semana. En segundo lugar, los tweets generados fuera de la residencia son siempre superiores a los realizados desde la misma y, además, la frecuencia durante el fin de semana es mucho más alta que la de entre semana.

Los resultados se relacionan con el comportamiento que se espera de una red social, donde la mayoría de tweets geolocalizados se crean durante actividades de carácter social, por lo que tendrá un registro amplio de actividades de ocio fuera de casa. Las actividades más susceptibles de crear un tweet geolocalizado son aquéllas diferentes a las habituales, para las que, además, el usuario suele querer mostrar una imagen, como se ha podido constatar en la cantidad de registros provenientes de la plataforma Instagram. Esto no supone un problema para detectar actividades cuotidianas, ya que el patrón espacial es distinto. Un lugar frecuentado habitualmente se repite en el espacio en varias ocasiones, aunque tenga menos posibilidades de generar un tweet cada vez que se acude. Un lugar poco frecuentado tiene más posibilidades de generar un tweet, pero no de repetirse espacialmente. Los lugares identificados de esta forma pueden ser vinculados a actividades de ocio, mientras que los que sí se repiten en el espacio fuera de casa serían considerados actividades cuotidianas, como por ejemplo el lugar de trabajo.

Finalmente, el mapa de la Figura 11 muestra los diferentes patrones espaciales de los tweets realizados desde el lugar de residencia y los emitidos desde fuera de ella. En este segundo caso la dispersión territorial es mayor, ya que el rango de actividades susceptibles de ser objeto de un tweet es mayor fuera de la residencia.

Figura 11. Tweets realizados dentro y fuera del lugar de residencia



Fuente: elaboración propia

4 Conclusiones

La investigación presentada se ha centrado en el empleo de datos geolocalizados de la red social Twitter para el estudio de las pautas de movilidad diarias, aplicado al caso del área urbana de Valencia.

El uso de esta red social aporta información muy significativa al conocimiento de los patrones de movilidad diaria, estudiados hasta la fecha empleando fuentes estadísticas tradicionales como encuestas, flujos de tráfico, estadísticas de transporte de pasajeros o censos. Twitter se revela como una fuente que posibilita seguimientos individualizados de sus usuarios, información sólo comparable a la realización de costosas encuestas. Si a esto se le añade el número de usuarios y la precisión espacial de la muestra, proporciona información no existente hasta la fecha. También engloba todos los tipos de movilidad, ya que se evidencian desplazamientos de personas sin determinar distancia, modo o finalidad. Otra característica a tener en cuenta es la constante actualización de los datos, ya que se descargan en tiempo real, por lo que la información puede ser objeto de estudios a largo plazo o detectar pequeñas variaciones diarias.

La importancia de la detección de residencias empleando la información espacio-temporal que ofrecen los tweets tiene un valor significativo para el estudio de las dinámicas de ciudad. En un primer lugar, permite conocer los desplazamientos, no sólo en distancia, sino también en la dirección en que se producen. Al inferir los distintos motivos de desplazamiento será posible conocer el uso de los distintos espacios de la ciudad de una forma cambiante en el tiempo. Además, se pueden añadir a cada residencia detectada variables socioeconómicas del lugar, como tipo de construcción, precio medio del suelo, edad media de los individuos…, de forma que se puedan asociar a un perfil de ciudadano cierto tipo de desplazamientos.

Por tanto, se considera que la realización de estudios geográficos a partir de datos masivos provenientes de la tecnología es una cuestión innovadora que, tomando como base las ventajas significativas que ofrece, requiere en la actualidad de un enfoque reflexivo y un análisis de conjunto. Así es como se ha desarrollado el trabajo que se presenta, que ha producido varias aportaciones.

La primera de ellas ha sido la confirmación de que la red social Twitter es adecuada para revelar patrones de movilidad de un área gracias a los tweets geolocalizados. Ahondar en la precisión espacial real de un tweet ha sido uno de los aspectos que más información ha revelado. La voluntad de cada usuario, que maneja el tipo de localización a incluir en su mensaje, es el factor clave que marca la precisión real. Esta información constituye la base para entender muchos de los resultados posteriores, y aplicar las correcciones y filtrados necesarios para mejorar la significación de los datos.

La segunda aportación es el diseño de una metodología que ha permitido convertir datos brutos de Twitter en información sobre la residencia de sus usuarios. El análisis de la fuente ha sido clave para comprender los problemas que afectaban al modelo, establecer requisitos de filtrado no incluidos inicialmente y, finalmente, crear la propuesta metodológica adecuada. Dicha adecuación se concluye tras los distintos procesos de validación de resultados, compuestos de un análisis cuantitativo, empleando el Padrón continuo de población, y otro cualitativo, basado en una lectura del contenido de los tweets.

La tercera aportación sustancial del trabajo se deriva del propio estudio de los resultados. Se observa cómo la diferenciación de los mensajes en función del lugar de emisión del tweet ofrece información sobre los patrones espaciales de uso de la plataforma. Aquí se revela que su utilización fuera de casa es muy superior a la de dentro, lo que implica que la fuente informa sobre una gran muestra de actividades de ocio. La clave para efectuar un estudio de movilidad se encuentra en tener en cuenta que el patrón espacial es distinto según el tipo de actividad. Una actividad de ocio tendrá muchas posibilidades de generar un tweet geolocalizado, pero pocas de repetirse, mientras que un tweet geolocalizado en un lugar frecuente tiene menos posibilidades de producirse, por lo habitual de la situación, pero muchas de repetirse en el tiempo. Este hecho se observa también al

analizar los tweets emitidos desde casa, que apenas se diferencian cuantitativamente entre semana y en fin de semana. Por el contrario, los registros fuera de casa en fin de semana son muy superiores a los existentes entre semana, situación obvia si se asocian con actividades de ocio. Finalmente, no se ha encontrado un momento temporal claro que pueda asociarse a la residencia o a estar fuera de ella. Este hecho puede deberse a varios motivos, cuya clarificación requiere de un análisis posterior a partir de una muestra de mayor tamaño que permita identificar un mayor número de actividades, incluyendo información sobre el centro de trabajo.

En conclusión, consideramos que los resultados obtenidos son positivos, y constituyen un primer avance en la aplicación de este tipo de datos al estudio de la movilidad en áreas urbanas. A partir de la detección de las residencias del usuario es posible parametrizar sus movimientos, lo que permitirá conocer sus pautas de movilidad habituales. La combinación de esta información con variables socioeconómicas y territoriales permitirá adquirir nueva información tanto sobre las pautas espaciotemporales de movilidad de los residentes urbanos como sobre la ocupación temporal del espacio urbano, línea de trabajo que estamos actualmente desarrollando.

Agradecimientos: Este artículo se ha elaborado en el marco del proyecto "Sostenibilidad social, conectividad global y economía creativa como estrategias de desarrollo en el Área metropolitana de Valencia" (CSO2016-74888-C4-1-R), financiado por la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER) dentro del Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad, incluido en el Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016, convocatoria de 2016. Carmen Zornoza cuenta con una ayuda para contratos predoctorales para la formación de doctores (BES-2014-067846) dentro del Subprograma Estatal de Formación del Programa Estatal de Promoción del Talento y su Empleabilidad, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016, financiado por el Ministerio de Economía y Competitividad y cofinanciado por el Fondo Social Europeo.

Declaración responsable: Las/os autoras/es declaran que no existe ningún conflicto de interés en relación a la publicación de este artículo. . Asimismo, las dos autoras declaran que han elaborado conjuntamente todos los apartados del artículo.

Bibliografía

Béjar, J., Álvarez, S., García, D., Gómez, I., Oliva, L., Tejeda, A., & Vázquez-Salceda, J. (2016). Discovery of spatio-temporal patterns from location-based social networks. *Journal of Experimental & Theoretical Artificial Intelligence, 28*(1–2), 313–329. https://doi.org/10.1080/0952813X.2015.1024492

Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S., & Ratti, C. (2015). Choosing the Right Home Location Definition Method for the Given Dataset (pp. 194–208). In T.Y. Liu, C. Scollon & W. Zhu (Eds.), *Social Informatics. SocInfo 2015. Lecture Notes in Computer Science*. Springer. https://doi.org/10.1007/978-3-319-27433-1_14

Frias-Martinez, V., & Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using Twitter activity. *Engineering Applications of Artificial Intelligence, 35,* 237–245. https://doi.org/10.1016/j.engappai.2014.06.019

Gabrielli, L., Rinzivillo, S., Ronzano, F., & Villatoro, D. (2014). From Tweets to Semantic Trajectories: Mining Anomalous Urban Mobility Patterns. In Jordi Nin & Daniel Villatoro (Eds.), *Citizen in Sensor Networks* (pp. 26–35). Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-04178-0_3

García-Palomares, J. C., Gutiérrez, J., & Mínguez, C. (2015). Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS. *Applied Geography, 63,* 408–417. https://doi.org/10.1016/j.apgeog.2015.08.002

Goodchild, M. (2007). Citiziens as sensors: the word of volunteered geography. *GeoJournal, 69,* 211–221. https://doi.org/10.1007/s10708-007-9111-y

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779–782. https://doi.org/10.1038/nature06958

Gutiérrez-Puebla, J., García-Palomares, J. C., & Salas-Olmedo, M. H. (2016). Big (Geo)Data in Social Sciences:Challenges and Opportunities. *Revista de estudios andaluces, 33,* 1–23. http://dx.doi.org/10.12795/rea.2016.i33.01

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013, August). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing* (pp. 1–8). Chicago. https://doi.org/10.1145/2505821.2505823

Huang, W., Li, S., Liu, X., & Ban, Y. (2015). Predicting human mobility with activity changes. *International Journal of Geographical Information Science, 29*(9), 1569–1587. https://doi.org/10.1080/13658816.2015.1033421

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., & Newth, D. (2015). Understanding human mobility from Twitter. *PloS one, 10(7)*, e0131469. https://doi.org/10.1371/journal.pone.0131469

Kwan, M. P. (1999). Gender, the home-work link, and space-time patterns of nonemployment activities. *Economic geography, 75*(4), 370–394. https://doi.org/10.1111/j.1944-8287.1999.tb00126.x

Masquenegocio. (2016, January). *Twitter users in Spain* [PDF report]. Retrieved from http://www.masquenegocio.com/wp-content/uploads/2016/01/Twitter-en-Espan%CC%83a.pdf

Song, C., Qu, Z., Blumm, N., & Barabási, A. L. (2010). Limits of predictability in human mobility. *Science, 327*(5968), 1018–1021. https://doi.org/10.1126/science.1177170

Li, S., Dragicevic, S., Castro, F., Sesterd, M., Wintere,S., Coltekin, A., … Chengi, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *Isprs journal of photogrammetry and remote sensing, 115*, 119–133. https://doi.org/10.1016/j.isprsjprs.2015.10.012

Llorente, A., Garcia-Herranz, M., Cebrian, M., & Moro, E. (2015). Social media fingerprints of unemployment. *PloS one, 10*(5), e0128692. https://doi.org/10.1371/journal.pone.0128692

Serrano Estrada, L., Serrano Salazar, S., & Álvarez Álvarez, F. J. (2014). Las redes sociales y los SIG como herramientas para conocer las preferencias sociales en las ciudades turísticas: el caso de Benidorm. Presented at the XVI Congreso Nacional de Tecnologías de Información Geográfica (pp. 1005–1012). Alicante, Spain, June 25–27. Madrid: Asociación de Geógrafos Españoles.