Geografía: Cambios, Retos y Adaptación
AGE y Universidad de La Rioja
pp. 917-923 (2023)

XXVIII Congreso de la Asociación Española de Geografía
ISBN 978-84-09-53925-3
DOI: 10.21138/CG/2023.lc

VINCULANDO LAS PERCEPCIONES DEL USO DEL METRO DE MADRID AL ESPACIO A PARTIR DE DATOS DE TWITTER

JOAQUÍN OSORIO ARJONA (id)[1]

[1]*Departamento de Geografía Humana, Facultad de Geografía e Historia Universidad de Sevilla, C. Doña María de Padilla, 41004 Sevilla*

Autor de correspondencia: josorio@us.es

Resumen. Las redes sociales son plataformas muy utilizadas por los viajeros quienes expresan sus opiniones sobre servicios como el transporte público. Esta comunicación presenta el valor de los textos de las redes sociales como fuente de datos para detectar la distribución espacial de los problemas dentro de una red de transporte público mediante la geolocalización de los sentimientos de los ciudadanos, y analiza los efectos que algunos factores como la población o los ingresos tienen sobre esa distribución espacial, con el objetivo de desarrollar un servicio de transporte público más inteligente y sostenible. Para ello se recogen datos de Twitter de la cuenta de Metro de Madrid durante un periodo de dos meses. A continuación, se utiliza un modelo de regresión geográficamente ponderada para explorar la causalidad de la distribución espacial de los usuarios que emiten quejas, utilizando fuentes de datos oficiales como variables exploratorias. Los resultados muestran que los usuarios de Twitter tienden a ser trabajadores de ingresos medios que residen en áreas periféricas y principalmente tuitean cuando viajan a sus lugares de trabajo. Los principales problemas detectados fueron la puntualidad y las averías en estaciones de transferencia o en zonas céntricas, principalmente en la mañana durante los días laborables.

Palabras clave: transporte público, Twitter, geolocalización, minería de texto, regresión geográficamente ponderada

LINKING PERCEPTIONS OF THE USE OF THE MADRID METRO TO SPACE FROM TWITTER DATA

Abstract. Social networks are platforms widely used by travellers who express their opinions about services such as public transport. This communication presents the value of texts from social networks as a source of data to detect the spatial distribution of problems within a public transport network through the geolocation of citizens' feelings and analyses the effects that some factors such as population or income have on that spatial distribution, with the aim of developing a more intelligent and sustainable public transport service. To do this, Twitter data is collected from the Madrid Metro account for a period of two months. Then, a geographically weighted regression model is used to explore the causality of the spatial distribution of users who issue complaints, using official data sources as exploratory variables. The results show that Twitter users tend to be middle-income workers residing in peripheral areas and mainly tweet when commuting to their workplaces. The main problems detected were punctuality and breakdowns at transfer stations or in downtown areas, mainly in the morning on weekdays.

Keywords: public transport, Twitter, geolocation, text mining, geographically weighted regression

1. INTRODUCCIÓN

El transporte público representa el principal modo de la movilidad urbana sostenible (Chen *et al.,* 2018). Sin embargo, el crecimiento de las áreas metropolitanas implica un aumento de la demanda de movilidad, lo que significa un aumento en el número de viajes, una intensificación de los viajes motorizados, y unas rutas más largas y que consumen más tiempo (Banister, 2011). Este aumento de la demanda ha provocado una congestión en los sistemas de transporte público. Ante esta situación, las agencias de transporte público necesitan tener información actualizada para detectar problemas en sus servicios (Ji *et al.,* 2018).

Las opiniones de los ciudadanos son fundamentales para entender las necesidades, motivaciones y sensibilidades del uso del transporte público, proporcionando información útil para el planteamiento de estos servicios (El-Diraby *et al.,* 2019). Sin embargo, los datos de las fuentes tradicionales parecen insuficientes debido a su alto coste, baja frecuencia de actualización y baja resolución espacial y temporal (Gutierrez-Puebla y García-Palomares, 2016; Miralles-Guasch y Martínez, 2013).

La gran variedad, velocidad y volumen de las nuevas fuentes de datos basadas en las Tecnologías de la Información y la Comunicación son valiosas para estudios de movilidad y usos del suelo, permitiendo realizar análisis en patrones espaciotemporales que no pueden ser realizados por medios tradicionales (Gutiérrez-Puebla y García-Palomares, 2016). Las agencias de transporte público han adoptado enfoques para comunicarse con los usuarios de internet, proporcionando información sobre sus servicios (Manetti *et al.,* 2017). Los mensajes y opiniones compartidos en las redes sociales pueden ser utilizados para detectar problemas en la red, pero también para conocer las opiniones de los usuarios sobre el servicio.

Los tweets son recursos de datos ricos para la extracción de opiniones y sentimientos (Kocich, 2017). Hay un amplio rango de investigaciones que emplean los textos de los tweets para obtener resultados valiosos en diversos campos de aplicación como el control del tráfico (Steiger *et al.,* 2015). Los tweets son datos con bajo coste de descarga, y los datos son producidos continuamente y casi en tiempo real, haciéndolos una alternativa interesante frente a fuentes de datos tradicionales como las encuestas de movilidad. Con los datos de Twitter es posible observar las necesidades específicas de un usuario sobre un tema, y obtener información útil sobre un sentimiento particular asociado a dichas necesidades (Collins, *et al.,* 2013).

El objetivo de esta comunicación es explorar las percepciones de los usuarios de Twitter cuando viajan en un sistema de transporte público como el Metro de Madrid, y mostrar la utilidad de Twitter para localizar problemas en el espacio que comprende una red de transporte. Los datos de Twitter presentan una alta resolución temporal que permite visualizar los resultados a lo largo del tiempo. Esta investigación busca también comprender los efectos de algunas variables (población, renta, densidad de puntos de interés o conexiones con otros servicios de transporte público) sobre la distribución espacial de las quejas de los usuarios de la red. Para ello, se usan los textos de los tweets para extraer palabras claves que identifiquen las estaciones de metro y se elabora un método de Regresión Geográficamente Ponderada (GWR) para explorar la causalidad de las variables que afectan espacialmente al número de usuarios con sentimientos negativos.

2. ÁREA DE ESTUDIO Y DATOS

El área de estudio incluye la red de Metro de Madrid. Esta red es el principal sistema de transporte público del Área Metropolitana de Madrid. Está compuesta por 12 líneas de tren y 242 estaciones, y conecta los 21 distritos de la capital además de alcanzar otros 12 municipios (Figura 1).

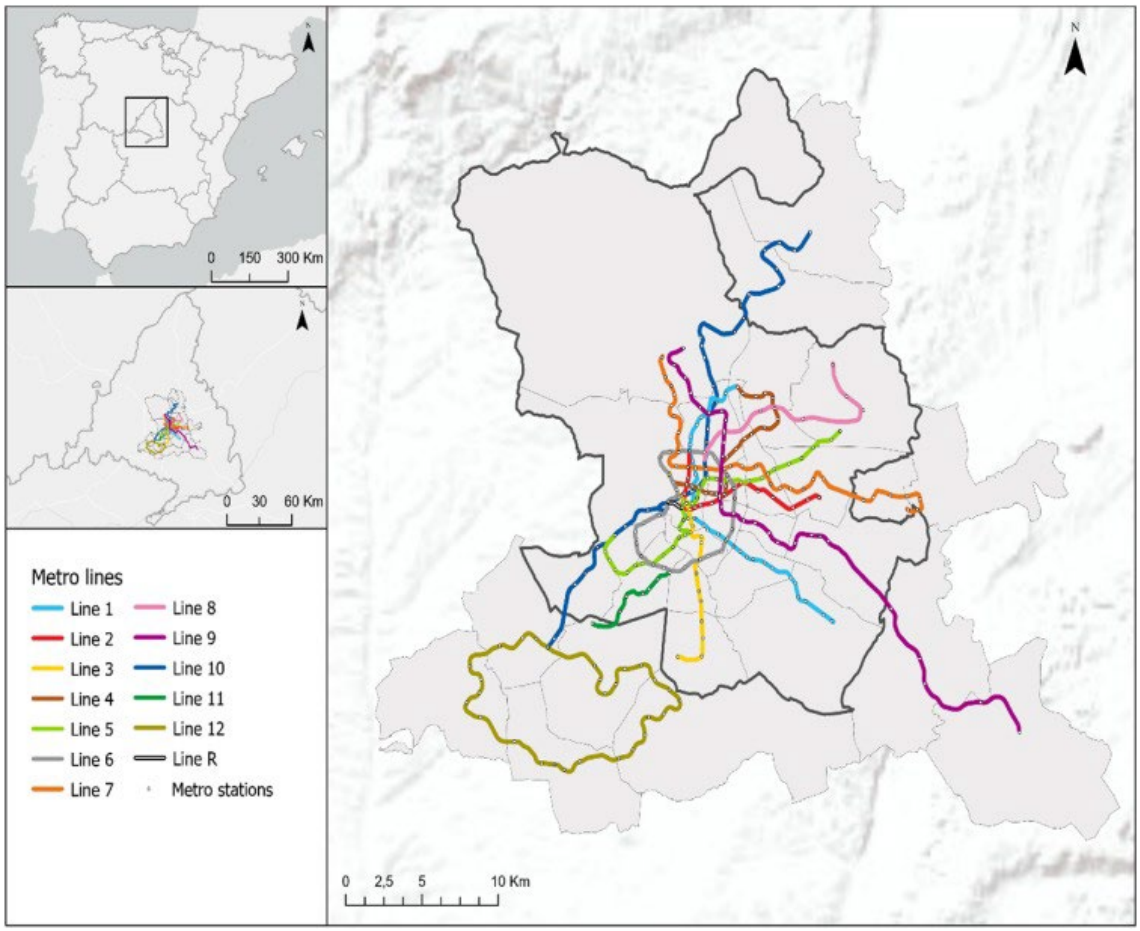
Los tweets utilizados son respuestas a la cuenta oficial del Metro de Madrid (*@metro_madrid*). Se han escogido estos tweets bajo la hipótesis de que los usuarios de Twitter tienden a contestar directamente a las cuentas de las agencias de transporte cuando quieren reportar un problema o expresar una queja acerca del servicio. Además, este tipo de textos tienden a dar información más detallada sobre el motivo de la queja o información relacionada con el servicio (Haghighi *et al.,* 2018).

La base inicial de tweets contiene 27.603 mensajes de 12.361 usuarios, recopilados a lo largo de un periodo de dos meses comprendido entre el 16 de septiembre y el 17 de noviembre del año 2019. Para identificar la estación desde donde se ha publicado un tweet, se ha usado un script de *Python* para geolocalizar el mensaje a partir de la identificación del nombre de la estación en el texto del tweet (Haghighi *et al.,* 2018).

Los datos espaciales de la red de metro están disponibles en la página web del Consorcio de Transportes de Madrid. Para el análisis GWR, los datos de la población residente han sido descargados del censo del año 2019 del Instituto Nacional de Estadística. Los datos de renta fueron suministrados por

el portal web del Ayuntamiento de Madrid a nivel de distritos, y por el Instituto de Estadística de la Comunidad de Madrid a nivel de municipios. El número de puntos de interés fue calculado a partir de los datos abiertos de *OpenStreetMap*.

Figura 1. Red de Metro de Madrid



Fuente: Elaboración propia.

3. METODOLOGÍA

Los textos de los tweets fueron procesados y limpiados a partir de la librería *Pandas* de *Python*. A continuación, se procedió a la geocodificación de los tweets empleando un diccionario de palabras claves con el nombre de todas las estaciones de metro. Con este método se logró geolocalizar 3.458 tweets de 2.418 usuarios (el 12,5% de la muestra inicial). Para los tweets geolocalizados en una estación ubicada en dos o más líneas de la red, se empleó un segundo diccionario de palabras claves para identificar una línea individual al tweet.

El siguiente paso fue desarrollar un análisis semántico para clasificar los textos en cuatro categorías: puntualidad, comodidad, averías y sobresaturación. Par ello se empleó un tercer diccionario con palabras claves asociadas a cada categoría (por ejemplo, ventilación, calor o sucio fueron utilizados como términos asociados a comodidad). Los tweets se clasificaron en estas cuatro categorías a partir del método de clusterización conocido como *Latent Dirichlet Allocation* (LDA) (Blei *et al.,* 2003).

Finalmente, un modelo GWR (Brundsdon *et al.,* 1996) fue efectuado para analizar las variables que potencialmente pueden afectar a la distribución espacial de las quejas de los usuarios de Twitter. El modelo GWR permite analizar variaciones en el espacio a partir de la obtención de coeficientes locales. Este análisis fue efectuado a una escala espacial que comprende los 21 distritos de Madrid y a los 12 municipios vecinos (32 unidades espaciales en total). Se utilizó como variable dependiente el número de usuarios de

Twitter de la muestra, y como unidades exploratorias la densidad de población en edad de trabajar, la renta media, la densidad de puntos de interés y el porcentaje de estaciones de transporte intermodal.

4. RESULTADOS

Las estaciones de metro más comentadas en la muestra corresponden con las estaciones más transitadas según datos oficiales. Estas estaciones están ubicadas en los distritos de la Almendra Central de Madrid o son estaciones de la línea 6 (línea circular que contiene varias estaciones de metro intermodales con otras líneas u otros servicios de transporte). Otras estaciones que destacar están situadas en zonas clave (como la estación ubicada en el aeropuerto de Barajas) o son estaciones periféricas que sirven de puntos de tránsito con el servicio en los municipios de la periferia (Figura 2).

La puntualidad y las averías son los problemas más reportados en la muestra. Mientras que la puntualidad es el principal problema reportado en el centro de Madrid, las averías son la queja más visible en la periferia. Los problemas de comodidad también son visualizados en el centro de la ciudad (Figura 3).

Analizando los resultados obtenidos en el modelo GWR, la variable de densidad de población influye positivamente en el número de usuarios que reportan problemas en el norte de la ciudad de Madrid (áreas de trabajo con baja densidad de población), mientras que tiene poco efecto sobre el sur del área metropolitana (áreas residenciales con alta densidad de población). Al mismo tiempo, hay poco impacto de esta variable en el centro de la ciudad, ya que allí los principales usuarios de la estación del metro son turistas. La variable de renta tiene un mayor poder explicativo en el sur del área metropolitana, que se corresponde con áreas residenciales habitadas por trabajadores de ingresos medios. Estos resultados también pueden demostrar que los ciudadanos de ingresos altos tienden a desplazarse al trabajo en coche. La variable de densidad de puntos de interés es la que más influye en el modelo. Sus resultados coinciden con las variables anteriores (los puntos de interés están relacionados con puntos ubicados en las áreas de destino de los viajes y los resultados obtenidos en la variable de densidad de población muestran que los usuarios de Twitter tienden a enviar mensajes mientras se desplazan al trabajo). La variable de puntos de interés tiene un alto poder explicativo en casi toda el área de estudio, especialmente en el centro de la ciudad, área con un mayor número de infraestructuras y servicios. Finalmente, los coeficientes de la variable de intermodalidad con otros servicios de transporte tienen una influencia alta en las zonas del norte de la ciudad. Esto puede interpretarse a que los usuarios de Twitter tienden a viajar en autobús o tren a los lugares de trabajo en el norte de Madrid, haciendo transbordo en una estación del centro o norte de la ciudad para completar su viaje (Figura 4).

5. CONCLUSIONES

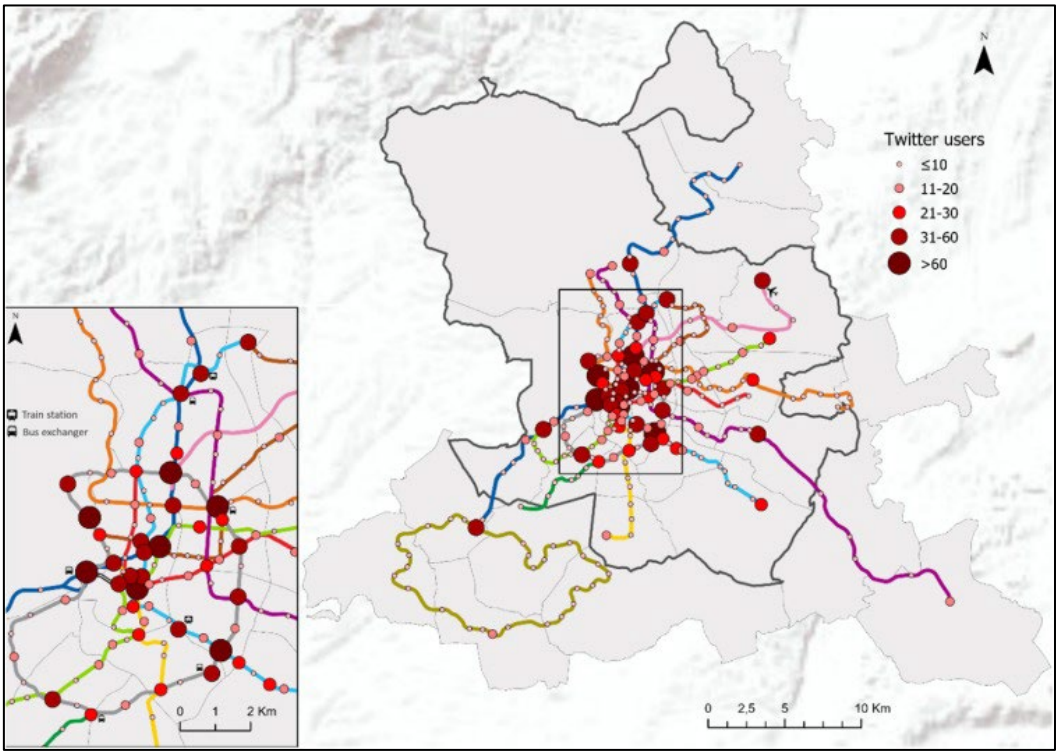
Esta comunicación ha usado datos de Twitter basados en respuestas a la cuenta oficial del Metro de Madrid para visualizar la distribución espacial de quejas en el espacio, los principales problemas de la red según la localización de las estaciones y el grado en el que determinadas variables afectan a esta distribución espacial mediante un modelo GWR.

Se ha observado que la variable principal que afecta a la distribución espacial de los tweets publicados en la muestra es la densidad de puntos de interés, asociada a servicios y equipamientos localizados en los lugares de destino. La puntualidad, el tema de queja mas frecuente, está relacionada con el nivel de renta en el sur del Área Metropolitana de Madrid, y con la densidad de estaciones intermodales en el norte de la ciudad. Estos resultados permiten visualizar el perfil de un usuario de Twitter que viaja en metro: trabajadores de renta media y que viven principalmente en los municipios del sur del área metropolitana, y que tienden a viajar a los lugares de trabajo del norte de Madrid, ya sea de forma directa o usando varios sistemas de transporte (en este último caso, hacen transbordo en las estaciones del centro de Madrid).

Hay que tener en cuenta que los datos de Twitter presentan una serie de limitaciones a tener en cuenta, como la falta de precisión espacial a la hora de utilizar el método utilizado en esta comunicación o la fiabilidad de las técnicas de minería de texto y detección de temas. Utilizar los textos de los tweets también conlleva posibles problemas de privacidad de datos, por lo que para minimizar este problema, los tweets de la muestra han sido agregados por estación de metro o municipio de localización.

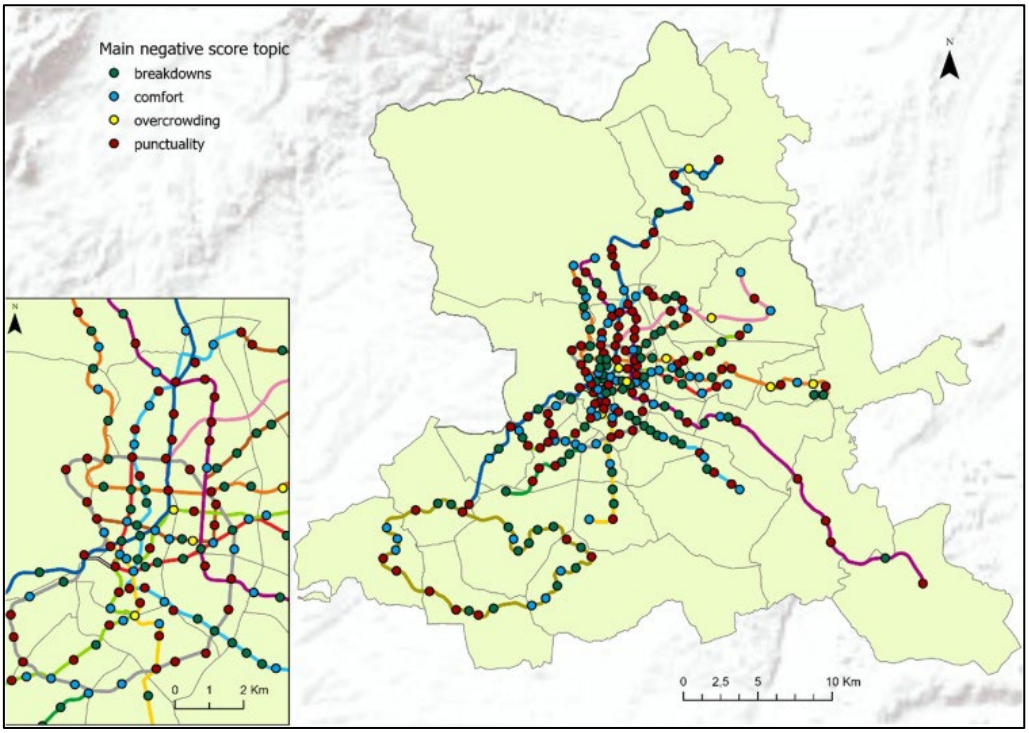
Agradecimientos: Este trabajo ha sido apoyado por la Comunidad de Madrid (SOCIALBIGDATA-CM, S2015/HUM-3427), el Ministerio de Educación, Ciencia y Universidades y el Fondo de Desarrollo Europeo Regional (DynMobility, RTI2018-098402-B-I00), y el Ministerio de Ciencia e Innovación (FJC2020-042912-I).

Figura 2. Distribución de usuarios de Twitter en la red de Metro de Madrid



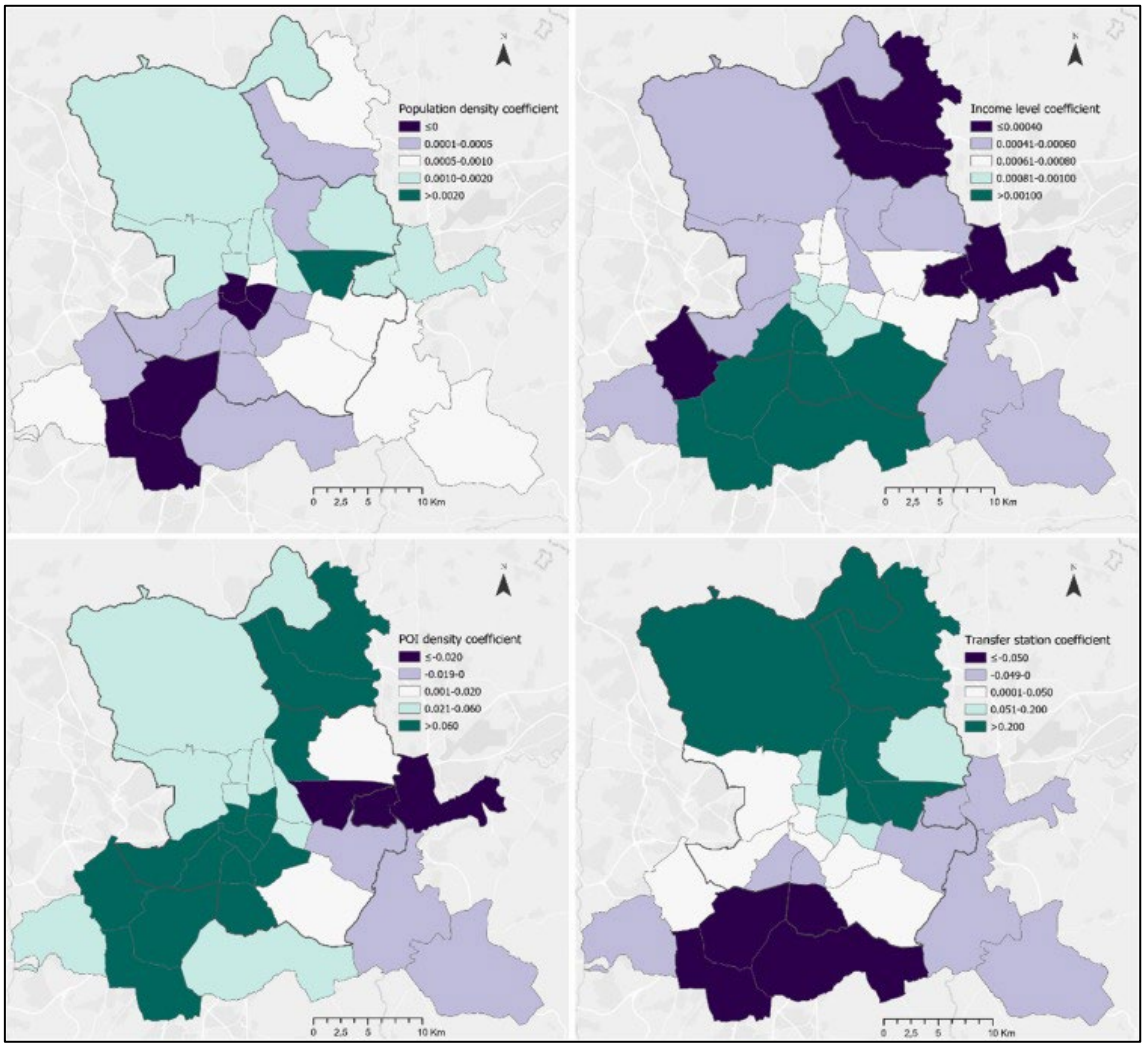
Fuente: Elaboración propia.

Figura 3. Problema principal detectado en las estaciones de la red de Metro de Madrid a partir de datos de Twitter



Fuente: Elaboración propia.

Figura 4. Coeficientes de distribución de las variables GWR en el Área Metropolitana de Madrid



Fuente: Elaboración propia.

REFERENCIAS

Banister, D. (2011). Cities, mobility and climate change. *Journal of Transport Geography*, 19(6), 1538–1546. https://doi.org/10.1016/j.jtrangeo.2011.03.009

Blei, D. M., Ng, A. Y., Edu, J. B. (2003). Latent dirichlet allocation Michael I. Jordan. *Journal of Machine Learning Research*, 3.

Brunsdon, C., Fotheringham, A. S., Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4), 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Chen, Y., Bouferguene, A., Li, H. X., Liu, H., Shen, Y., Al-Hussein, M. (2018). Spatial gaps in urban public transport supply and demand from the perspective of sustainability. *Journal of Cleaner Production*, 195, 1237–1248.

Collins, C., Hasan, S., Ukkusuri, S. V. (2013). A novel transit rider satisfaction metric a novel transit rider satisfaction metric: Rider sentiments measured from online social media data. J*ournal of Public Transportation*, 16(2), 21–45.

El-Diraby, T., Shalaby, A., Hosseini, M. (2019). Linking social, semantic and sentiment analyses to support modeling transit customers' satisfaction : Towards formal study of opinion dynamics. *Sustainable Cities and Society*, 49. https://doi.org/10.1016/j.scs.2019.101578

Gutiérrez-Puebla, J., García-Palomares, J. C. (2016). Big (Geo) Data en Ciencias Sociales: Retos y Oportunidades. *Revista de Estudios Andaluces,* 33(331), 1–23. https://doi.org/10.12795/rea.2016.i33.0

Haghighi, N. N., Liu, X. C., Wei, R., Li, W., Shao, H. (2018). Using Twitter data for transit performance assessment: A framework for evaluating transit riders' opinions about quality of service. *Public Transport*, 10(2), 363–377. https://doi.org/10.1007/s12469-018-0184-4

Ji, T., Fu, K., Self, N., Lu, C.-T., Ramakrishnan, N. (2018). Multi-task learning for transit service disruption detection. ASONAM 2018 : *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

Kocich, D. (2017). Multilingual sentiment mapping using twitter, Open source tools, and dictionary-based machine translation approach. *GIS Ostrava 2017*. https://doi.org/10.1007/978-3-319-61297-3_16

Manetti, G., Bellucci, M., Bagnoli, L. (2017). Stakeholder engagement and public information through social media : A study of Canadian and American public transportation agencies. *The American Review of Public Administration*, 47(8), 991–1009. https://doi.org/10.1177/0275074016649260

Miralles-Guasch, C., Martínez, M. (2013). Las fuentes de informaci´on sobre movilidad: La visión de los profesionales. Ejemplo de aplicación de metodología DELPHI. *Revista Transporte y Territorio*, (8), 99–116.

Steiger, E., de Albuquerque, J. P., Zipf, A. (2015). An advanced systematic literature review on spatiotemporal analyses of twitter data. *Transactions in GIS*, 19(6), 809–834. https://doi.org/10.1111/tgis.12132