# ESTUDIO DE LA MOVILIDAD INTERPROVINCIAL EN ESPAÑA MEDIANTE LA FUSIÓN DE DATOS DE TELEFONÍA MÓVIL CON OTRAS FUENTES DE DATOS

Javier Torres, Carlos Olivos, Miguel Picornell, Oliva García Cantú, Ricardo Herranz
Nommon Solutions and Technologies
nommon@nommon.es

## RESUMEN

Este artículo presenta un caso práctico del uso de datos provenientes de telefonía móvil para el estudio de la movilidad de media y larga distancia en España. El estudio es relevante tanto por su escala como por sus aportes metodológicos en áreas como la corrección de errores debidos a torres de telefonía mal posicionadas, la construcción de una muestra que incluye usuarios nacionales e internacionales, y la inferencia de diarios de viajes determinando origen, destino, duración, propósito, modo y ruta. Los resultados son validados mediante la comparación con la última encuesta nacional de transporte disponible.

*Palabras clave: matrices origen-destino, telefonía móvil, big data*

## ABSTRACT

This article presents a case study on the use of mobile network data in the analysis of medium- and long-distance mobility in Spain. The relevance of the study lies both in its scope and its methodological contributions in areas like the correction of errors due to mispositioned antennas, the construction of a sample that includes Spanish residents and foreign visitors, and the inference of travel diaries identifying the origin, destination, duration, purpose, mode and route of the trips. The results are validated through comparison with the latest national travel survey.

*Keywords: origin-destination matrices, mobile network data, big data*

## 1. INTRODUCCIÓN

La planificación y gestión de los sistemas de transporte requiere información precisa, fiable y actualizada acerca de la demanda de viajes. El perfil socioeconómico de los viajeros, el origen y destino de sus viajes, el propósito de los mismos, el modo de transporte y la ruta escogida son datos esenciales tanto para las autoridades de transporte, como para empresas operadoras y concesionarias de infraestructuras y servicios de transporte, con objeto de caracterizar la situación actual y modelizar el comportamiento futuro de la demanda. Los métodos tradicionales de recogida de información, basados en encuestas, aportan información detallada sobre la movilidad, pero presentan también algunas limitaciones importantes. En primer lugar, su realización consume mucho tiempo y presenta un alto coste, por lo que la muestra analizada es limitada, tanto geográficamente como en el tiempo, lo que repercute en la calidad de la información obtenida. Por el mismo motivo, a menudo no se dispone de información detallada sobre eventos especiales que influyen en la movilidad (fines de semana, periodos vacacionales, etc.). Finalmente, es común que las personas simplifiquen sus respuestas, por lo que la información puede ser no veraz o incorrecta. Estas limitaciones a menudo conducen a que proyectos de infraestructuras y servicios de transporte se planifiquen sobre la base de información incompleta o desactualizada. La pandemia de COVID-19 ha puesto aún más de manifiesto estas limitaciones, tanto por la imposibilidad de realizar trabajos de campo convencionales, como por la necesidad de monitorizar de manera continua unos patrones de actividad y movilidad cada vez más cambiantes.

Distintas tecnologías desarrolladas en los últimos años han abierto la puerta a la recolección pasiva de datos, sin necesidad de interactuar con los usuarios. Esta característica genera nuevas posibilidades para recoger información de movilidad a un coste muy inferior al de los métodos tradicionales, lo que ha producido gran interés entre los modeladores de transporte (Lee et al., 2016). Entre las fuentes más comunes de recolección pasiva de información de viajes podemos destacar los datos obtenidos vía GPS, sensores de *wifi* o *bluetooth*, tarjetas inteligentes de transporte y datos de telefonía celular procedentes de la conexión de los dispositivos móviles con la red de telefonía. Si bien cada una de estas fuentes presenta ventajas y limitaciones, los datos de telefonía móvil resultan particularmente interesantes, gracias a la posibilidad de obtener muestras de gran tamaño de prácticamente todos los segmentos de población, así como por proporcionar una resolución espacio-temporal suficientemente alta para la reconstrucción fiable de los patrones de movilidad de la población (Alexander et al., 2015; Picornell, 2017; García-Albertos et al., 2017). En Chen et al. (2016) se presenta una extensa revisión de la literatura y se analiza la integración y posibles sinergias entre las disciplinas de análisis de datos y los estudios de transporte tradicionales con vistas a entender de forma detallada los patrones de movilidad de los individuos.

Por otro lado, el análisis de los datos de telefonía móvil plantea nuevos desafíos. Si tomamos como referencia el modelo de transporte de cuatro etapas (Ortúzar y Willumsen, 2011), existen numerosos estudios que abordan de forma efectiva el uso de datos de telefonía móvil para la recogida de información acerca de las etapas de generación/atracción y de distribución de viajes (Dewulf et al., 2016; Windham et al., 2015). En cuanto a las etapas de reparto modal y asignación, existen dos grandes categorías de técnicas para la identificación del modo de transporte y las rutas seguidas por los viajes: *map matching* y aprendizaje no supervisado. Las técnicas de *map matching* se basan en superponer las posiciones de los registros de un usuario sobre la red de infraestructura de transporte (Brakatsoulas et al., 2005). La principal desventaja de este método es que requiere de información detallada sobre la infraestructura y los servicios de transporte en el área de estudio. Chen y Bierlaire (2014), Bonnetain et al. (2019) y Sakamanee et al. (2020), entre otros, han reportado el desarrollo de algoritmos de *map matching*

con resultados satisfactorios. Las técnicas de aprendizaje no supervisado pretenden discernir el modo y la ruta exclusivamente en función de los datos obtenidos de los registros móviles. Wang et al. (2010), por ejemplo, utilizan un algoritmo de tipo *k-means* para identificar el modo de transporte en función de los tiempos de viaje, agrupando tiempos similares en la misma categoría; Chen et al. (2020) han desarrollado un algoritmo que agrupa distintos modos en función de sus velocidades típicas, asumiendo que las velocidades de cada grupo siguen una distribución normal. Si bien estos métodos no necesitan datos sobre la red y los servicios de transportes, tienen la desventaja de que es necesario definir el número de *clusters* de forma arbitraria. Independientemente de la técnica escogida para la identificación de modo y ruta, se puede mejorar la predicción obtenida mediante el uso de datos de billetaje. Esto es lo que se hace en Braz et al. (2018), donde se toman datos de la infraestructura de la red de transporte y datos pasivos de localización de usuarios para estimar matrices origen-destino que luego se ajustan con datos de billetaje.

En este contexto, en 2017 el Ministerio de Fomento de España (hoy renombrado como Ministerio de Transportes, Movilidad y Agenda Urbana) encarga un análisis de la movilidad interprovincial de viajeros en el país. En el pasado, dicho estudio se realizaba a través de las encuestas Movilia, las cuales suponían un importante esfuerzo económico, técnico y humano. La última encuesta de este tipo, Movilia 2007, data de hace ya casi 15 años. En el año 2017, el Ministerio decidió sustituir la tradicional fuente de datos de la encuesta Movilia (preferencias declaradas), por información obtenida a partir de registros anonimizados de telefonía móvil (preferencias reveladas).

El objetivo del estudio era la obtención de matrices origen-destino de viajes interprovinciales segmentados por modo de transporte, para diferentes periodos del año 2017 (normal y estival), utilizando como fuente de datos principal registros generados por los terminales de telefonía móvil. Para cada uno de los días objeto de estudio, se analizó la movilidad de diferentes franjas temporales (punta de la mañana, periodo valle, punta de la tarde y periodo nocturno). En relación a los modos de transporte, se diferenció entre modo carretera (diferenciando a su vez entre vehículo privado y autobús), ferrocarril, aéreo y marítimo. En cuanto a la zonificación, se emplearon 59 zonas correspondientes a las 47 provincias de la España peninsular, las 10 islas principales de los archipiélagos de Baleares y Canarias, 2 zonas adicionales correspondientes a las ciudades autónomas de Ceuta y Melilla, y una última zona denominada ‘Extranjero’ para caracterizar los viajes con origen o destino fuera de España. Solo los viajes interprovinciales de más de 50 km fueron analizados, salvo para las provincias de Madrid, Barcelona, Alicante y Vizcaya, para las que se analizaron también los viajes de entre 10 y 50 km.

Además de presentar algunas consideraciones metodológicas que avanzan el estado del arte en la reconstrucción de diarios de movilidad a partir de la fusión de datos de telefonía móvil con otras fuentes de datos, el propósito de este artículo es presentar un caso práctico en el que se utilizan datos de telefonía móvil a gran escala para el estudio de la movilidad de un país. En primer lugar, el artículo presenta una descripción de los datos empleados. En segundo lugar, se describe la metodología diseñada para la generación de matrices origen-destino a partir de datos de telefonía móvil y su fusión con otras fuentes de datos. A continuación, se discute la validación de la metodología diseñada. Finalmente, se presentan las conclusiones del estudio, destacando las contribuciones de esta investigación, las limitaciones encontradas y las futuras líneas de trabajo propuestas.

## 2. DATOS

### 2.1 Telefonía móvil

La principal fuente de datos del estudio la constituyen los registros de telefonía móvil proporcionados por Orange España. En el año 2017, Orange registró la información de algo más de 14 millones de líneas móviles, lo que supone un 27,3% de cuota de mercado (fuente: CNMC, julio de 2017).

Los datos proporcionados pueden clasificarse en tres categorías:

- **Datos de eventos registrados**: datos asociados a los registros de comunicación del dispositivo con la red de telefonía móvil. Estos registros están constituidos principalmente por CDRs (Call Detail Records), que proporcionan información de la posición del usuario cada vez que éste interactúa con la red para efectuar una llamada, enviar o recibir un mensaje SMS, o hacer uso de una conexión de datos. A estos registros se les unen determinados eventos pasivos (actualización periódica de la posición del dispositivo, cambios de áreas de cobertura, etc.), dando lugar a unos "CDRs enriquecidos" con mayor granularidad temporal. Los datos de telefonía proporcionan una granularidad temporal muy elevada (especialmente en el caso de los usuarios de 'smartphones', que, en España, en el periodo de estudio, constituían más del 90% de los usuarios de telefonía móvil), que permite determinar con alto nivel de detalle la localización del dispositivo a lo largo del día. En cuanto a la granularidad espacial, se dispone de información de localización del dispositivo móvil a nivel de celda de telefonía, lo que supone una precisión espacial de decenas o cientos de metros en ciudad y hasta varios kilómetros en zonas rurales. En general, para un día medio, se dispone de alrededor de 1.000 millones de registros.
- **Datos de la topología de la red de telefonía móvil**: datos sobre la red de telefonía, incluyendo la localización de las torres y la orientación de las antenas.
- **Datos sociodemográficos**: información sociodemográfica de los clientes del operador. En este proyecto, la única información utilizada fueron los datos de edad y género.

Para el estudio se recopiló información de dos periodos del año 2017, un periodo laboral (octubre) y un periodo vacacional (julio-agosto).

### 2.2 Otros datos

Adicionalmente, se utilizaron datos de usos de suelo para mejorar la caracterización y la localización espacial de las actividades identificadas a partir de los datos de telefonía móvil; datos del Padrón Municipal de Habitantes de 2017 para la elevación muestral de la población residente en España; datos de movimientos turísticos para la elevación muestral de la población extranjera; datos de rutas por carreteras para la identificación de modo y ruta; datos de oferta y billetaje de servicios de autobuses interurbanos, para la detección y ajuste de la demanda de autobús; y datos de infraestructura, oferta y billetaje de transporte ferroviario, aéreo y marítimo, para la detección y ajuste de la demanda en estos modos.

## 3. METODOLOGÍA

La secuencia completa de los subprocesos para la obtención de las matrices origen-destino es la siguiente: (i) extracción y anonimización de los registros de telefonía móvil; (ii) pre-procesado y limpieza de los datos; (iii) construcción de la muestra; (iv) extracción de diarios de actividades y viajes; (v) elevación de la muestra al total de la población; (vi) ajuste de los resultados con datos de billetaje; (vii) generación de matrices origen-destino. Cada uno de estos subprocesos, así como los algoritmos asociados, se describen a continuación.

### 3.1 Extracción y anonimización de los registros de telefonía móvil

El primer subproceso consiste en la extracción y anonimización de los CDRs, realizada por el operador móvil. La anonimización de los CDRs está basada en la utilización de una función hash unidireccional, es decir, una función que permite el cálculo de un identificador anonimizado a partir del identificador original de tal forma que resulta imposible realizar el proceso a la inversa. Como resultado del proceso, se obtiene un conjunto de registros que, además de otros campos que no resultan de utilidad para el propósito de este estudio, incluyen la siguiente información:

ID_ dispositivo_anonimizado | ID_celda | fecha (DD/MM/YYYY) | hora (HH:MM:SS)

Por otro lado, se dispone de la localización de la torre de telefonía y la orientación de la antena correspondiente a cada identificador "ID_celda" (lo que permite estimar el área de cobertura de cada celda) y las características sociodemográficas (género, edad) asociadas al identificador "ID_ dispositivo_anonimizado".

### 3.2 Pre-procesado y limpieza de los datos

Una vez recibidos los datos anonimizados del operador, se realiza un pre-procesado de los mismos, ordenando los registros de la forma más eficiente para su análisis. Adicionalmente, se llevan a cabo distintos procesos de limpieza y depuración de errores.

El principal error a corregir es el registro incorrecto de la localización de ciertas torres. Algunos errores de posicionamiento pueden ser muy evidentes (por ejemplo, torres ubicadas en el mar), pero no siempre es así. Para depurar estos errores, se analizan temporalmente los registros, buscando saltos geográficos imposibles de ser realizados. Por ejemplo, un registro en Barcelona precedido y sucedido por una secuencia de registros muy cercanos en el tiempo localizados en el entorno de Madrid, con una diferencia de tiempo incompatible con un viaje en cualquier modo de transporte, indica un error en el posicionamiento de la correspondiente torre de telefonía, que es reubicada en el entorno de las torres en las que se registran los eventos inmediatamente anteriores y posteriores. Si bien este tipo de errores son poco frecuentes, pueden generar viajes espurios con una influencia importante en determinadas celdas de la matriz de viajes, por lo que su depuración resulta fundamental.

### 3.3 Construcción de la muestra

Se realiza una selección de los usuarios cuya actividad telefónica proporciona información suficiente para inferir su diario de actividades y viajes de manera fiable, para evitar la inclusión de usuarios que realicen actividades y viajes imposibles de detectar que puedan por tanto afectar a la calidad de las matrices origen-destino. La inclusión de eventos pasivos aumenta drásticamente la información espacio-temporal disponible de los usuarios y, por tanto, la muestra útil es muy elevada, pero aun así es posible, por ejemplo, que haya dispositivos que estén apagados durante un -día determinado, y por tanto no sea posible reconstruir su secuencia completa de actividades y viajes. Es importante también diferenciar estos casos de los dispositivos que desaparecen temporalmente de la red como consecuencia, por ejemplo, de un viaje en avión o un viaje al extranjero, que sí deben ser incluidos en la muestra. Esta consideración representa un aspecto habitualmente ignorado en los estudios disponibles en la literatura.

### 3.4 Extracción de diarios de actividades y viajes

El subproceso de generación de los diarios de actividades y viajes es el encargado de transformar los registros de telefonía móvil en información de movilidad. El objetivo es generar

un diario de actividades y viajes para cada identificador anonimizado incluido en la muestra y para cada uno de los días de estudio.

La información asociada a cada actividad incluye: (i) localización, (ii) tipo de actividad (distinguiendo entre hogar, trabajo, otras actividades frecuentes, y otras actividades no frecuentes), (iii) hora de inicio y (iv) hora de finalización.

La información asociada a cada viaje incluye: (i) origen (igual a la localización de la actividad inmediatamente anterior), (ii) destino (igual a la localización de la actividad inmediatamente posterior al viaje), (iii) hora de inicio (igual a la hora de finalización de la actividad anterior), (iv) hora de finalización (igual hora de inicio de la actividad siguiente), (v) si se trata de un viaje multietapa, las paradas intermedias (localización, hora de inicio de la estancia, hora de finalización de la estancia), (vi) modo de transporte empleado para cada etapa del viaje y (vi) en el caso de las etapas por carretera, ruta elegida para cada etapa.

Con el fin de describir en detalle los algoritmos utilizados para inferir esta información, es útil dividir dichos algoritmos en los siguientes grupos: (i) identificación y caracterización de estancias y actividades; (ii) identificación y caracterización de viajes y etapas.

### 3.4.1. Identificación de estancias y actividades

El primer paso consiste en identificar las estancias del usuario, es decir, su permanencia durante un cierto tiempo en un lugar determinado. Para ello, se analizan los registros consecutivos en una misma celda de telefonía y se identifican como estancias aquellas localizaciones en las que el individuo es detectado durante un tiempo superior a 5 minutos. A continuación, se identifican aquellas estancias que corresponden a actividades, distinguiéndolas de aquellas que corresponden a estancias intermedias subordinadas a un viaje y que son realizadas entre etapas del mismo. Para ello, los algoritmos desarrollados combinan distintos criterios basados en:

- Duración de la estancia: una estancia por encima de una determinada duración, se considera actividad. Para ello se utiliza un umbral adaptativo en función del tipo de viaje, que tiene en cuenta factores como la localización de la estancia y la distancia viajada. Por ejemplo, una parada de 30 minutos en un viaje urbano es considerada una actividad, mientras que, en un viaje interurbano, una parada de 30 minutos precedida y sucedida por desplazamientos de 2 horas podría corresponder a una parada intermedia.
- Localización de la estancia: por ejemplo, una estancia de más de 3 horas en un lugar de un núcleo urbano con toda probabilidad es una actividad, mientras que una estancia de 3 horas en un aeropuerto puede corresponder a una conexión entre vuelos y por tanto no constituir una actividad.
- Patrones longitudinales de comportamiento: por ejemplo, el lugar donde el usuario pernocta habitualmente se etiqueta como ‘Hogar’.
- Los itinerarios de los desplazamientos: por ejemplo, una estancia que no forma parte de ningún trayecto lógico entre las estancias anterior y posterior debe constituir una actividad, aunque no cumpla el criterio de duración mínima.

Si bien los criterios de tiempo de estancia han sido utilizados con anterioridad para la distinción entre actividades y estancias en distintos estudios (Hariharan & Toyama, 2004; Wang et al, 2010; Windham et al, 2015), la coherencia en el itinerario de los desplazamientos no ha sido explotada con anterioridad. Este criterio permite identificar actividades de corta duración que se perderían si se aplicase solo un criterio de duración y constituye una de las aportaciones metodológicas del estudio.

Utilizando estos criterios, se refina la localización de actividades y estancias mediante la fusión con datos de usos del suelo, asociándolas a un área dentro de la correspondiente celda de telefonía (por ejemplo, si el hogar de un usuario se sitúa en una celda de telefonía que incluye suelo residencial y suelo forestal, la localización de la actividad 'Hogar' se asigna a la zona de suelo residencial), y se clasifican las actividades en cuatro categorías principales: "Hogar" (incluye las actividades realizadas en el lugar de residencia habitual), "Trabajo/Estudio" (actividad recurrente de una duración significativa (6-8 horas) que se realiza en un lugar diferente al de residencia), "Otras actividades frecuentes" (actividades que se realizan de manera recurrente en una misma localización distintas de "Hogar" y "Trabajo/Estudio") y "Actividades No Frecuentes" (actividades que se realizan sin una frecuencia preestablecida).

### 3.4.2. Caracterización de viajes y etapas. Identificación de modo y ruta

A partir de las estancias identificadas, se caracterizan los viajes realizados entre las distintas actividades y las etapas que los componen, determinando la hora de inicio y de finalización, el modo de transporte y la ruta. Los algoritmos utilizados para determinar estos atributos se basan en la siguiente aproximación:

- Análisis de la oferta: utilizando la información de la red (red de carreteras, red ferroviaria, aeropuertos, puertos y estaciones de autobuses) y la oferta de servicios (conexiones disponibles, horarios, etc.), se identifican todas las alternativas posibles para desplazarse desde el origen al destino, las etapas que las componen, y su duración.
- Determinación de modo y ruta: los registros espacio-temporales generados durante el viaje se superponen con los itinerarios correspondientes a las distintas alternativas identificadas (trazado de las carreteras y vías de ferrocarril, localización de puertos y aeropuertos) mediante técnicas de *map-matching*, seleccionando el modo de transporte (y, en el caso de los viajes por carretera y tren, la ruta) compatible con dichos registros. Aunque es poco común en viajes de media y larga distancia, pueden presentarse casos en los que aparezca alguna ambigüedad (por ejemplo, si el trazado de una carretera transcurre muy cerca al del ferrocarril de principio a fin, o si dos carreteras alternativas discurren muy próximas). En esos casos, las técnicas de *map-matching* se complementan con criterios basados en velocidades y tiempos de viaje estimados. Si la ambigüedad persiste y siguen existiendo dos o más modos/rutas compatibles con los registros observados, lo cual normalmente ocurre en un porcentaje pequeño de los desplazamientos, el modo y la ruta se asignan de acuerdo con las distribuciones obtenidas del resto de la muestra para ese par origen-destino. Al final de este proceso, los resultados quedan clasificados según el grado de certeza de la identificación de modo de viaje, distinguiendo cuando se tiene certeza, cuando existe un subconjunto posible de modos y cuando el modo de viaje no es identificable. Posteriormente, cuando se realice el juste con datos de billetaje, se terminará de asignar el modo entre las cuatro opciones posibles: carretera, ferroviario, aéreo o marítimo.
- Determinación de la hora de inicio y fin de cada etapa. Combinando la información sobre la duración de cada etapa para el modo y ruta identificados y la distribución espacio-temporal de los registros generados durante el viaje, se determina la hora de inicio y fin de cada etapa del viaje.
- Clasificación de vehículos para los viajes por carretera. A partir de los patrones de movilidad de los usuarios a lo largo de varias semanas (kilómetros recorridos, tiempos de parada en viajes largos y días de "inactividad", estancia en centros logísticos, etc.), se divide a los usuarios en dos grupos: usuarios con movilidad característica de transportistas (vehículos de mercancías, aquellos con un mayor número de kilómetros recorridos y con descansos comparables con aquellos requeridos por la legislación) y resto de usuarios

(vehículos ligeros y autobús), con el fin de excluir los viajes de vehículos de mercancías de la matriz de viajes de pasajeros. El uso de indicadores obtenidos de los patrones de movilidad para la distinción de vehículos es otra de las novedades en el estado del arte sobre reconstrucción de la movilidad a partir de telefonía móvil. Finalmente, se emplea la información de oferta y demanda de las concesiones de autobuses para realizar la segmentación entre vehículo privado y autobús, empleando técnicas de *map-matching* similares a las utilizadas para identificar los modos ferroviario, aéreo y marítimo.

### 3.5 Elevación de la muestra al total de la población

Cabe distinguir dos casos:

- Residentes en España: la elevación de la muestra de viajeros residentes en España se realizará tomando como marco muestral la población residente en el país, según datos del Padrón de Habitantes proporcionados por el Instituto Nacional de Estadística (INE). Se aplican factores de expansión por lugar de residencia y estrato sociodemográfico (género y franja de edad) a nivel de sección censal. Los factores corresponden a la división de la población efectiva y la muestra obtenida, para cada sección censal y grupo muestral.
- Visitantes residentes en el extranjero: la elevación de la muestra se realiza empleando factores de expansión basados en nacionalidad y tipo de visitante (turista o excursionista), tomando como marco muestral los datos de la encuesta de movimientos en frontera (FRONTUR) proporcionados por el INE. Una vez más, los factores corresponden a la división del marco muestral por la muestra obtenida para cada país o grupo de países y tipo de visitante.

### 3.6 Ajuste de resultados con datos de billetaje

Los resultados obtenidos a partir de telefonía móvil se han ajustado utilizando los datos de billetaje disponibles para los modos autobús, ferroviario, aéreo y marítimo. En función del grado de certeza con la que se determina el modo de transporte a partir de los datos de telefonía móvil, los viajes identificados se clasifican en tres grupos:

- Determinados: son aquellos casos para los que se determina con certeza el modo de transporte a partir de técnicas de *map-matching*.
- Indeterminados: son aquellos casos en los que se conoce con certeza el subconjunto de opciones de transporte que son compatibles con la información de telefonía móvil generada a lo largo del viaje. Por ejemplo, es posible determinar que un viaje Madrid-Barcelona no se ha realizado en avión, pero no se puede determinar con certeza si el viaje se ha realizado en ferrocarril o por carretera.
- Desconocidos: son aquellos casos en los que cualquiera de las opciones de transporte es válida en base a la información de telefonía móvil disponible.

Para definir el método de ajuste de resultados, se han tenido en cuenta los siguientes criterios:

1. Tomar como punto de partida la información obtenida de los datos de telefonía móvil, con el fin de aprovechar al máximo la información que no es posible obtener por otras fuentes, en particular los orígenes y destinos puerta-a-puerta y los viajes en vehículo privado. Con objeto de no distorsionar artificialmente los resultados obtenidos, se ha decidido emplear un método de ajuste que proporcione valores compatibles con los resultados obtenidos del análisis de los datos de telefonía móvil.
2. Utilizar los datos de billetaje para ajustar los resultados dentro del intervalo de valores posibles proporcionado por la telefonía.
3. Emplear una aproximación desagregada, incorporando de manera individual los viajes ‘indeterminados’ y ‘desconocidos’ a cada modo de transporte, en lugar de aplicar un

factor de ajuste agregado, con el objeto de conservar las características de relaciones origen-destino puerta-a-puerta de los viajes muestrales y de mantener la coherencia entre los distintos indicadores de demanda (matriz de viajes, matriz de etapas, etc.).

Teniendo en cuenta estos criterios, el método de ajuste seleccionado es el siguiente:

- Para cada par origen-destino, el número mínimo de viajes en un modo concreto de transporte lo fija el valor de viajes 'determinados' identificados en dicho modo a partir de los datos de telefonía, mientras que el máximo viene determinado por el máximo valor compatible con la información obtenida mediante telefonía. Así, una vez aplicado el ajuste, el volumen de viajes en cada modo para un determinado par origen-destino estará contenido entre: (i) el número de viajes 'determinados' para ese modo; (ii) la suma de los viajes 'determinados' para dicho modo, los viajes 'indeterminados' para los que el modo es una alternativa posible, y los viajes con modo 'desconocido'.
- Si el volumen de viajes 'determinados' es superior al dato de billetaje, se mantiene el valor de viajes determinados obtenido a partir de los datos de telefonía. De esta forma, el método es robusto frente a posibles deficiencias o falta de información en los datos de demanda procedentes de fuentes alternativas a la telefonía móvil, permitiendo realizar estimaciones que superan las limitaciones de las otras fuentes de datos (por ejemplo, los viajes en transporte marítimo pueden incluir viajes en embarcaciones privadas que no están registrados en los datos de billetaje proporcionados por Puertos del Estado).
- Si el volumen de viajes determinados para un modo concreto es inferior al reportado por los datos de billetaje, el valor se ajusta al alza teniendo en cuenta el volumen de viajes 'indeterminados' en dicho modo y los casos 'desconocidos', incorporando dichos viajes al grupo de 'determinados'. Los viajes 'indeterminados' y 'desconocidos' se ordenan según su grado de similitud con los distintos modos, de forma que los viajes con mejores métricas de similitud para cada modo son los primeros en pasar al grupo de 'determinados' de dicho modo. El ajuste de cada modo finaliza una vez se alcanza el volumen de referencia proporcionado por los datos de billetaje o cuando ya no existen, en base a los resultados de telefonía, más viajes compatibles con el modo en cuestión.

### 3.7 Generación de indicadores de demanda de transporte

Una vez los resultados han sido ajustados, se generan las matrices origen-destino especificadas con el nivel de agregación espacial y temporal requerido.

## 4. RESULTADOS

### 4.1 Ejemplos de los resultados obtenidos

Los resultados del estudio incluyen matrices de etapas y viajes para todos los días del periodo de estudio, segmentadas por hora del día, modo de transporte, propósito del viaje y lugar de residencia de los viajeros.

A modo ilustrativo, las Figuras 1 y 2 muestran algunos de los resultados del estudio.

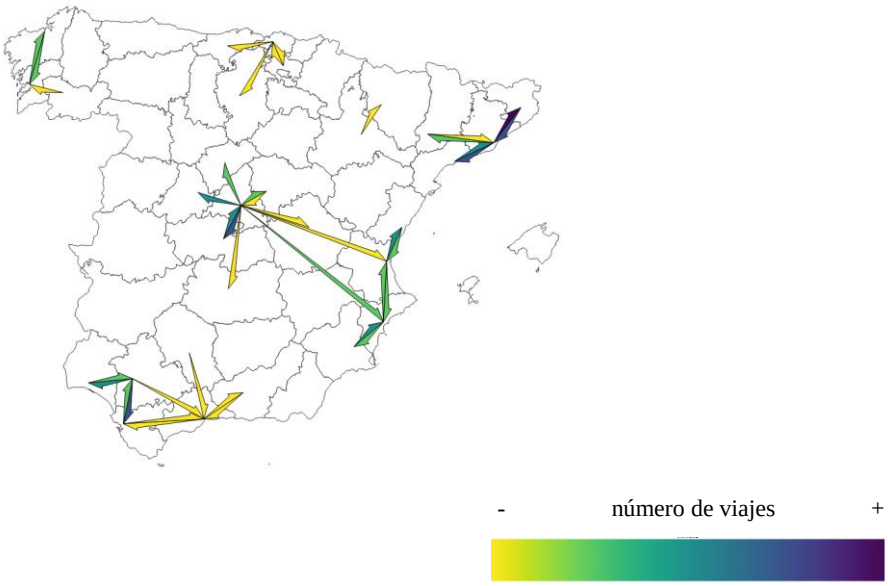


**Figura 1**. Ejemplo de los resultados obtenidos:
flujos por par origen-destino para el día 14 de julio de 2017.

El día 14 de julio de 2017 corresponde a una "operación salida" desde las principales ciudades de España por el comienzo de las vacaciones de muchos residentes, generando muchos viajes turísticos. En la Figura 1 podemos observar, por ejemplo, como grandes flujos se desplazan fuera de Madrid, principalmente hacia el entorno cercano, pero también hacia la costa. Podemos también contrastar estos patrones con la Figura 2, donde se observa un jueves promedio de octubre de 2017, presentando una menor cantidad de viajes saliendo de Madrid y menor número de viajes hacia la costa. Además, en la Figura 2 se observa un flujo considerable entre Madrid y Barcelona, que no se aprecia en la Figura 1, correspondiente a viajes de negocios normalmente ejecutados en el día.

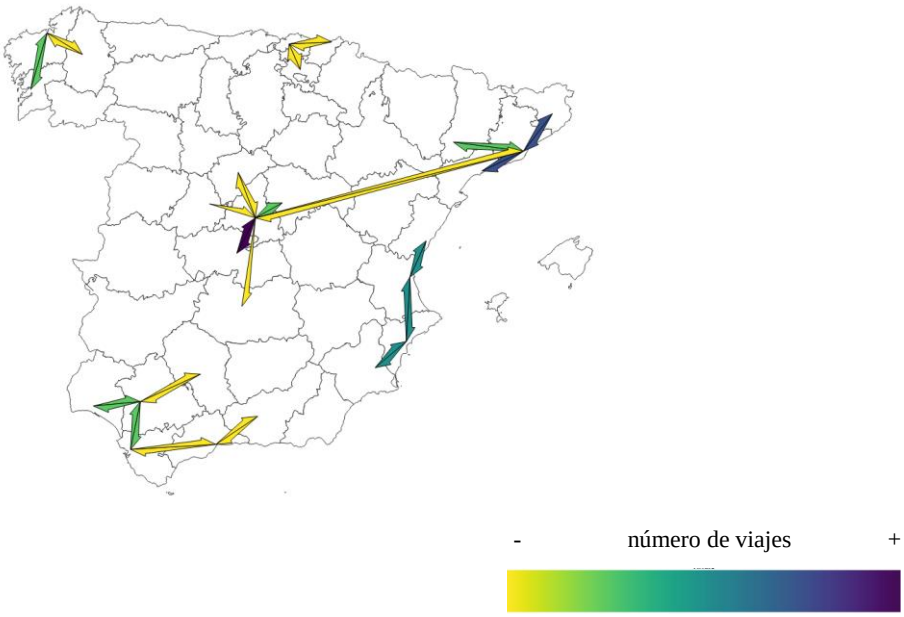


**Figura 2**. Ejemplo de los resultados obtenidos:
flujos por par OD para un jueves promedio de octubre de 2017.

Parte importante de la propuesta de valor de este estudio es la posibilidad de estudiar un elevado número de días particulares en un mismo año, lo que resulta económicamente inviable mediante métodos tradicionales como los de Movilia 2007. Los resultados completos del estudio están disponibles como datos abiertos a través de la página web del Ministerio de Transportes, Movilidad y Agenda Urbana, en el enlace https://observatoriotransporte.mitma.es/estudio-experimental.

### 4.2 Validación con estudios anteriores

La validación de este estudio normalmente implicaría que existiese otro con el cual contrastar. En este sentido, el estudio más similar se corresponde a la encuesta Movilia de 2007. No obstante, al realizar la comparativa hay que tener en consideración que (i) existe una diferencia de diez años entre ambos estudios durante los que se han inaugurado grandes infraestructuras en España (por ejemplo, la línea de Alta Velocidad Madrid-Levante) que influyen en la estructura de la matriz; y (ii) la publicación de los resultados se realizó a nivel de las 17 Comunidades Autónomas (nivel administrativo superior a provincia) por lo que los viajes intra-Comunidad Autónoma no se pueden comparar. Pese a estas limitaciones, dado que no existe una encuesta reciente de las dimensiones del estudio, se hicieron validaciones indirectas comparando las estructuras de los flujos con la anterior encuesta Movilia del 2007.

El orden de magnitud de los volúmenes de ambas encuestas, para cada par origen-destino se presenta en la Figura 3. Se puede observar que los volúmenes de viajes obtenidos con telefonía móvil son en promedio tres veces los obtenidos en el estudio anterior.


**Figura 3**. Volúmenes de viajes de ambos estudios.

Para validar, se categorizaron los pares origen-destino de cada estudio en función de su importancia relativa en volumen de viajes. De este modo, el par origen-destino con mayor volumen de viajes se encuentra en la primera posición, el siguiente en la segunda, y así sucesivamente. La Figura 4 muestra la correlación entre las posiciones de cada par origen-destino en cada estudio.

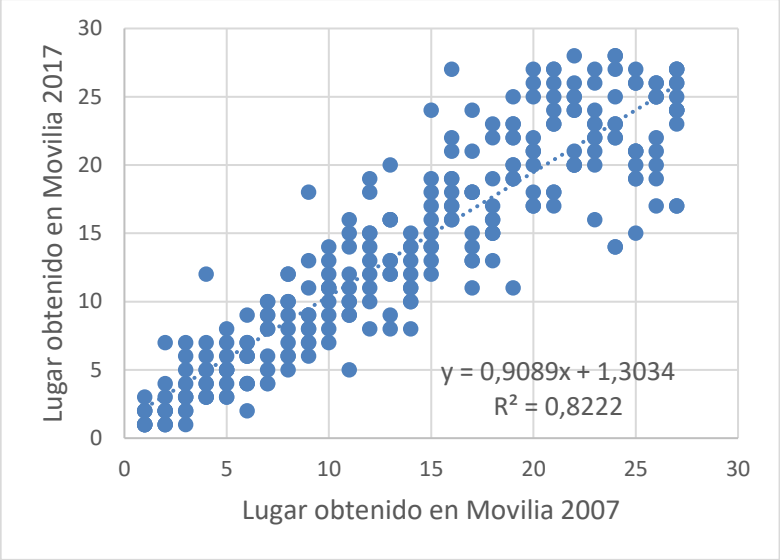


**Figura 4**. Comparación de ambos estudios: importancia relativa de los distintos pares origen-destino

Se puede observar que, aun cuando el volumen de viajes cambió en el transcurso de los diez años que pasaron entre ambos estudios, la jerarquía de pares origen-destino no sufre grandes cambios. La mayoría de los pares origen-destino en 2017 mantienen un lugar muy cercano al que tenían en 2007; el cambio promedio de lugar es de -0.05 puestos, con una desviación estándar de 0.37 puestos.

Es importante destacar que hubo múltiples instancias de validación durante el transcurso del proyecto basadas en el juicio de expertos, en las que participaron los principales administradores de infraestructuras de España (ADIF, AENA, Puertos del Estado) y algunos operadores como RENFE. Finalmente, los técnicos del Ministerio hicieron una validación global de los resultados, tomando en cuenta las validaciones indirectas y por modos antes mencionados. Los resultados fueron considerados válidos y se están usando actualmente como base para la planificación estratégica de infraestructuras en el marco del crecimiento a futuro del país, incluyendo el desarrollo del modelo nacional de transporte.

### 4.3 Validación de resultados de billetaje

Para validar los resultados obtenidos, se realizó una comparativa entre la información de demanda de transporte generada únicamente a partir de los datos de telefonía móvil (asignando los viajes ‘indeterminados’ y ‘desconocidos’ probabilísticamente, según la distribución observada en los viajes ‘determinados’, de acuerdo al método descrito en la sección 3.4.2) y la información de demanda proporcionada por los operadores. En concreto, se compararon los volúmenes de viaje obtenidos a partir de los datos de telefonía móvil en modo ferroviario y aéreo con los datos de billetaje proporcionados por los operadores y gestores de la infraestructura (ADIF y RENFE para modo ferroviario y AENA para modo aéreo).

En la Figura 2 se muestra la comparativa de los resultados para los modos ferrocarril y aéreo antes y después del proceso de ajuste para un día tipo.

Previamente al ajuste, puede observarse que para el modo ferroviario la correlación es elevada (R^2= 0.76) y la pendiente de la recta cercana a la unidad (1.01), lo que muestra que los volúmenes en media obtenidos con telefonía móvil y billetaje son similares. Cabe señalar que, en determinados pares origen-destino, para los que los datos de billetaje no incluyen información de los trenes de Cercanías, es esperable que el número de viajes estimados con la telefonía sea superior al volumen del billetaje, que solo incluye servicios de media y larga distancia. Para el caso del modo aéreo puede observarse como la correlación es igualmente elevada (R^2= 0.84) y la pendiente de la recta es también cercana a la unidad (0.89). Si bien los resultados de telefonía móvil por sí solos ya proporcionan una información de calidad, los resultados pueden mejorarse mediante la fusión con datos de billetaje, como se aprecia en los valores de correlación obtenidos después del ajuste. Los valores de correlación tras el ajuste aumentan a 0.90 para el modo ferroviario y a 0.95 para el modo aéreo, con pendientes cercanas a la unidad y disminuyendo la dispersión de los resultados.

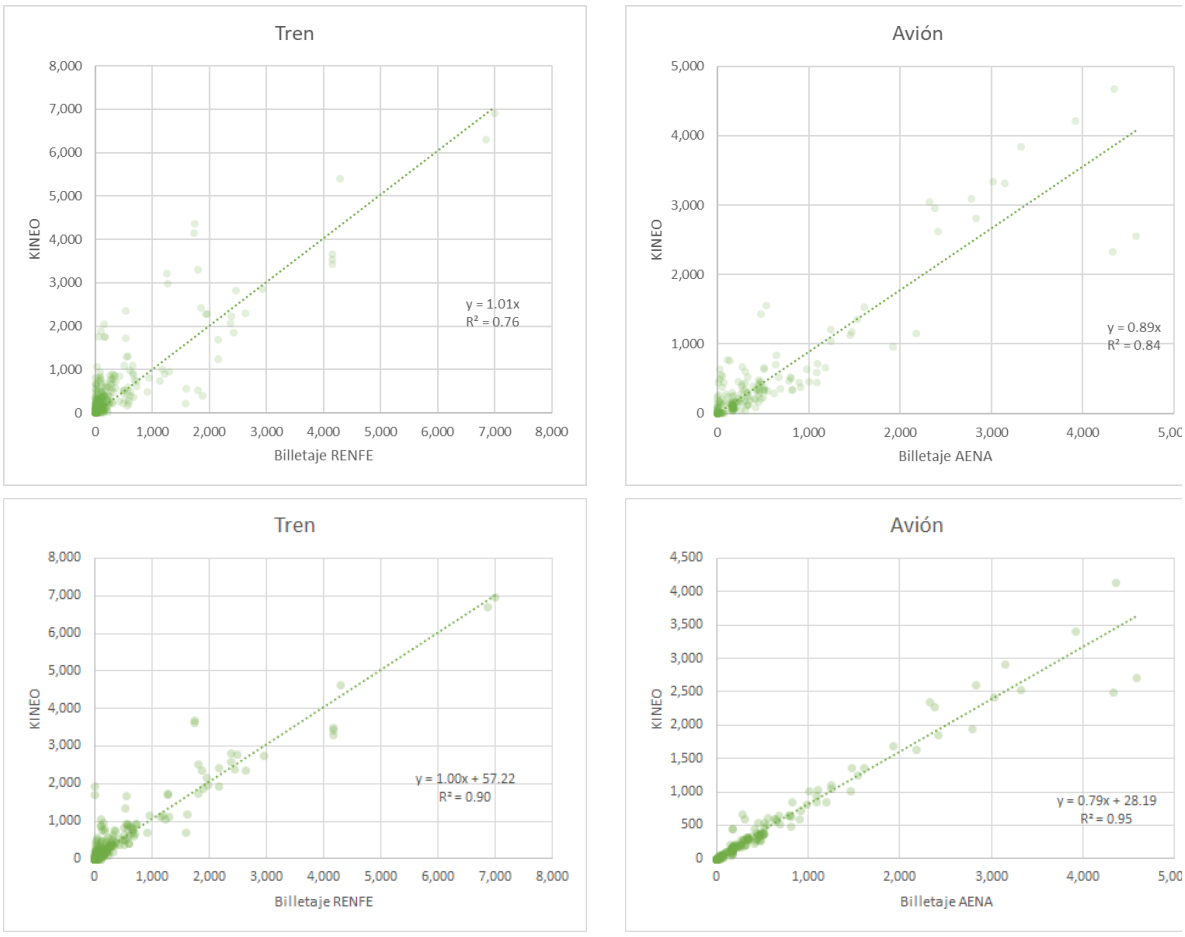


**Figura 2**. Comparativa de los resultados para los modos ferrocarril y aéreo antes (fila superior) y después (fila inferior) del proceso de ajuste para un día tipo (miércoles promedio de octubre de 2017).

## 5. CONCLUSIONES

Las principales conclusiones del estudio se enumeran a continuación:

- En base a los análisis de validación realizados y la bondad de los resultados obtenidos, se puede afirmar que la información de demanda de transporte generada a partir de los datos de telefonía móvil es de gran calidad. La metodología empleada permite identificar de manera satisfactoria el modo de transporte en la mayoría de los viajes analizados, siendo de especial relevancia la determinación del volumen y estructura de los viajes por carretera, para los que no existe ninguna otra fuente fiable.
- Los datos de telefonía móvil permiten obtener un tamaño de muestra muy superior al que es posible obtener a un coste razonable mediante otras metodologías como las encuestas domiciliarias. Es previsible que este mayor tamaño muestral resulte en una mejor estructura de viajes, así como una mejor caracterización de determinadas variables como el reparto modal puerta a puerta.
- Los costes y plazos de ejecución del estudio llevado a cabo han sido muy inferiores a los que se hubieran requerido en caso de emplear otras metodologías como las basadas en encuestas y trabajos de toma de datos en campo.

### 5.1 Limitaciones

A pesar de las ventajas que presenta el uso de datos de telefonía móvil como fuente de datos principal de demanda de transporte, es importante señalar también algunas limitaciones:

- Error muestral: a pesar de disponer de una muestra de usuarios muy superior a la de los métodos tradicionales basados en encuestas, es importante tener en cuenta que el análisis se realiza sobre una muestra de usuarios de la población y no sobre el total, por lo que, al igual que en una encuesta, los resultados presentan un error muestral intrínseco a este tipo de aproximaciones de inferencia poblacional. La metodología de ajuste con otras fuentes de datos ayuda a paliar los posibles errores cometidos en la inferencia.
- Errores en los marcos muestrales: para los procesos de elevación muestral, es necesario emplear marcos muestrales asociados a la población objeto de estudio. Estas fuentes de datos no están exentas de errores que posteriormente tendrán repercusión en la calidad de los resultados obtenidos.
- Resolución espacio-temporal de los datos: las características espacio-temporales de los datos suelen ser suficientes para determinar las características de los viajes en la mayoría de los casos. No obstante, para algunos viajes de corta distancia puede ocurrir que la información de telefonía móvil no permita identificar el modo de transporte. Del mismo modo, en viajes de media-larga distancia donde la oferta de transporte es similar en cuanto a trazado, horario, tiempos de viaje, etc., la precisión espacio-temporal de los datos de telefonía móvil no permite en determinadas ocasiones distinguir de manera certera el modo y/o ruta del viaje. Para solventar estas limitaciones, se emplean las técnicas de ajuste con otras fuentes de datos (billetaje).
- Calidad de la información sociodemográfica: la información sociodemográfica obtenida a partir de los datos de telefonía móvil es limitada en comparación a la que normalmente se obtiene a través de encuestas. Adicionalmente, la información del dispositivo móvil suele ir asociada al titular del contrato, y no al usuario del dispositivo, lo que puede generar ciertos errores en las estimaciones donde se consideren las variables de edad y género proporcionadas por el operador.

## 5.2 Futuras líneas de trabajo

En base a los resultados y conclusiones extraídos del estudio, se han identificado las siguientes futuras líneas de trabajo:

- Una de las fuentes de datos complementarias a los datos de telefonía móvil que influyen de forma más significativa en la calidad de los resultados del estudio son los datos de red y oferta de transporte. Aunque una parte importante de los recursos del proyecto se ha destinado a recoger información de calidad al respecto, existe aún margen de mejora para depurar esta información. Por ejemplo, mejorar los datos de ubicación de las paradas de autobús para algunas concesiones, lo que resulta fundamental para la correcta identificación del modo autobús, ayudaría a refinar los resultados obtenidos.
- Incorporación de datos procedentes de sondas de red, que recogen, de forma pasiva cada pocos segundos la localización de todos los usuarios sin necesidad de que el teléfono realice ninguna interacción con la red. Es de esperar que estos datos permitan aumentar el tamaño de la muestra útil y la precisión de los algoritmos de *map-matching*, aumentando así la calidad de los resultados.
- Ampliar el alcance del estudio para analizar de movilidad en ámbitos metropolitanos. Esto requerirá la revisión de las técnicas de *map-matching* y/o la incorporación de nuevas aproximaciones para la estimación de modo y ruta en ámbitos urbanos.
- Desarrollar algoritmos para la determinación del perfil sociodemográfico de los usuarios a partir del análisis de sus patrones de comportamiento (e.g. distancia de sus viajes, lugar de trabajo), mejorando la caracterización del viajero y los procesos de elevación muestral. También se está estudiando la incorporación de otros datos disponibles para el operador (por ejemplo, perfiles de navegación web de los usuarios) que permitan refinar la estimación de sus características sociodemográficas.

## REFERENCIAS


Alexander, L., Jiang, S., Murga, M., & 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.

Bonnetain, L., Furno, A., Krug, J., & Faouzi, N. E. E. (2019). Can We Map-Match Individual Cellular Network Signaling Trajectories in Urban Environments? Data-Driven Study. *Transportation Research Record*, *2673*(7), 74-88.

Brakatsoulas, S., Pfoser, D., Salas, R., & Wenk, C. (2005, August). On map-matching vehicle tracking data. In *Proceedings of the 31st International Conference on Very Large Data Bases*, 853-864.

Braz, T., Maciel, M., Mestre, D. G., Andrade, N., Pires, C. E., Queiroz, A. R., & Santos, V. B. (2018). Estimating inefficiency in bus trip choices from a user perspective with schedule, positioning, and ticketing data. *IEEE Transactions on Intelligent Transportation Systems*, *19*(11), 3630-3641.

Chen, J., & Bierlaire, M. (2015). Probabilistic multimodal map matching with rich smartphone data. *Journal of Intelligent Transportation Systems*, *19*(2), 134-148.

Chen, C., Ma, J., Susilo, Y., Liu, 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.

Chen, X., Xu, X., & Yang, C. (2020). Trip mode inference from mobile phone signaling data using logarithm Gaussian mixture model. *Journal of Transport and Land Use*, *13*(1), 429-445.

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

García-Albertos, P., Cantú Ros, O. G., Herranz, R., & Ciruelos, C. (2017). Understanding door-to-door travel times from opportunistically collected mobile phone records. *SESAR Innovation Days 2017*.

Hariharan, R., & Toyama, K. (2004, October). Project Lachesis: parsing and modeling location histories. In *International Conference on Geographic Information Science* (pp. 106-124). Springer, Berlin, Heidelberg.

Lee, R. J., Sener, I. N., & Mullins III, J. A. (2016). An evaluation of emerging data collection technologies for travel demand modeling: from research to practice. *Transportation Letters*, *8*(4), 181-193.

Ortúzar, J. D., & Willumsen, L. G. (2011). *Modelling Transport*. John Wiley & Sons.

Picornell Tronch, M. (2017). *Metodología para la extracción de patrones de movilidad urbana mediante el análisis de registros de actividad telefónica (Call Detail Records)* (Tesis Doctoral, Universitat Politècnica de València).

Sakamanee, P., Phithakkitnukoon, S., Smoreda, Z., & Ratti, C. (2020). Methods for Inferring Route Choice of Commuting Trip From Mobile Phone Network Data. *ISPRS International Journal of Geo-Information*, *9*(5), 306.

Wang, H., Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010, September). Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *13th International IEEE Conference on Intelligent Transportation Systems* (pp. 318-323). IEEE.

Widhalm, P., Yang, Y., Ulm, M., Athavale, S., & González, M. C. (2015). Discovering urban activity patterns in cell phone data. *Transportation*, *42*(4), 597-623.