Secretaría de Estado de Transportes, Movilidad y Agenda Urbana

Análisis de la movilidad en España con tecnología Big Data durante el estado de alarma para la gestión de la crisis del COVID-19

Informe metodológico

14 de abril de 2021

Índice

1. Introducción

1.1 Antecedentes

El Ministerio de Transportes, Movilidad y Agenda Urbana considera necesario analizar los cambios que se están produciendo en la movilidad de los españoles durante la crisis del COVID-19, con el objetivo de generar información que sirva tanto para evaluar el efecto de las medidas de restricción de la movilidad impuestas a la ciudadanía como para otros análisis y estudios que ayuden a la gestión y a la posterior salida de esta crisis. Dada su capacidad para producir información de calidad y con un elevado nivel de detalle en plazos de tiempo muy cortos, el Ministerio ha considerado que la solución óptima para responder a esta necesidad es el empleo de soluciones basadas en el análisis de datos masivos, aprovechando la experiencia adquirida en el proyecto 'Estudio de la Movilidad Interprovincial de Viajeros aplicando la Tecnología Big Data' realizado en 2018, que por su alcance y complejidad fue pionero en este campo a nivel internacional. Al igual que en el estudio anteriormente mencionado, el presente estudio emplea como fuente de datos principal registros anonimizados procedentes de las redes de telefonía móvil. Dichos registros, generados originalmente a efectos de facturación o de gestión de la red, proporcionan muestras de gran tamaño y con una elevada resolución espacio-temporal de prácticamente todos los segmentos de población y se han usado con éxito en numerosos estudios sobre movilidad y demanda de transporte a lo largo de los últimos años, por lo que resultan particularmente apropiados para el propósito del presente estudio. El estudio se apoya en una solución técnica y una metodología similares a las del estudio de movilidad interprovincial realizado por el Ministerio en 2018. La información procedente de las redes de telefonía móvil se ha fusionado con otras fuentes de datos para generar matrices origen-destino y otros indicadores de movilidad y presencia de población anónimos y agregados, garantizando el estricto cumplimiento con los requisitos de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (LOPD-GDD). Este documento describe los datos utilizados en el proyecto, la metodología y los algoritmos de análisis de los datos, y los indicadores generados.

1.2 Objeto y alcance del estudio

El proyecto contempla la generación de matrices origen-destino y otros indicadores de movilidad.

Las especificaciones del estudio se detallan a continuación.

Tabla 1. Especificación del estudio

Especificaciones del estudio	
Población de estudio	Población residente en España.
Zonificación	Los indicadores se calculan para una zonificación específica definida por el MITMA ("zonas MITMA"). En la mayoría de los casos, las zonas MITMA se corresponden con municipios. Los municipios más pequeños se agregan siguiendo el mismo criterio empleado por el INE en el estudio del análisis de las relaciones casa-trabajo a partir de datos de telefonía móvil realizado por el INE en 2019. Esto permite realizar las agregaciones pertinentes para proporcionar también indicadores a nivel de provincia, a nivel de CCAA y a nivel nacional.
Días de estudio	Se analiza la movilidad y la distribución de la población en el territorio tras la aplicación del real decreto 463/2020, de 14 de marzo, de manera diaria hasta la finalización del estado de alarma. Asimismo, el estudio incluye también las dos semanas anteriores al estado de alarma. Una vez finalizado el estado de alarma, se analizarán las semanas posteriores a dicha finalización hasta observar un nivel estable de movilidad. El estudio incluye también un análisis de una semana tipo (del 14 al 20 de febrero de 2020), con objeto de evaluar la reducción de la movilidad con respecto a un conjunto de días con niveles habituales de movilidad.
Viajes objeto de estudio	Se analizan todos los viajes de más de 500 metros con origen y destino dentro de España.

Indicadores	Se proporcionan los siguientes indicadores:
	- Matrices origen-destino, segmentando los viajes:
	  ○ en tramos de 1 hora según la hora de inicio del viaje;
	  ○ según la distancia ortodrómica entre el origen y el destino, distinguiendo 6 rangos de distancia: 0,5-2 km, 2-5 km, 5-10 km, 10-50 km, 50-100 km y más de 100 km.
	Para cada elemento de la matriz de viajes, se proporciona también el total de viajeros-km correspondiente a dicho par origen-destino, según la distancia ortodrómica entre el origen y el destino.
	- Distribución del número de viajes por persona, distinguiendo entre las personas que no realizan ningún viaje, las que realizan 1 viaje, las que realizan 2 viajes, y las que realizan más de 2 viajes.

2. Fuentes de datos utilizadas

2.1 Registros anonimizados de telefonía móvil

La principal fuente de datos la constituyen los registros anonimizados de telefonía móvil. El estudio parte de una muestra de datos de más de 13 millones de líneas móviles proporcionada por un operador móvil, que podría ser incrementada a lo largo del proyecto, en la medida que se cuente con los datos de más operadores.

Los datos de partida pueden clasificarse en dos categorías:

- **Datos de eventos registrados**[1]: datos anonimizados asociados a los registros de conexión de los dispositivos móviles con la red de telefonía móvil. Estos registros incluyen tanto eventos activos como pasivos. Los eventos activos están constituidos por lo que se denomina CDRs (Call Detail Records), que proporcionan un registro cada vez que el dispositivo interactúa con la red (llamadas, envío de mensajes de texto, sesiones de datos). A estos registros se les unen datos de eventos pasivos (actualización periódica de la posición del dispositivo, cambios de áreas de cobertura, etc.), proporcionando una granularidad temporal muy elevada. En cuanto a la resolución espacial, se dispone de información de localización 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.
- **Datos de la topología de la red de telefonía móvil**: datos sobre la red de telefonía, incluyendo la localización de las torres de comunicación y la orientación de las antenas.

2.2 Usos del suelo

Se han utilizado también datos de usos del suelo, para mejorar la caracterización y la localización espacial de las actividades identificadas a partir de los datos de telefonía móvil. Los datos de usos del suelo proceden del Sistema de Información sobre Ocupación del Suelo de España (SIOSE) y de otras bases de datos disponibles a nivel autonómico.

2.3 Datos de población

Para los procesos de elevación de la muestra se han utilizado datos procedentes del Padrón Municipal de Habitantes.

2.4 Datos de la red de transporte

Los algoritmos empleados para la identificación de viajes emplean también información de la red de transporte (por ejemplo, localización de aeropuertos, red ferroviaria, etc.), con el objetivo de refinar la distinción entre actividades y paradas intermedias entre etapas de un mismo viaje.

[1] Los datos de eventos registrados se procesan en un entorno seguro en la infraestructura del operador móvil para generar información agregada y por tanto anonimizada, con el fin de cumplir con lo establecido en la LOPD-GDD.

3. Solución técnica y metodología

3.1 Extracción de los registros de telefonía móvil

El primer subproceso consiste en la extracción y pseudonimización de los registros de telefonía móvil. La pseudonimización de los registros está basada en la utilización de una función hash unidireccional, es decir, una función que permite el cálculo de un identificador anonimizado (similar a un texto aleatorio) a partir del identificador original (habitualmente el IMSI, en el caso de un operador de telefonía) de tal forma que resulta imposible realizar el proceso a la inversa. Se utiliza lo que se conoce como funciones hash perfectas, que por su diseño evitan las colisiones, es decir, evitan que dos identificadores originales diferentes den como resultado un mismo identificador anonimizado. Una vez anonimizados, los registros de telefonía se almacenan en un entorno seguro dentro de la infraestructura del operador móvil, en el que se instala el software necesario para generar los indicadores agregados y anonimizados especificados en la sección 1.2.

3.2 Generación de los indicadores de movilidad

La generación de los indicadores de movilidad se ha llevado a cabo utilizando un software especializado desarrollado a tal efecto. Este software ha sido empleado en más de 80 proyectos en distintos países en los que se han utilizado datos anonimizados de telefonía móvil para la caracterización de la movilidad urbana e interurbana, tanto para clientes públicos (agencias estadísticas, autoridades de transporte, etc.) como privados (empresas concesionarias de autopistas, operadores de autobuses interurbanos, consultoras de transporte, etc.). Entre estos proyectos se incluye el ya mencionado 'Estudio de la Movilidad Interprovincial de Viajeros aplicando la Tecnología Big Data' llevado a cabo por el Ministerio de Transportes, Movilidad y Agenda Urbana en el año 2018.

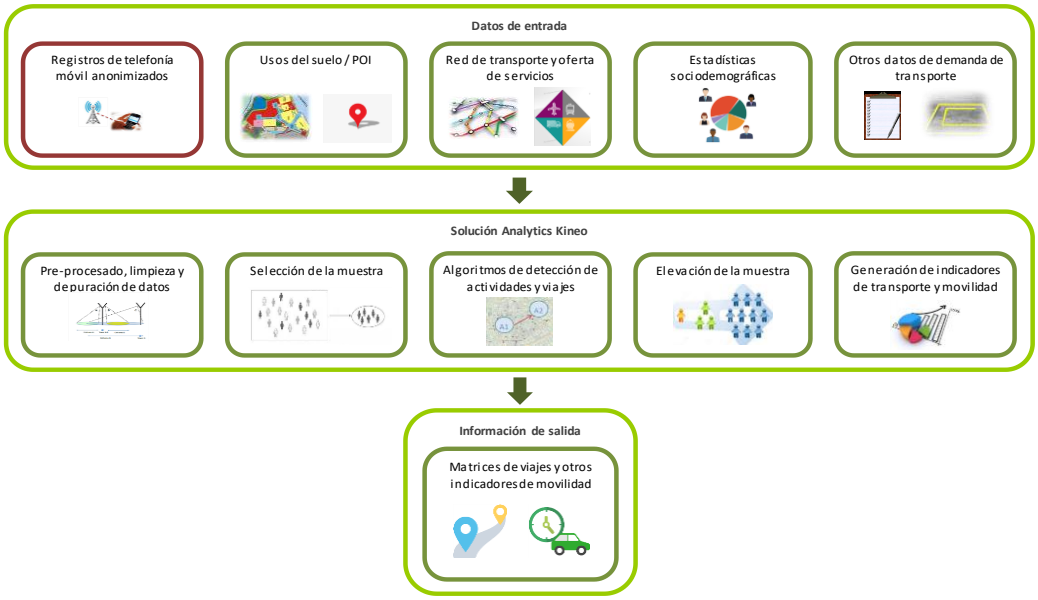

Figura 1 – Esquema de alto nivel de la solución técnica empleada en el proyecto

La Figura 1 muestra un esquema de alto nivel de la solución técnica. El procesado y análisis de los datos consta de los subprocesos principales descritos a continuación. Como ya se ha señalado anteriormente, todos estos procesos se llevan a cabo dentro de la infraestructura del operador móvil, de manera que la información generada y entregada al Ministerio es ya información agregada y anonimizada.

1. **Pre-procesado y limpieza de los datos**. En primer lugar se realiza un pre-procesado de los datos de telefonía para facilitar su gestión, ordenando y agrupando los registros de la forma más conveniente para su posterior análisis. También se lleva a cabo un análisis de integridad de los datos para eliminar posibles errores en los datos del operador móvil. Este proceso resulta esencial para asegurar la calidad de los datos, evitando que posibles errores de origen desvirtúen los resultados que se obtienen con los algoritmos de extracción de patrones de actividad y movilidad[2].

2. **Construcción de la muestra**. Para construir la muestra se realiza una selección de los usuarios válidos para proporcionar información relativa a sus desplazamientos. Dicha selección se realiza de acuerdo con distintos criterios relacionados con su actividad telefónica, de manera que ésta sea suficiente para establecer sus patrones de comportamiento con un nivel de fiabilidad adecuado. La construcción de la muestra supone un compromiso entre cantidad y calidad. Ejercicios de validación llevados a cabo en proyectos anteriores demuestran la importancia de seleccionar una muestra de buena calidad, aún a costa de reducir ligeramente el tamaño muestral, para evitar la inclusión de usuarios que realicen actividades y viajes imposibles de detectar y que puedan por tanto afectar a la calidad de las matrices origen-destino y del resto de indicadores a generar.

3. **Identificación del lugar de residencia habitual y el lugar de pernoctación**. A partir del análisis de los hábitos de comportamiento de los usuarios a lo largo de varias semanas se identifica su lugar de residencia habitual, el cual se utilizará posteriormente en el proceso de elevación muestral. Asimismo, se identifica también el lugar de pernoctación de los usuarios en el día de estudio.

4. **Extracción de actividades y viajes**. Para identificar actividades y viajes, se emplea una combinación de criterios basados en los tiempos de estancia, los itinerarios de los desplazamientos y los patrones de comportamiento a lo largo del periodo de estudio, filtrando las estancias intermedias subordinadas al viaje y realizadas entre etapas del mismo (por ejemplo, una parada intermedia para realizar un transbordo entre autobuses). El resultado de este proceso es la secuencia de actividades y viajes realizados por cada usuario en los días de estudio. La información asociada a cada actividad incluye su localización (a nivel de celda de telefonía móvil), la hora de inicio de la actividad y la hora de finalización. La información asociada a cada viaje incluye origen (localización de la actividad inmediatamente anterior al viaje), destino (localización de la actividad inmediatamente posterior al viaje), hora de inicio del viaje (hora de finalización de la actividad anterior) y hora de finalización (hora de inicio de la actividad siguiente).

5. **Elevación de la muestra**. La expansión de la muestra se realiza tomando como marco muestral la población residente en el país, según los datos del Padrón de Habitantes proporcionados por el INE. Se emplean procedimientos estándar de elevación muestral (similares a los que se emplean, por ejemplo, en una encuesta domiciliaria de movilidad), aplicando factores de expansión por lugar de

[2] A partir del 25 de octubre de 2020 inclusive, se incorpora una mejora en los algoritmos de depuración de la topología de antenas para minimizar los efectos de desactualización o errores en los ficheros de inventario de las antenas.

residencia a nivel de distrito censal, buscando un compromiso entre resolución espacial y homogeneidad de la muestra disponible por unidad censal. Además, se aplica un criterio mínimo de tamaño muestral, descartando aquellos distritos para los que la muestra es inferior al 2% de la población (es decir, para los que el factor de expansión es superior a 50), evitando así que factores de elevación excesivamente altos puedan distorsionar los indicadores de movilidad.

6. **Generación de indicadores**. Finalmente, la información obtenida se agrega con la resolución espacial y temporal requerida y las segmentaciones deseadas para generar las matrices origen-destino y el resto de indicadores de movilidad. La agregación se realiza de tal forma que el tamaño poblacional de los distintos grupos de población analizados garantice la imposibilidad de reidentificar a ningún individuo mediante un hipotético proceso de fusión con otras fuentes de datos, de acuerdo con los requisitos de la LOPD-GDD. Por otro lado, teniendo en cuenta el criterio de limitación de los factores de elevación muestral descrito en el punto 5, cuando para una determinada zona se ha descartado más del 25% del marco muestral, no se proporcionan los indicadores correspondientes a dicha zona.

3.3 Fiabilidad de los resultados y error muestral

Se asume que la muestra de los usuarios de uno de los tres principales operadores en cada zona del territorio y para cada estrato sociodemográfico se aproxima razonablemente bien a una muestra aleatoria de la población residente en dicha zona, salvo por las limitaciones intrínsecas asociadas a la tecnología (ausencia de niños de muy corta edad, que no disponen de teléfono móvil, y menor representación de los ancianos de edad avanzada, algunos de los cuales tampoco son usuarios de líneas móviles). En estas condiciones, y en base a la experiencia de numerosos estudios de movilidad llevados a cabo en los últimos años por numerosas autoridades de transporte a nivel nacional, autonómico y municipal, se considera que la muestra utilizada, de más de 13 millones de líneas móviles, proporcionará un alto nivel de fiabilidad para los indicadores de movilidad a nivel de CCAA y provincia, así como para la movilidad de los municipios de mayor tamaño y las principales relaciones de movilidad entre municipios, suficiente para cumplir con los objetivos del estudio. El error muestral aumentará a medida que se toman resultados más desagregados (por ejemplo, movilidad en municipios pequeños), así como en las relaciones con menor número de viajes.

3.4 Equivalencia entre días de estudio

La comparación de indicadores entre diferentes días de estudio resulta fundamental a la hora de monitorizar la movilidad, especialmente con respecto a aquellos días previos a la crisis del COVID-19 que se toman como referencia de una movilidad habitual. Para garantizar que los indicadores proporcionados tienen en cuenta el mismo marco muestral y son por tanto comparables, se aplican dos criterios adicionales a los descritos en la sección 3.2:

- Se ha relajado el criterio de factor de elevación máximo permitido durante los días de estudio frente a los días de referencia (50 para los días de referencia y 70 para los días de estudio), de forma que se reduce drásticamente la probabilidad de eliminar alguno de los distritos censales considerados en los días de referencia.
- Se han eliminado de los días de estudio los distritos censales que habían sido eliminados en su correspondiente día de referencia.

4. Entregables

Los resultados producidos para cada día de estudio son los siguientes:

1. **Indicadores de movilidad y distribución de la población:**

   1.1 **Matrices de viajes**. Cada elemento se proporciona de acuerdo al siguiente formato:

   - **fecha**: día de estudio en formato "AAAAMMDD"
   - **origen:** identificador de la zona donde se origina el viaje
   - **destino**: identificador de la zona donde finaliza el viaje
   - **periodo**: indicador de la hora en que se origina el viaje en formato "HH". "00" indica la franja horaria comprendida entre las 00:00 y 00:59
   - **distancia**: rango de distancia del viaje con los siguientes valores:
     - 0005-002: viajes de comprendidos entre 500 metros y 2 km
     - 002-005: viajes comprendidos entre 2 y 5 km
     - 005-010: viajes comprendidos entre 5 y 10 km
     - 010-050: viajes comprendidos entre 10 y 20 km
     - 050-100: viajes comprendidos entre 50 y 100 km
     - 100+: viajes de más de 100 km
   - **viajes**: número de viajes
   - **viajes-km**: número de viajeros*kilómetro.

   1.2 **Distribución del número de viajes por persona**, de acuerdo al siguiente formato

   - **fecha**: día de estudio en formato "AAAAMMDD"
   - **zona:** identificador de la zona de pernoctación de los viajeros
   - **número de viajes**: número de viajes realizados: '0', '1', '2' o 'más de 2'
   - **personas**: número de personas

   Estos indicadores se generan a nivel de municipio (o agregación de municipios), así como distintas agregaciones a otros niveles (por ejemplo, a nivel autonómico y provincial) para facilitar la presentación de la información.

2. **Visualización interactiva de datos**, disponible en la página web del Ministerio.