GOBIERNO
DE ESPAÑA

MINISTERIO
DE TRANSPORTES, MOVILIDAD
Y AGENDA URBANA

Secretary of State for Transport, Mobility and Urban Agenda

Analysis of mobility in Spain with Big Data technology during the state of alarm for COVID-19 crisis management.

Methodological report

April 14, 2021

Index

1. Introduction

1.1 Background

The Ministry of Transport, Mobility and Urban Agenda considers it necessary to analyze the changes that are taking place in the mobility of Spaniards during the COVID-19 crisis, with the aim of generating information that will serve both to evaluate the effect of the mobility restriction measures imposed on citizens and for other analyses and studies that will help in the management and subsequent exit from this crisis. Given its capacity to produce quality information with a high level of detail in very short time frames, the Ministry has considered that the optimal solution to respond to this need is the use of solutions based on the analysis of massive data, taking advantage of the experience acquired in the project 'Study of Interprovincial Mobility of Travelers applying Big Data Technology' carried out in 2018, which due to its scope and complexity was a pioneer in this field at an international level. As in the aforementioned study, the present study uses anonymized records from cell phone networks as its main data source. Such records, originally generated for billing or network management purposes, provide large samples with high spatio-temporal resolution of virtually all population segments and have been successfully used in numerous mobility and transport demand studies over the last few years, making them particularly suitable for the purpose of the present study. The study relies on a technical solution and methodology similar to those of the interprovincial mobility study conducted by the Ministry in 2018. Information from cell phone networks has been merged with other data sources to generate origin-destination matrices and other anonymous and aggregated mobility and population presence indicators, ensuring strict compliance with the requirements of Organic Law 3/2018, of December 5, on the Protection of Personal Data and Guarantee of Digital Rights (LOPD-GDD). This document describes the data used in the project, the methodology and data analysis algorithms, and the indicators generated.

1.2 Purpose and scope of the study

The project includes the generation of origin-destination matrices and other mobility indicators. The specifications of the study are detailed below.

Table 1. Specification of the study

Study specifications	
Study population	Resident population in Spain.
Zoning	The indicators are calculated for a specific zoning defined by MITMA ("MITMA zones"). In most cases, MITMA zones correspond to municipalities. Smaller municipalities are aggregated following the same criterion used by INE in the study of the analysis of home-work relationships from cell phone data conducted by INE in 2019. This allows the relevant aggregations to be made in order to also provide indicators at the province level, at the Autonomous Community level and at the national level.
Study days	The mobility and distribution of the population in the territory after the application of Royal Decree 463/2020 of March 14 is analyzed on a daily basis until the end of the state of alarm. The study also includes the two weeks prior to the state of alarm. Once the state of alarm has ended, the weeks following the end of the state of alarm will be analyzed until a stable level of mobility is observed. The study also includes an analysis of a typical week (February 14-20, 2020), in order to evaluate the reduction in mobility with respect to a set of days with usual levels of mobility.
Trips under study	All trips of more than 500 meters with origin and destination within Spain are analyzed.
Indicators	The following indicators are provided: - Origin-destination matrices, segmenting trips: ○ in 1-hour segments according to the starting time of the trip; ○ according to the orthodromic distance between the origin and the destination, distinguishing 6 distance ranges: 0.5-2 km, 2-5 km, 5-10 km, 10-50 km, 50-100 km and over 100 km. For each element of the trip matrix, the total number of traveler-km corresponding to that origin-destination pair is also provided, according to the orthodromic distance between origin and destination. - Distribution of the number of trips per person, distinguishing between those who make no trips, those who make 1 trip, those who make 2 trips, and those who make more than 2 trips.

2. Data sources used

2.1 Anonymized mobile telephony records

The main source of data is anonymized mobile telephony records. The study is based on a data sample of more than 13 million mobile lines provided by a mobile operator, which could be increased throughout the project, as data from more operators become available.

The input data can be classified into two categories:

- **Recorded event data**[1] : anonymized data associated with the connection records of mobile devices to the mobile network. These records include both active and passive events. Active events are made up of what are called CDRs (Call Detail Records), which provide a record each time the device interacts with the network (calls, sending text messages, data sessions). These records are joined by passive event data (periodic update of the device position, changes in coverage areas, etc.), providing a very high temporal granularity. In terms of spatial resolution, location information is available at the telephony cell level, which means a spatial accuracy of tens or hundreds of meters in cities and up to several kilometers in rural areas.
- **Mobile network topology data**: data on the mobile network, including the location of communication towers and the orientation of antennas.

2.2 Land uses

Land use data have also been used to improve the characterization and spatial location of the activities identified from the cell phone data. The land use data come from the Spanish Land Use Information System (SIOSE) and other databases available at the regional level.

2.3 Population data

Data from the Municipal Register of Inhabitants have been used for the sample elevation processes.

2.4 Transport network data

The algorithms used for trip identification also employ information from the transportation network (e.g., airport locations, rail network, etc.) to refine the distinction between activities and intermediate stops between legs of the same trip.

[1] The recorded event data are processed in a secure environment in the mobile operator's infrastructure to generate aggregated and therefore anonymized information, in order to comply with the provisions of the LOPD-GDD.

3. Technical solution and methodology

3.1 Extraction of cell phone records

The first sub-process consists of the extraction and pseudonymization of cell phone records. The pseudonymization of the records is based on the use of a one-way hash function, i.e. a function that allows the calculation of an anonymized identifier (similar to a random text) from the original identifier (usually the IMSI, in the case of a telephone operator) in such a way that it is impossible to carry out the process in reverse. What are known as perfect hash functions are used, which by design avoid collisions, i.e., they prevent two different original identifiers from resulting in the same anonymized identifier. Once anonymized, the telephony records are stored in a secure environment within the mobile operator's infrastructure, where the necessary software is installed to generate the aggregated and anonymized indicators specified in section 1.2.

3.2 Generation of mobility indicators

The generation of mobility indicators has been carried out using specialized software developed for this purpose. This software has been used in more than 80 projects in different countries in which anonymized cell phone data have been used to characterize urban and interurban mobility, both for public (statistical agencies, transport authorities, etc.) and private clients (highway concession companies, intercity bus operators, transport consultants, etc.). These projects include the aforementioned 'Study of Interprovincial Passenger Intercity Mobility applying Big Data Technology' carried out by the Ministry of Transport, Mobility and Urban Agenda in 2018.

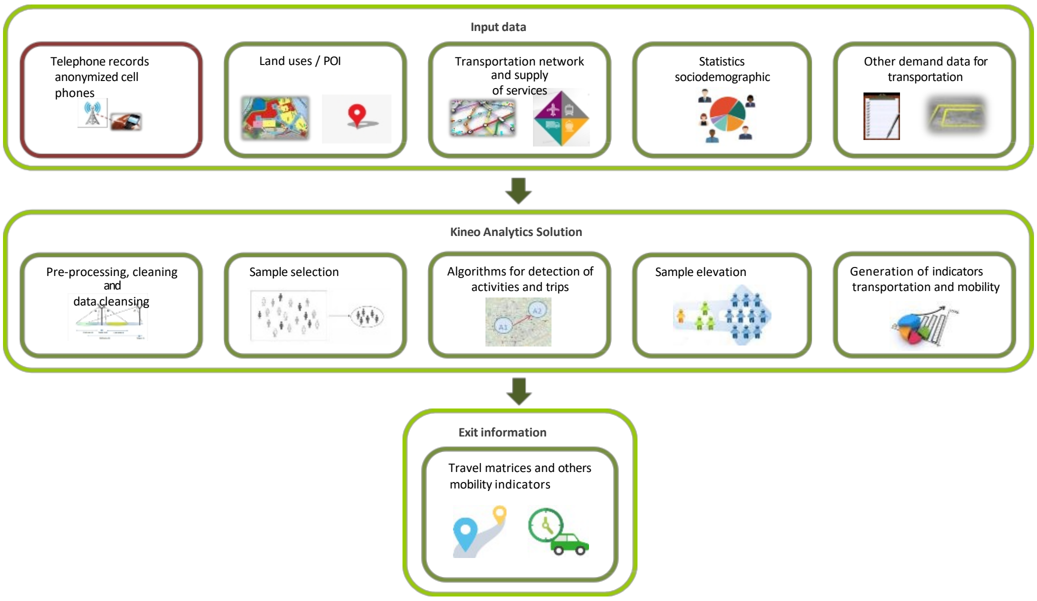
Input data
Telephone records anonymized cell phones
Land uses / POI
Transportation network and supply of services
Statistics sociodemographic
Other demand data for transportation
Kineo Analytics Solution
Pre-processing, cleaning and data cleansing
Sample selection
Algorithms for detection of activities and trips
Sample elevation
Generation of indicators transportation and mobility
Exit information
Travel matrices and others mobility indicators

Figure 1 - High-level schematic of the technical solution used in the project

Figure 1 shows a high-level schematic of the technical solution. Data processing and analysis consists of the main sub-processes described below. As noted above, all these processes are carried out within the mobile operator's infrastructure, so that the information generated and delivered to the Ministry is already aggregated and anonymized information.

1. **Pre-processing and data cleaning**. First of all, telephony data is pre-processed to facilitate its management, sorting and grouping the records in the most convenient way for subsequent analysis. A data integrity analysis is also performed to eliminate possible errors in the mobile operator data. This process is essential to ensure the quality of the data, preventing possible source errors from distorting the results obtained with the activity and mobility pattern extraction algorithms[2] .

2. **Sample construction**. To construct the sample, a selection of valid users is made to provide information related to their trips. This selection is made according to different criteria related to their telephone activity, so that it is sufficient to establish their behavior patterns with an adequate level of reliability. The construction of the sample involves a compromise between quantity and quality. Validation exercises carried out in previous projects demonstrate the importance of selecting a good quality sample, even at the cost of slightly reducing the sample size, in order to avoid the inclusion of users who carry out activities and trips that are impossible to detect and that may therefore affect the quality of the origin-destination matrices and the rest of the indicators to be generated.

3. **Identification of the usual place of residence and the place of overnight stay**. Based on the analysis of users' behavioral habits over several weeks, their usual place of residence is identified, which will be used later in the sample elevation process. Likewise, the users' place of overnight stay on the study day is also identified.

4. **Extraction of activities and trips**. To identify activities and trips, a combination of criteria based on dwell times, trip itineraries and behavior patterns is used throughout the study period, filtering out intermediate stays subordinate to the trip and made between trip stages (e.g., an intermediate stop to transfer between buses). The result of this process is the sequence of activities and trips made by each user on the study days. The information associated with each activity includes its location (at the cell phone cell level), the start time of the activity and the end time. The information associated with each trip includes origin (location of the activity immediately preceding the trip), destination (location of the activity immediately following the trip), trip start time (end time of the previous activity) and end time (start time of the next activity).

5. **Sample expansion**. The expansion of the sample is carried out by taking the resident population of the country as the sample frame, according to the data of the Register of Inhabitants provided by the INE. Standard sample expansion procedures are used (similar to those used, for example, in a household mobility survey), applying expansion factors by place of residence.

[2] From and including October 25, 2020, an improvement in the antenna topology debugging algorithms is incorporated to minimize the effects of out-of-date or errors in the antenna inventory files.

residence at the census district level, seeking a compromise between spatial resolution and homogeneity of the available sample per census unit. In addition, a minimum sample size criterion is applied, discarding those districts for which the sample is less than 2% of the population (i.e., for which the expansion factor is greater than 50), thus avoiding excessively high elevation factors that could distort the mobility indicators.

6. **Generation of indicators**. Finally, the information obtained is aggregated with the required spatial and temporal resolution and the desired segmentations to generate the origin-destination matrices and the rest of the mobility indicators. The aggregation is carried out in such a way that the population size of the different population groups analyzed guarantees the impossibility of re-identifying any individual through a hypothetical process of merging with other data sources, in accordance with the requirements of the LOPD-GDD. On the other hand, taking into account the criterion for limiting the sample elevation factors described in point 5, when for a given area more than 25% of the sample frame has been discarded, the indicators corresponding to that area are not provided.

3.3 Reliability of results and sampling error

It is assumed that the sample of users of one of the three main operators in each area of the territory and for each socio-demographic stratum is reasonably close to a random sample of the resident population in that area, except for the intrinsic limitations associated with the technology (absence of very young children, who do not have a cell phone, and lower representation of the elderly, some of whom are not mobile line users either). Under these conditions, and based on the experience of numerous mobility studies carried out in recent years by numerous transport authorities at national, regional and municipal level, it is considered that the sample used, of more than 13 million mobile lines, will provide a high level of reliability for mobility indicators at the Autonomous Community and provincial level, as well as for the mobility of the largest municipalities and the main mobility relationships between municipalities, sufficient to meet the objectives of the study. The sampling error will increase as more disaggregated results are taken (e.g. mobility in small municipalities), as well as in the relationships with lower number of trips.

3.4 Equivalence between study days

The comparison of indicators between different study days is fundamental when monitoring mobility, especially with respect to those days prior to the COVID-19 crisis that are taken as a reference for usual mobility. To ensure that the indicators provided take into account the same sample frame and are therefore comparable, two additional criteria to those described in section 3.2 are applied:

- The maximum elevation factor criterion allowed during study days versus reference days has been relaxed (50 for reference days and 70 for study days), thus drastically reducing the probability of eliminating any of the census tracts considered on reference days.
- Census districts that had been eliminated on their corresponding reference day have been removed from the study days.

4. Deliverables

The results produced for each study day are as follows:

1. **Mobility and population distribution indicators:**

   1.1 **Trip matrices**. Each element is provided according to the following format:

   - **date**: day of study in format "YYYYYMMDD".
   - **origin:** identifier of the zone from which the trip originates
   - **destination**: identifier of the zone where the trip ends
   - **period**: indicator of the time at which the trip originates in "HH" format. "00" indicates the time zone between 00:00 and 00:59.
   - **distance**: travel distance range with the following values:
     - 0005-002: trips of between 500 meters and 2 km
     - 002-005: travel between 2 and 5 km
     - 005-010: trips between 5 and 10 km
     - 010-050: trips between 10 and 20 km
     - 050-100: trips between 50 and 100 km
     - 100+: trips of more than 100 km
   - **trips**: number of trips
   - **trips-km**: number of passengers*kilometer.

   1.2 **Distribution of the number of trips per person**, according to the following format

   - **date**: day of study in format "YYYYYMMDD".
   - **zone:** identifier of the travelers' overnight stay zone
   - **number of trips**: number of trips made: '0', '1', '2' or 'more than 2'.
   - **persons**: number of persons

   These indicators are generated at the municipality level (or aggregation of municipalities), as well as different aggregations at other levels (for example, at the autonomous community and provincial levels) to facilitate the presentation of the information.

2. **Interactive data visualization**, available on the Ministry's website.