GOBIERNO DE ESPAÑA
MINISTERIO DE TRANSPORTES Y MOVILIDAD SOSTENIBLE

**Secretary of State for Transportation and Sustainable Mobility**

# Study of national passenger mobility using Big Data technology.

## Methodological report

Version 8 - December 11, 2024

# Index

# Revision log

| Version | Date | Modification | Sections affected |
|---|---|---|---|
| 1 | 20/12/2021 | Initial version. | |
| 2 | 26/07/2023 | The "zone_residence" field of the overnight stays indicator becomes the census district of residence, according to INE coding. | 4 |
| 3 | 01/02/2024 | Information regarding the complete studies is included.<br>The information of the deliverables is presented in tables for easy reading. | 1, 2, 3, 4 |
| 4 | 10/07/2024 | MITMS comments are resolved and the description of deliverables is expanded. | All |
| 5 | 12/07/2024 | Revision of content and expansion of the description of deliverables. | All |
| 6 | 13/07/2024 | Correction in the number of zoning zones. | 1.2.1 y 4.1 |
| 7 | 25/09/2024 | Correction of the name of the Secretary of State.<br>Correction travel including on cross-border products.<br>Specification of recurrence only for national users in mandatory mobility.<br>Specification of routes with origin and destination in national territory in route matrix. | Cover page, headlines.<br>1.2.2, 4.3.4, 4.3.6, 4.4.1. |
| 8 | 11/12/2024 | Review of the object and scope of the study.<br>Review of currently used data sources. | 1 y 2 |

# 1. Introduction

## 1.1 Background

In its efforts to improve efficiency in the planning and management of transport infrastructures and services, the Ministry of Transport and Sustainable Mobility (hereinafter MITMS), through the Secretary of State for Infrastructures, Transport and Housing, has among its functions the performance of the so-called "Demand Foresight Studies". These studies are considered essential for assessing the impact of infrastructure, services or regulatory actions on the evolution of the transport system, and constitute information of great value for different types of users, both public and private. One of these studies is the **analysis of internal passenger mobility**.

In the past, the study of passenger mobility in Spain was carried out through the Movilia surveys, which involved a significant economic, technical and human effort. The Movilia 2007 survey is the last survey available for this purpose, and is more than a decade old. In recent years, the proliferation of new sources of geolocated data from personal mobile devices has led to new ways of studying people's mobility and obtaining transportation demand information in a reduced time frame and at a significantly lower cost than traditional methods, eliminating or mitigating some of the main limitations of mobility surveys. Data from the cell phone network are particularly interesting for this purpose, thanks to the possibility of obtaining large samples with a high spatial and temporal resolution of virtually all segments of the population. In order to take advantage of these opportunities, in 2017 the Secretary of State for Infrastructure, Transport and Housing launched the **"Study of interprovincial passenger mobility applying big data technology "**[1] , a pioneering project that used the records generated by mobile terminals when connected to telephone networks as the main source of data to study passenger mobility on a national scale. Building on the experience gained in the previous study, in 2020 MITMS carried out the study **"Analysis of mobility in Spain with big data technology during the State of Alarm for the management of the COVID-19 crisis "**[2] , in which cell phone data were used to monitor mobility on a daily basis and support different decision-making processes related to the management of the health emergency caused by the COVID-19 pandemic.

With the precedent of these two projects, MITMS has launched the present project called **"Study of Travelers' Mobility at national level applying Big Data technology"** which will give continuity to the previous ones, broadening its scope and deepening in this line of work.

[1] https://observatoriotransporte.mitma.es/estudio-experimental

[2] https://www.mitma.gob.es/ministerio/covid-19/evolucion-movilidad-big-data

## 1.2 Purpose and scope of the study

The purpose of the project is to analyze passenger mobility at the national

level. Three types of studies are generated:

1. Basic daily study
2. Complete study (monthly)
3. Study of road routes

The scope and content of each type of study is detailed below.

### 1.2.1 Basic daily study

Giving continuity to the daily study carried out in the project "Analysis of mobility in Spain with *big data* technology during the State of Alarm for COVID-19 crisis management", a **basic daily study** is carried out, according to the following specifications:

- **Trips under study**: trips of more than 500 meters with origin and/or destination in the national territory. Trips abroad whose route in national territory is less than 500 meters are not included.
- **Study population**: residents in Spain.
- **Study period**: the duration of the study period is a minimum of 18 months, which has been extended for an additional 18 months. Mobility is analyzed from January 1, 2022 onwards.
- **Zoning**: the base zoning of the study is composed of a set of zones that constitute aggregations of census districts for the territory of Spain and NUTS-3 zones for France and Portugal. This zoning presents a total of 3,909 zones. For trips with origin or destination outside the study zoning, an external zone covering the rest of the world has been defined. Starting from this base zoning, the results are also presented aggregated in two additional zonings to facilitate the exploitation of the information: one at the municipal level with a total of 2,735 zones and another at the level of large urban areas (GAUs) with a total of 2,203 zones.
- **Time resolution**: trips are segmented by time slots of 1 hour duration. As a criterion for assigning the time of the trip, the start time of the trip is considered in the case of trips from Spain. For trips originating outside the national territory, the time of entry into Spain is considered as the time of the trip.
- **Trip distance**: is determined as the orthodromic distance between the origin and destination of the trip and the following intervals are used: 500 meters-2 km, 2-10 km, 10-50 km and >50 km. For trips to/from abroad, the distance corresponds to the distance between the origin/destination of the trip and the point of departure/entry into the national territory.
- **Segmentation by trip purpose**: for each trip, the type of activity at origin and destination is determined, distinguishing between home, work/study, other frequent activities, and sporadic activities. In addition, a supplementary indicator is incorporated for

indicate whether the activity could potentially be related to the purpose of the study. It is important to note that the category "work/study" contains mainly full-time work/study at fixed locations. Part of the forced mobility may also be collected in the group "other frequent activities", for those cases of work/study activities that have not been classified in the previous group because they have a pattern also compatible with other types of frequent activities (e.g. frequent leisure activities).

- **Segmentation by socio-demographic** profile[3]: trips are segmented according to the following traveler characteristics:
  - Place of residence at the provincial level
  - Age, according to the following age bands: 0-24, 25-44, 45-64 and ≥ 65.
  - Sex, differentiating between "male" and "female".
  - Average household income level, according to the following bands in thousands of euros: < 10, between 10 and 15 and >15.

The following main products are generated for each study day:

- **Basic trip matrix**, which provides the number of trips and trip-kilometers for each combination of origin-destination pair, time of day, trip distance interval, activity at origin, activity at destination and socio-demographic profile of travelers.
- **Trips per person**, which provides the number of people who make a given number of trips per day by area of overnight stay, segmented by age and sex.
- **Overnight stays**, which allow us to obtain the number of people who spend the night in each area by identifying their place of residence.

A detailed description of each of the products/files generated is presented in section 4 of this document.

### 1.2.2 Complete study

A **complete study** is carried out **on a monthly basis**, with a higher level of detail than the basic daily study, according to the following specifications:

- **Trips under study**: trips of more than 500 meters with origin and/or destination in the national territory, excluding trips made by professional drivers (drivers of goods vehicles, delivery drivers, cab drivers, airline crews, etc.). Unlike the basic study, this study includes trips abroad with a distance of less than 500 meters on national territory.
- **Study population**: residents in Spain and visitors residing abroad.

[3] Some of the sociodemographic variables may not be provided in some cases due to privacy criteria; in such cases, the value "NA" will be displayed.

- **Study period**: the duration of the study period is the same as the duration of the basic studies. For each month of study, a typical week and a set of up to 10 singular days are analyzed.
- **Zoning**: the same zoning is used as in the basic study.
- **Identification of stages and mode of transport**: trips whose orthodromic distance between origin and destination is greater than 5 km are segmented into their component stages and the mode of transport of each stage is identified by distinguishing the following modes: (i) road, differentiating between trips in private vehicles and bus in the case of trips with origin and destination in different Autonomous Communities; (ii) rail, distinguishing between high speed and conventional rail; (iii) air transport; and (iv) maritime transport.
- **Time resolution**: trips and stages are segmented by time slots of 1 hour duration. As a criterion for assigning the time of the trip, the start time of the trip/stage is considered. For trips originating outside the national territory, the time of entry into Spain is considered as the time of the trip,
- **Trip distance**: the distance of each trip/stage is determined as the distance between the origin and destination along the route followed over the transport infrastructure and the following intervals are used: 500 meters-2 km, 2-5 km, 5-10 km, 10-50 km and >50 km. For trips to/from abroad, the distance corresponds to the distance between the origin/destination of the trip and the point of departure/entry into the national territory.
- **Travel time**: the time between the beginning and end of each stage and of the trip as a whole is determined.
- **Segmentation by trip purpose**: for each trip, the type of activity at origin and destination is determined, distinguishing between home, work/study, other frequent activities, and sporadic activities.
- **Segmentation by** sociodemographic **profile4**: trips/stages of Spanish residents are segmented according to the following traveler characteristics:
    - Place of residence at the provincial level.
    - Age, according to the following age bands: 0-24, 25-44, 45-64 and ≥ 65.
    - Sex, differentiating between "male" and "female".
    - Average household income level, according to the following bands in thousands of euros: < 10, between 10 and 15 and >15.
    - Country of residence.

The following products will be generated for each complete monthly study:

- **Stage matrix**, which provides trip data disaggregated by stage, indicating the mode of transport for each stage. The stage matrix is univocally linked to the trip and tour matrices so that the stages and modes that make up each set of trips or tours can be identified.

[4] Some of the sociodemographic variables may not be provided in some cases due to privacy criteria; in such cases, the value "NA" will be displayed.

- **Complete travel matrix**, which in addition to the information contained in the basic travel matrix, includes segmentation by mode of transport, number of connections or transfers, and travel time.
- **Tour matrix**, which identifies the travel sequences that make up tours, segmented according to the modes of transportation used in such travel sequences.
- **Cross-border mobility matrix**, which provides information on travel by land border and road from/to France or Portugal indicating the national border area crossed.
- **Matrices of the area of influence of the main public transport terminals**, which provide information on the trips that make use of these terminals, characterizing the modes of access and dispersion.
- **Recurrent forced mobility matrix**, which provides information on the frequency of trips by domestic users based on the origin-destination pair.

### 1.2.3. Road route studies

A study is carried out of the **routes followed in road travel and the total number of travelers using each route**. The geometry of the roads is that of Open Street Maps (OSM), and information is provided for the sections that are within the study area. The study area is the vicinity of the MITMS roads of interest.

Results are generated according to the following specifications:

- **Trips under study**: inter-municipal road trips with origin and/or destination in the national territory.
- **Study population**: residents in Spain and visitors residing abroad.
- **Study period**: a typical winter week and a typical summer week, to be defined jointly with MITMS, are analyzed.
- **Zoning**: the route data are segmented according to their origins and destinations, according to the zoning used in the basic study and the complete study.
- **Segmentation by route used and travel distance**: a distinction is made between the different road routes used, according to the georeferenced information of the road network.

The following products are generated:

- **Section geometry**, which provides the geometry of each section of the road network.
- **Route-section relationship,** which provides the set of sections that make up each route.
- **Section information,** which provides the number of travelers using each section of the road network.
- **Route matrices,** which provide the number of travelers using a specific route in national territory.

A detailed description of each of t h e products/files generated is presented in section 4 of this document.

## 1.3 Glossary of terms

| | |
|---|---|
| **Stay** | A user's permanence during a period of time in a given location. Consecutive records in the same cell, or set of adjacent telephone cells, are analyzed for spatial resolution. When the mobile device connects to different nearby antennas without the user having moved, a "signal hop" occurs, which is filtered by the solution. |
| **Displacement** | Records not associated with stays. |
| **Activity** | An individual's stay in a given place that does not meet the stop criteria. The activities are originated by an interaction or set of interactions with the environment (including interactions with other people, with a certain service, etc.) that motivate the individual to travel there. |
| **Stage** | Travel between the locations of two consecutive stays by the same mode of transport. The use of collective modes of transport generally involves an access/return stage to/from the corresponding terminal or stop. Within a trip in the same mode of transport, transfers may occur between different means or vehicles, which are recorded as consecutive stages. In this project, this is the case for commuter train trips (one stage) that continue on a long-distance train (another stage), and also for airplane trips with a stopover (the first flight is one stage, and the second flight is another). On the other hand, road trips by private vehicle are considered single-stage. That is, the road trip is not separated according to intermediate stops (e.g., at a gas station for refueling). |
| **Tour** | A sequence of activities and trips within a single day that begins and ends in the same location, in this project the place of overnight stay according to the zoning of interest. A tour is composed of two or more trips (e.g., a sequence of home-work-shopping-home trips). |
| **Travel** | Travel between the locations of two consecutive activities. A trip has a main purpose determined by the activity at origin, the activity at destination or a combination of both. A trip may be single-modal or multimodal and may consist of one or more stages. |

# 2. Data sources used

Solution input data can be classified into 5 main categories:

1. Data from the cell phone network and geolocated mobile devices.
2. Sociodemographic statistics.
3. Land use data and points of interest.
4. Data on the transportation network and the supply of transportation services.
5. Transportation demand data.

## 2.1 Anonymized telephony records cell phones

The main source of data is the anonymized mobile telephony records. The study is based on a data sample of more than 13 million mobile lines provided by Orange Spain.

The input data can be classified into three categories:

- **Recorded event** [data5]: anonymized data associated with the connection records of mobile devices to the mobile telephony network. These logs include both passive and active events. Passive events are made up of what are called **network probe data**, which provide a very high temporal granularity, generating a record in Spain every time the device changes antenna. These records are joined by active events called **CDRs** (Call Detail Records), which provide a record every time the device interacts with mobile networks (calls, sending text messages, data sessions). 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.
- **Sociodemographic data**: data on the gender and age ranges of mobile line holders.

*It is important to point out at this point that the data used in this project are substantially different from the data used in previous studies, presenting an improvement in quality in terms of temporal granularity. Annex I discusses the aspects to be taken into consideration when comparing the results of this project with previous studies, pointing out the influence of the change in the data source for this project.*

[5] 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.

## 2.2 Statistics sociodemographic

Data from the INE, specifically, the Municipal Register of Inhabitants and the Continuous Population Statistics for residents in Spain and data from the FRONTUR surveys for foreign visitors, were used for the sample elevation processes.

Income data have been obtained from the Atlas of Household Income Distribution, also from the INE.

## 2.3 Land uses and points of interest

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. Additionally, data on the location of educational centers from the POIS (Points of Interest) of the BTN product (National Topographic Base) of the National Geographic Institute have been used to characterize the possible activities associated with the study.

## 2.4 Transport network data and supply and demand of services

The technical solution integrates data from the transport network and information on the supply of services that are essential for the determination of the transport mode and travel route. The data sources used for the different transport modes are described below.

### 2.4.1 Road transportation

The data sources used to obtain road transport data can be divided into two main groups: (i) information associated with the national road network and (ii) information associated with bus transport services between autonomous communities.

In relation to the road network information, this is obtained through the use of the OpenStreetMap APIs (www.openstreetmap.org), obtaining information on the different route options to travel by private vehicle from point 'A' to point 'B' of the territory. The information is obtained by considering different options for route calculation (e.g. no restrictions, avoiding tolls, avoiding main roads, etc.), so as to obtain a sufficiently wide route tree.

On the other hand, for bus travel information, data is available on transportation supply associated with VACs (Passengers, Authorization and Concession) contract dispatches, which contains quarterly information on intercity bus routes associated with state concessions. These data are non-public and have been provided by MITMS for use as part of this project. The data are available in Excel format and contain, for each of the VACs, the ordered sequence of stops traveled by buses on each dispatch. For all stops there is a unique identifier code, the name of the stop, its address and the municipality in which it is located. In addition, for a high proportion of the stops, a unique identifier code is available.

The longitude and latitude are also available, which makes it possible to locate them precisely in the territory.

### 2.4.2 Rail transport

The data sources used to obtain the rail transport supply data are as follows:

- **Information on the location of stations and regular services offered by RENFE AVLD** (medium distance, long distance and AVE), **RENFE Cercanías, RENFE FEVE, FGC, FGV, Euskotren, OUIGO and Iryo**. Monthly information is available on departures and arrivals of rail services in national territory broken down by timetable, as well as the location of the stations. Although these data are public for some periods and can be downloaded from the National Access Point (NAP) of transport data, the project has access to historical data from the NAP (provided by MITMS), so that the data used are always those closest to the month of study. The RENFE AVLD, Cercanías, FEVE, FGC, FGC, FGV, Euskotren and OUIGO regular rail service files are in GTFS (General Transit Feed Information) format and provide service information through several CSV files. In particular, there is a file with georeferenced information of the railroad stations and a file describing the ordered sequence of stops that each railway dispatch runs.
- **Rail network infrastructure data**, containing geo-referenced information on rail sections and dependencies. These data are not public and are provided quarterly by ADIF specifically for the execution of the project. The rail infrastructure data are provided in ESRI GeoDataBase format and contain the geometry of the rail sections and dependencies, as well as the relationships between geometries and identifier codes.
- **Rail traffic information,** which specifies the ordered sequence of sections traveled between each pair of origin and destination stations for commuter, long and medium distance services. As in the case of rail infrastructure data, this data is provided by ADIF on a quarterly basis for exclusive use in this project, as it is not public. The rail traffic file is in Excel format and describes, for a given date, the ordered sequence of rail sections that a train travels between a station of origin and a station of destination. The same Excel file also provides the relationship between sections and railway dependencies.

### 2.4.3 Air transportation

The data source used to obtain information on air transport supply is the information provided by AENA specifically for the execution of this project, which consists of data on passengers transported on flights with origin and/or destination at airports in Spanish territory. These files contain information on the pair of airports connected by a specific service, the actual date and time of departure of the service, the type of service (scheduled vs. non-scheduled) and whether the service corresponds to a departure or arrival of a flight. With the information of all the pairs present

supply can be generated on the demand side, such as possible connections between airports in the national territory.

### 2.4.4 Maritime transportation

The offer used in the generation of indicators for the years 2022 and 2023 was a list of ports and connections developed in collaboration with the MITMS team.

The data source used to obtain the shipping supply for the generation of the results obtained as of January 2024 is the Hermes database together with the following list of additional ports:

- Port of Tangier
- Port of Corralejo (Fuerteventura)
- Port of Playa Blanca (Lanzarote)
- Puerto de las Nieves (Gran Canaria)
- Port of Playa de Santiago (La Gomera)
- Puerto de la Quema (La Gomera)
- Port of Sant Antoni de Portmany (Ibiza)
- Ciutadella Port (Menorca)
- Port of Denia (peninsula)
- Tarajal Port (Fuerteventura)
- Port of Morro Jable (Fuerteventura)

In order to generate the offer between these ports, it has been decided to use only those inter-island connections (conceiving the peninsula as an island).

## 2.5 Demand data from transport

Finally, Nommon's analytics platform integrates information from other transportation demand data sources that is used to adjust modal split results estimated from mobile data for trips longer than 5 km.

Additionally, demand and ticketing data have been used to validate the transport supply data and, if necessary, complete the supply to be aligned with demand.

### 2.5.1 Road transportation

For intercity bus transport, we have used the transport demand data associated with VAC contract shipments used as a source of supply, which consists of quarterly demand data.

These quarterly demand data for VAC contracts are transformed into daily bus demand data using (i) INE data on the number of passengers transported per month.

in intercity buses to obtain the monthly demand from the quarterly one and (ii) the data from the basic studies to find the daily demand taking into account the monthly demand.

For road transport by private vehicle, no demand data (gauges) have been used.

### 2.5.2 Rail transport

Rail demand data is used to calibrate the adjustment of data obtained with mobile telephony in rail services without commercial supply.

The demand for services without commercial offer has been obtained as a subset of the demand collected in the PSO (Public Service Obligation) data provided by Renfe for commuter and long distance services. These data are daily and are provided on a monthly basis.

In this case, unlike in the case of air and bus transport, demand cannot be used to correct for supply, since demand data comprise only a subset of the services offered.

### 2.5.3 Air transportation

The source of data used to obtain information on air transport demand is the information provided by AENA on a monthly basis, which contains daily information on domestic flights, indicating for each of these operations whether it is an arrival/departure operation, the airport of origin/destination of the trip, the time at which the operation takes place and the number of passengers.

# 3. Technical solution and methodology

## 3.1 Extraction of telephone records cell phones

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.

## 3.2 Generation of the indicators of mobility

The generation of mobility indicators has been carried out using specialized software developed for this purpose. This software has been used in more than 150 projects in different countries in which anonymized cell phone data have been used for the characterization of 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 Mobility applying Big Data Technology" carried out for the Ministry of Transport, Mobility and Urban Agenda in 2018 and the "Analysis of mobility in Spain with big data technology during the State of Alarm for COVID 19 crisis management".

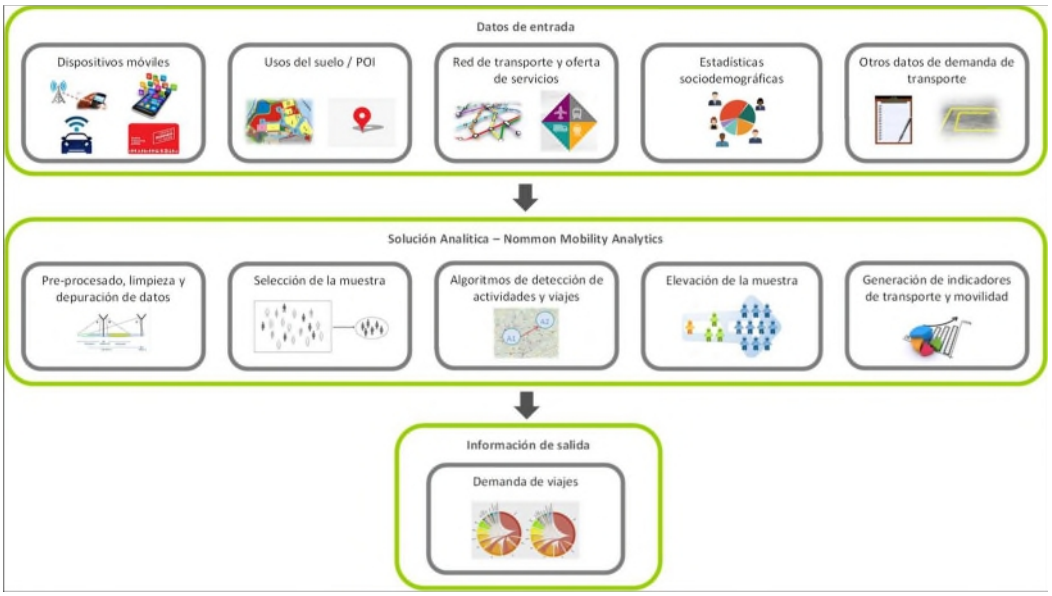


**Figure 1 - High-level schematic of the technical solution used in the project**

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, a pre-processing of the telephony data is performed to facilitate its management, sorting and grouping the records in the most convenient way for further 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. **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 frequent activities and place of overnight stay**. Based on the analysis of users' behavioral habits over several weeks, frequent activities performed by the user are identified, such as their usual place of residence (which will be used later in the sample elevation process), their place of work/study, other frequent activities or sporadic activities. In addition, in order to differentiate between work activities and those associated with study, patterns associated with the proximity of the activity to educational centers, the time of the activity and the socio-demographic profile of the user were analyzed in order to assess the potential for the activity to be of an educational nature. Finally, we also identified the activity associated with the users' place of overnight stay on the study day.

4. **Extraction of activities and trips**. To identify activities and trips, a combination of criteria based on dwell times, travel itineraries and behavioral patterns throughout the study period is used, filtering out intermediate stays subordinated to the trip and made between trip stages (for example, an intermediate stop to transfer between buses). For the identification of transport mode and route, *map matching* techniques are used that take into consideration criteria based on travel time, speed, topology of the transport network and compatibility with the supply of transport services, making it possible to obtain a similarity metric for each of the transport alternatives. Each trip is assigned *a priori* the transport alternative that maximizes the probability of occurrence of the observed records (e.g., a set of records along the route).

The result of this process is the sequence of activities and trips made by each user on the study days, which will correspond to a stage on high-speed rail.) 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 type of activity (home, work/study, etc.), the start time of the activity and the end time. The information associated with each trip includes origin (location of the activity immediately prior to the trip), destination (location of the activity immediately after the trip), trip start time (end time of the previous activity), end time (start time of the next activity) and the sequence of stages that make up the trip, detailing the mode and route used.

5. **Raising the sample**. Once the activity and travel diaries with all their attributes have been constructed, it is necessary to increase the sample of users to the total population under study. Two different cases can be distinguished: the expansion of the sample of residents in Spain and the expansion of the sample of non-residents. In the case of residents in Spain, the expansion of the sample is carried out by taking the population resident in the country as the sampling frame, according to data from the Register of Inhabitants provided by the INE. Standard sampling procedures are used (similar to those used, for example, in a household mobility survey), applying expansion factors stratified by place of residence at the census district, age and gender levels. In addition, a minimum sample size criterion is applied, discarding those strata for which the sample is less than 1% of the population (i.e., for which the expansion factor is greater than 100), thus avoiding excessively high elevation factors that could distort the mobility indicators. In the case of non-residents, the expansion of the sample is carried out using data from the Border Tourism Movement Statistics (FRONTUR) prepared by the INE, applying expansion factors stratified by country of residence and length of stay.

6. **Adjustment with true baseline demand data**. The results obtained from the mobile telephony data and their fusion with other data sources are adjusted using certain reference demand data for air, train and bus. The solution used is based on an adjustment at a disaggregated level, evaluating the degree of goodness of fit in the estimation of the modes of transport for each trip based on the techniques described in section 4 and adjusting the results taking into account this degree of goodness of fit.

7. **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 20% of the sample frame has been discarded, the indicators corresponding to that area are not provided. Additionally, for each product, additional privacy criteria can be applied (see detail in section 4.Deliverables).

## 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 approximates reasonably well 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). For this reason, an adjustment is made using algorithms that allow us to refine the information on the age and gender of the users based on their behavioral patterns. For each of the users in the sample, metrics related to the types of activity performed (for example, if they have a work/study activity), the profile and distance of trips per hour of the day, etc. are obtained. 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 levels, 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 AC and province 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. Similarly, it is important to note that the results of mobility at the national level are more reliable than the results of trips with origin or destination abroad, due to the lower information available in these cases (much smaller number (or nonexistent for some users) of network events outside the national territory and a much lower spatial precision).

# 4. Deliverables

The deliverables consist of zoning files, origin-destination matrices and summary tables for each study (basic, complete and routes), and are described in detail below:

## 4.1 Zoning

Most of the deliverables of this project refer to the origin and destination zones of the trips, and/or the place of residence of the users. In order to know to which specific zones the identifiers in these columns refer, zoning files in shapefile format are provided.

In the project, apart from the main study zoning (3,909 districts and aggregations thereof), a set of relevant aggregations have been defined in order to facilitate the exploitation of the results. Specifically, aggregations have been made at the municipal level (or aggregations of municipalities) and at the level of large urban areas (GAUs). The following zoning files have been generated in digital format:

- **"zonificacion_distritos.shp"**, zoning of greater detail composed of districts or aggregations of districts.
- **"zonificacion_municipios.shp"**, aggregation of detailed zoning at the municipal level or aggregations of municipalities.
- **"zonificacion_GAU.shp"**, aggregation of detailed zoning at the level of large urban areas and municipalities.

Associated with these files are the zone name and population files:

- **"relacion_ine_zonificacionMitma.csv",** file that relates the INE codification with the different codifications used in the different zonifications of the project**.**
- **"poblacion.csv"**, a file with population information by census district and its relationship with municipality and province according to INE codification.
- **"poblacion_[zona].csv"**, a file for each zoning ("distritos", "municipios" and "GAU") that provides for each zone the resident population.
- **"names_[zone].csv"**, a file for each zoning ("districts", "municipalities" and "GAU") that provides for each zone its name.

## 4.2 Studies basic

### 4.2.1 Travel matrices

The trip matrices provide data on the number of trips made by users between each origin-destination pair and according to a series of population and trip characteristics. Specifically, they are files in "CSV" format with the daily mobility information of residents in Spain, for trips of more than 0.5 km. The trip and trip-kilometer metrics are segmented by period, origin, destination, distance, activity and socio-demographic indicators. Different files are generated according to the different zonings.

The file nomenclature is "YYYYYMMDD_Trips_[zone].csv.gz", where "YYYYMMDD" is the year, month and day of study and "[zone]" the type of zoning used, which can take the values "districts", "municipalities" and "GAU". The travel matrix presents the following fields:

| Fields | Description |
|---|---|
| date | Date of study in YYYYMMDD format. |
| period | Time period in 24 time slots. 00' means mobility with origin between 00:00:00:00 and 00:59:59. For trips originating outside the national territory, the time of the trip is the time of entry into Spain. |
| origin | Origin zone identifier. For trips originating outside the study zoning, the "external" origin is defined. |
| destination | Destination zone identifier. For trips with destination outside the study zoning, the destination "external" is defined. |
| distance | Orthodromic distance between trip origin and destination in ranges: 0.5 km-2 km, 2-10 km, 10-50 km and >50 km. For trips with origin or destination outside the study zoning, the value of the distance traveled on national territory is provided |
| source_activity | Type of activity carried out in the area of origin of the trip, differentiating between:<br>- "home"<br>- "work_study"<br>- "frequent"<br>- "not_frequent". |
| target_activity | Type of activity carried out in the destination area of the trip, differentiating between:<br>- "home"<br>- "work_study"<br>- "frequent"<br>- "not_frequent". |
| possible_origin_study | Takes the value 'yes' if the activity in origin can refer to an educational activity. It takes the value 'no' otherwise. |
| possible_destination_study | Takes the value 'yes' if the target activity can refer to an educational activity. It takes the value 'no' otherwise. |
| residence | Place of residence at the provincial level, according to INE coding. |
| income | Income level (in thousands of euros) of travelers in the following ranges:<br>- <10: less than 10,000 euros.<br>- 10-15: between 10,000 and 15,000 euros.<br>- >15: more than 15,000 euros. |
| age | Age range of travelers:<br>- 0-24: 0-24 years<br>- 25-44: 25-44 years old<br>- 45-64: 45-64 years old<br>- >65: > 65 years old<br>Take the value "NA" when the information cannot be provided for privacy reasons. |
| sex | Take the value "NA" when the information cannot be provided for privacy reasons. |
| travel | Number of expanded trips. |
| travel_km | Product of 'trips' by the orthodromic distance of the trips. For trips with origin or destination outside the study zoning, only the distance within the national territory is counted. |

### 4.2.2 Trips per person

This product indicates, for each area, how many people have stayed overnight in the area each day, indicating their age, sex and number of trips made.

These are files in "CSV" format with information on the number of trips made per person according to date, zone, sex and age. Different files are generated according to the different zonings.

The file nomenclature is "YYYYYMMDD_Persons_dia_[zone].csv" where "YYYYMMDD" is the year, month and day of study and "[zone]" the type of zoning used, which can take the values "districts", "municipalities" and "GAU". The file contains the following fields:

| Fields | Description |
|---|---|
| date | Date of study in YYYYMMDD format |
| overnight_zone | Overnight area identifier |
| age | Age range of travelers:<br>- 0-24: 0-24 years<br>- 25-44: 25-44 years old<br>- 45-64: 45-64 years old<br>- >65: > 65 years old<br>Take the value "NA" when the information cannot be provided for privacy reasons. |
| sex | Gender of travelers: 'male' and 'female'.<br>Take the value "NA" when the information cannot be provided for privacy reasons. |
| number of trips | Number of trips made by individuals in the following groups:<br>- '0'<br>- '1'<br>- '2'<br>- '2+' |
| people | Number of people. |

### 4.2.3 Overnight stays

Overnight stay files, unlike per person travel files, indicate where people have spent the night each day according to their place of residence.

These are files in "CSV" format with information on the number of people staying overnight in each zone according to date and zone of residence. Different files are generated according to the different zonings.

The file nomenclature is "YYYYYMMDD_[zone].csv.gz" where "YYYYMMDD" is the year, month and day of study and "[zone]" the type of zoning used, which can take the values "districts", "municipalities" and "GAU". The file contains the following fields:

| Fields | Description |
|---|---|
| date | Date of study in YYYYMMDD format. |
| residence_zone | Identifier of the area of residence. |
| overnight_zone | Overnight stay area identifier. |
| people | Number of people. |

### 4.2.4 Discarded areas

These files, in "CSV" format, contain the information of the zones, according to the zoning of "districts", discarded for each date due to lack of sample on the study day.

The file is named "YYYYYMMDD_districts_discarded.txt" where "YYYYMMDD" is the year, month and day of study. The file presents the following fields:

| Fields | Description |
|---|---|
| date | Date of study in YYYYMMDD format. |
| zone | Zone identifier. |

## 4.3 Studies complete

### 4.3.1 Stage matrices

Trips can be divided into stages according to the mode of transportation used in each one. For example, a trip from Guadalajara by car to Adolfo Suárez Madrid Barajas airport, and from there by plane to Tenerife-Norte, and then by car to La Laguna, has 3 stages:

- Stage 1: origin Guadalajara, destination Adolfo Suárez Madrid Barajas airport, car mode.
- Stage 2: origin Adolfo Suárez Madrid Barajas airport, destination Tenerife-North airport, airplane mode.
- Stage 3: Tenerife-North airport origin, destination La Laguna, car mode.

These are files in "CSV" format with the information of trip stages of more than 0.5 km segmented by mode and means of transport (in the case of trip stages of more than 5 km), as well as by temporal, spatial and socio-demographic variables. Stage and stage-kilometer metrics are provided.

Different files are generated according to the different zonings. The nomenclature of the files is "YYYYYMMDD_Stages_[zone].csv.gz", where "YYYYMMDD" is the year, month and day of study and "[zone]" the type of zoning used, which can take the values "districts", "municipalities" and "GAU". The stage matrix presents the following fields:

| Fields | Description |
|---|---|
| date | Date of study in YYYYMMDD format. |
| day_type | Code that identifies the day of the week and whether it is a day that is part of a standard week of complete studies ("type" day) or whether it is a day that is studied because its mobility presents special patterns ("special" day), such as holidays, eve of long weekends, etc.<br>It can take the values "LT", "MT", "XT", "JT", "VT", "ST", "DT" (Monday Type, Tuesday Type, Wednesday Type, etc.) for standard days or "LE", "ME", "XE", "JE", "VE", "SE", "DE" (Monday Special, Tuesday Special, Wednesday Special, etc.) for special days. |
| ID | Unique stage identifier for each combination of origin zone, destination zone, time period, mode, and medium.<br>This code is used to relate the complete trip matrices and the stages. |
| period | Time of origin of the stage, in "HH" format. |
| origin | Identifier of the area of origin. |
| destination | Identifier of the destination zone. |
| mode | Mode of transport used, differentiating between "C" (road mode), "F" (rail mode), "A" (air mode) and "M" (sea mode).<br>The "mode" field can take the value "NA" for trips of less than 5 km that do not have origin and destination within an urban area to be studied in detail by surveys. |
| medium | Means of transport used:<br>- If the value of the "mode" field is "A" or "M", the value of this field will be "A" or "M", respectively.<br>- If the value of the "mode" field is "F", the value of this field may be "FC" (conventional rail) or "AV" (high-speed rail). |

| Fields | Description |
|---|---|
| | The "conventional" type of service includes Intercity, Media Distancia, Regional, Regional Express and Euromed services; the "high-speed" type of service includes ALVIA, AVANT, AVCity, AVE, AVE-INT, AVE-TGV, AVLO and the services of the Ouigo and Iryo companies.<br>- If the value of the "mode" field is "C", the value of this field may be "PR" (private vehicle) or "AU" (bus). The bus mode is only distinguished for trips between autonomous communities.<br>The value of the "medium" field is "NA" for stages where the "mode" field is also equal to "NA". |
| distance | Distance between origin and destination measured in km along the route followed over the transport infrastructure. In case the mode is "NA", the distance corresponds to the straight line distance between origin and destination.<br>The "distance" field can take the following values:<br>- 0-0.5: trips of less than 0.5 km.<br>- 0.5-2: trips between 0.5 km and 2 km.<br>- 2-5: trips between 2 km and 5 km.<br>- 5-10: trips between 5 km and 10 km.<br>- 10-50: trips between 10 km and 50 km.<br>- >50: trips of more than 50 km. |
| duration | Travel time of the stage.<br>The "duration" field can take the following values:<br>- 0-1: trips of less than 1 hour.<br>- 1-2: trips between 1 hour and 2 hours.<br>- 2-3: 2 to 3 hour trips.<br>- 3-4: trips between 3 hours and 4 hours.<br>- >4: trips of more than 4 hours. |
| country | Identifier code of the country or group of countries where the traveler resides. The "country" field can take the following values:<br>- ES (Spain)<br>- DE (Germany)<br>- BE (Belgium)<br>- FR (France)<br>- IE (Ireland)<br>- IT (Italy)<br>- NL (Netherlands)<br>- PNO (Nordic Countries)<br>- PT (Portugal)<br>- GB (United Kingdom)<br>- REU (Rest of Europe)<br>- RMU (Rest of the World). |
| residence | Place of residence at the provincial level, according to INE coding. |
| income | Income level (in thousands of euros) of travelers in the following ranges:<br>- <10: less than 10,000 euros<br>- 10-15: between 10,000 and 15,000 euros<br>- >15: more than 15,000 euros<br>It takes the value 'NA' for foreign resident travelers. |
| age | Age range of travelers:<br>- 0-24: 0-24 years<br>- 25-44: 25-44 years old<br>- 45-64: 45-64 years old<br>- >65: > 65 years old<br>Takes the value 'NA' for foreign resident travelers or when the information cannot be provided for privacy reasons. |
| sex | Sex of travelers; can take the values "male" or "female".<br>May take the value 'NA' for foreign resident travelers or when the information cannot be provided for privacy reasons. |
| travel | Number of stages expanded, without considering professional carriers. |

| Fields | Description |
|---|---|
| travel_km | Product of the "trips" field by the distance of the stage. |

### 4.3.2 Complete trip matrices

Unlike the stage matrices, the trip matrices consider as origin the place where the complete trip starts, and as destination the place where the trip ends. That is, a trip from Guadalajara by car to Adolfo Suárez Madrid Barajas airport, and from there by plane to Tenerife-Norte airport, and then by car to La Laguna, will have Guadalajara as the origin of the trip and La Laguna as the destination.

These are files in "CSV" format with information on the number of trips and trip-kilometers segmented by mode and means of transport, and by variables on the day and time of the trip, zoning of origin, destination and residence, itineraries, distance and sociodemographic variables, among others. Trip metrics and trip-kilometer metrics are available. Different files are generated according to the different zonings.

The file nomenclature is "YYYYYMMDD_Trips_complete_[zone].csv.gz", where "YYYYMMDD" is the year, month and day of study and "[zone]" the type of zoning used, which can take the values "districts", "municipalities" and "GAU". The travel matrix presents the following fields:

| Fields | Description |
|---|---|
| date | Date of study in YYYYMMDD format. |
| day_type | Code that identifies the day of the week and whether it is a day that is part of a standard full week of studies ("type" day) or whether it is a day that is studied because it is a day of the week.<br>mobility presents special patterns ("special" day), as is the case of holidays, eve of long weekends, etc.<br>It can take the values "LT", "MT", "XT", "JT", "VT", "ST", "DT" (Monday Type, Tuesday Type, Wednesday Type, etc.) for standard days or "LE", "ME", "XE", "JE", "VE", "SE", "DE" (Monday Special, Tuesday Special, Wednesday Special, etc.) for special days. |
| ID | Unique trip identifier for each combination of origin zone, destination zone, time period, main mode and main means.<br>This code is used to relate the travel matrix to the tour matrix, transport terminal influence matrix and cross-border mobility matrix. |
| period | Time of origin of the trip, in "HH" format |
| itinerary | Orderly sequence of the stages that make up the trip, concatenated by semicolons ";".<br>Each stage is referenced by the value of the "ID" field of the associated stage matrix "YYYYMMDD_Trips_stages_districts.csv.gz". |
| connections | Number of connections/transfers that make up the trip. |
| origin | Identifier of the area of origin. |
| destination | Identifier of the destination zone. |
| source_activity | Type of activity carried out in the area of origin of the trip, differentiating between:<br>- "home"<br>- "work_study"<br>- "frequent"<br>- "not_frequent" |
| target_activity | Type of activity carried out in the destination area of the trip, differentiating between:<br>- "home"<br>- "work_study"<br>- "frequent"<br>- "not_frequent". |
| pppal_mode | Mode of transport used in the longest distance leg of the journey, differentiating between "CCP" (road mode by a professional driver), "CPA" (road mode by a passenger), "F" (road mode by a professional driver), "CPA" (road mode by a passenger), "F" (road mode by a professional driver) and "CPA" (road mode by a |

| | |
|---|---|
| | professional driver). |

| Fields | Description |
|---|---|
| | rail), "A" (air mode) and "M" (sea mode).<br>The "mode_ppal" field can take the value "NA" for trips of less than 5 km that do not have origin and destination within an urban area to be studied in detail by surveys. |
| medium_ppal | Means of transportation used in the longest distance leg of the trip:<br>- In case the value of the "mode_ppal" field is "A" or "M", the value of this field will be "A" or "M", respectively.<br>- If the value of the "mode_ppal" field is "F", the value of this field may be "FC" (conventional rail) or "AV" (high-speed rail).<br>- If the value of the "mode_ppal" field is "C", the value of this field can be "PR" (private vehicle) or "AU" (bus). The "bus" mode is only distinguished for trips between autonomous communities.<br>- If the value of the "ppal_mode" field is "NA", the value of this field will be "NA". |
| modes | Orderly sequence of modes of the stages that make up the trip.<br>It is equal to the concatenation of the modes of each of the stages that make up the trip, separated by dashes "-". |
| media | Orderly sequence of modes of the stages that make up the trip.<br>It is equal to the concatenation of the means of each of the stages that make up the trip, separated by dashes "-". |
| distance | Distance between origin and destination measured in km along the route followed over the transport infrastructure. In case the "mode_ppal" field is "NA", the distance corresponds to the straight line distance between origin and destination.<br>The "distance" field can take the following values:<br>- 0-0.5: trips of less than 0.5 km.<br>- 0.5-2: trips between 0.5 km and 2 km.<br>- 2-5: trips between 2 km and 5 km.<br>- 5-10: trips between 5 km and 10 km.<br>- 10-50: trips between 10 km and 50 km.<br>- >50: trips of more than 50 km. |
| duration | Travel time.<br>The "duration" field can take the following values:<br>- 0-1: trips of less than 1 hour.<br>- 1-2: trips between 1 hour and 2 hours.<br>- 2-3: 2 to 3 hour trips.<br>- 3-4: trips between 3 hours and 4 hours.<br>- >4: trips of more than 4 hours. |
| travel | Number of expanded trips, excluding professional carriers. |
| travel_km | Product of the "trips" field by the distance of the trip. |

### 4.3.3 Tour matrices

A tour is the concatenation of activities and trips within the same day and that begins and ends in the same location. In this sense, a tour is a minimum of two trips, for example:

- Trip 1: from home to work.
- Trip 2; from work to home.

This information is delivered in files in "CSV" format with the information of tours based on the place of overnight stay carried out during the study day. The metrics of people and people-kilometers are segmented by variables related to the day, itinerary and zones visited, distances and durations among others. Different files are generated according to the different zonings.

The file nomenclature is "YYYYYMMDD_Tours_[zone].csv.gz", where "YYYYMMDD" is the year, month and day of study and "[zone]" the type of zoning used, which can take the values "districts", "municipalities" and "GAU". The tour matrix presents the following fields:

| Fields | Description |
|---|---|
| date | Date of study in YYYYMMDD format. |
| day_type | Code that identifies the day of the week and whether it is a day that is part of a standard week of complete studies ("type" day) or whether it is a day that is studied because its mobility presents special patterns ("special" day), such as holidays, eve of long weekends, etc.<br>It can take the values "LT", "MT", "XT", "JT", "VT", "ST", "DT" (Monday Type, Tuesday Type, Wednesday Type, etc.) for standard days or "LE", "ME", "XE", "JE", "VE", "SE", "DE" (Monday Special, Tuesday Special, Wednesday Special, etc.) for special days. |
| itinerary | Ordered sequence of trip identifiers that make up the tour, concatenated by comma ",".<br>The trip identifiers that appear in this field refer to the value of the "ID" field in the associated trip matrix "YYYYMMDD_Trips_complete_districts.csv.gz". |
| origin | Identifier of the area of origin of the tour.<br>The area of origin of the tour will always be the place where a traveler stays overnight. |
| areas | Ordered sequence of the areas visited in the tour, concatenated by commas ",". |
| modes_ppales | Ordered sequence of the main modes of travel that make up the tour, concatenated by comma ",". |
| media_ppales | Orderly sequence of the main means of travel that make up the tour, concatenated by comma ",". |
| distances | Set of distances of the trips that make up the tour.<br>It is equal to the ordered sequence of the distances of each of the trips of the tour, concatenated by comma ",". |
| durations | Set of durations of the trips that make up the tour.<br>It is equal to the ordered sequence of the durations of each of the trips of the tour, concatenated by comma ",". |
| people | Number of expanded people on the tour, excluding professional transporters. |
| people_km | Product of the "people" field by the distance of each tour. |

### 4.3.4 Cross-border mobility matrices

This product contains information on land border trips by road to France and Portugal, indicating, among others, the access point. It can be linked to the trip matrix from a common identifier.

Files are generated in "CSV" format with information on road trips to/from France or Portugal with indication of the national border area crossed, origin, destination, date and distance. The metrics are trips and trip-kilometers. Different files are generated according to the different zonings.

The file nomenclature is "YYYYYMMDD_Transborder_[zone].csv.gz", where "YYYYMMDD" is the year, month and day of study and "[zone]" the type of zoning used, which can take the values "districts", "municipalities" and "GAU". The cross-border mobility matrix presents the following fields:

| Fields | Description |
|---|---|
| date | Date of study in YYYYMMDD format. |
| day_type | Code that identifies the day of the week and whether it is a day that is part of a standard full week of studies ("type" day) or whether it is a day that is studied |

| Fields | Description |
|---|---|
| | because their mobility presents special patterns ("special" day), as is the case of holidays, eve of long weekends, etc.<br>It can take the values "LT", "MT", "XT", "JT", "VT", "ST", "DT" (Monday Type, Tuesday Type, Wednesday Type, etc.) for standard days or "LE", "ME", "XE", "JE", "VE", "SE", "DE" (Monday Special, Tuesday Special, Wednesday Special, etc.) for special days. |
| trip_id | Unique identifier of the trip that has origin and/or destination in a cross-border area. Refers to the "ID" field of the complete trip matrix. "YYYYMMDD_Trips_complete_districts.csv.gz" |
| origin | Identifier of the area of origin. |
| destination | Identifier of the destination zone. |
| entry_point | Border zone of entry to Spain.<br>The border area crossed by road is indicated.<br>If it is not a trip originating abroad, this field takes the value 'NA'. |
| exit_point | Border zone of exit from Spain.<br>The border area crossed by road is indicated.<br>If the trip is not to a foreign destination, this field takes the value 'NA'. |
| distance | Distance between origin and destination measured in km along the route followed over the transport infrastructure. In case the "mode_ppal" field is "NA", the distance corresponds to the straight line distance between origin and destination.<br>The "distance" field can take the following values:<br>- 0-0.5: trips of less than 0.5 km.<br>- 0.5-2: trips between 0.5 km and 2 km.<br>- 2-5: trips between 2 km and 5 km.<br>- 5-10: trips of between 5 km and 10 km.<br>- 10-50: trips between 10 km and 50 km.<br>- >50: trips of more than 50 km. |
| travel | Number of expanded trips, excluding professional carriers. |
| travel_km | Product of 'travel' due to the distance of travel. |

### 4.3.5 Matrices of the area of influence of the main public transportation terminals

These matrices contain information on the public transport terminal used in the stages carried out in these modes. It can be related to the trip and stage matrices from the corresponding identifiers.

These are files in "CSV" format with information on travelers using these terminals, characterizing the modes of access and dispersion. It provides information on trips and trip-kilometers for the date, trip and stage, origin and destination of the trip, distance, and terminals used. Different files are generated according to the different zonings.

The nomenclature of the files is "YYYYYMMDD_Influence_TP_TP_[zone].csv.gz", where "YYYYMMDD" is the year, month and day of study and "[zone]" the type of zoning used, which can take the values "districts", "municipalities" and "GAU". The cross-border mobility matrix presents the following fields:

| Fields | Description |
|---|---|
| date | Date of study in YYYYMMDD format |
| day_type | Code that identifies the day of the week and whether it is a day that is part of a standard week of complete studies ("type" day) or whether it is a day that is studied because its mobility presents special patterns ("special" day), such as holidays, eve of long weekends, etc. |

| Fields | Description |
|---|---|
| | It can take the values "LT", "MT", "XT", "JT", "VT", "ST", "DT" (Monday Type, Tuesday Type, Wednesday Type, etc.) for standard days or "LE", "ME", "XE", "JE", "VE", "SE", "DE" (Monday Special, Tuesday Special, Wednesday Special, etc.) for special days. |
| trip_id | Unique trip identifier of a trip that has some stage with origin and/or destination in the terminal, being that trip a trip of the deliverable "AAAAMMDD_Complete_GAU_Trips_GAU.csv.gz". |
| stage_id | Unique identifier of the leg of the trip that has origin and/or destination at the terminal, being that leg a leg of the deliverable. "YYYYMMDD_Travel_stages_districts.csv.gz" |
| terminal_type | Terminal type:<br>- airport<br>- port<br>- train<br>- bus. |
| station | Identifier of the airport, port, bus station or railroad station. The "Station Code" column is used as the station identifier. |
| terminal | Terminal identifier in cases where the station is divided into terminals. |
| stage_type | Relationship that exists between the stage and the transport terminal under consideration. This field can take the following values:<br>- "C" (connection): if the previous leg performed by the traveler is a leg with destination at the terminal in the mode associated with the terminal and the next leg performed by the traveler is a leg with origin at the terminal in the mode associated with the terminal.<br>- "A" (access): if the leg is not a connection and the next leg performed by the traveler is a leg originating at the terminal in the mode associated with the terminal.<br>- "D" (dispersion): if the leg is not a connection and the previous leg performed by the traveler is a leg with destination at the terminal in the mode associated with the terminal. |
| origin_travel | Identifier of the zone of origin of the trip. |
| travel_destination | Identifier of the destination zone of the trip. |
| country | Identifier code of the country or group of countries of residence of the traveler. The "country" field can take the following values:<br>- ES (Spain)<br>- DE (Germany)<br>- BE (Belgium)<br>- FR (France)<br>- IE (Ireland)<br>- IT (Italy)<br>- NL (Netherlands)<br>- PNO (Nordic Countries)<br>- PT (Portugal)<br>- GB (United Kingdom)<br>- REU (Rest of Europe)<br>- RMU (Rest of the World) |
| travel | Number of travelers expanded, excluding professional carriers. |
| travel_km | Product of the value of the "trips" field by the distance of the trip. |

### 4.3.6 Recurrent forced mobility matrix

In these matrices we take 14 typical days of each study month and calculate the recurrence of domestic users in those 14 days for each origin-destination pair and additional segmentations. The result is the number of people making each type of trip (origin, destination and other segmentations) 1 time, 2 times, etc. in those two weeks.

These are files in "CSV" format with information on the frequency of trips for each OD pair based on district zoning. Segmentation is provided by origin, destination, age, sex, residence and recurrence. The unit of measurement is persons.

The file nomenclature is "YYYYYMMDD_Mobligated_mobility_districts.csv.gz", where "YYYYMMDD" is the year, month and day of study. The forced mobility matrix presents the following fields:

| Fields | Description |
|---|---|
| month | Month of study in AAAAMM format |
| origin | Identifier of the zone of origin based on "zonificacion_distritos". |
| destination | Target zone identifier based on "zoning_districts". |
| age | Age range of travelers:<br>- 0-24: 0-24 years<br>- 25-44: 25-44 years old<br>- 45-64: 45-64 years old<br>- >65: > 65 years old<br>May take the value 'NA' for foreign resident travelers or when the information cannot be provided for privacy reasons. |
| sex | Sex of travelers; can take the values "male" or "female".<br>May take the value 'NA' for foreign resident travelers or when the information cannot be provided for privacy reasons. |
| residence | Place of residence at the provincial level, according to INE coding. |
| recurrence | Number of days that users make a trip from "origin" to "destination" during the two weeks of the month analyzed. |
| people | Number of persons expanded to meet the segmentation given by the previous fields, without considering professional carriers. |

## 4.4 Route studies by road

Route studies aim to visualize the information of road mode trips passing through the different road segments of interest. A route is defined as the unique set of section identifiers, and in the case of road mode, it consists of the spatial representation of the route. In this way, detailed geometric information is obtained for each complete route. A leg can have one or more routes. For example, one can go from home to work in a city through some streets in winter, and through others, looking for shade, in summer. These matrices are calculated from the stage matrices of the complete studies.

### 4.4.1 Routing matrices

For each origin and destination, the road routes that have been followed in trips of more than 5 km by road with origin and destination in national territory are indicated, as well as the average distance and the median duration of the stages that have used that route. The spatial representation of the number of trips can be done through the route identification, the route-leg relationship file and the leg geometry file, as explained in the following sections.

Files in "CSV" format are provided with information on private vehicle trips on the roads of interest, segmented by route, distance and duration (in seconds) of the stage.

The nomenclature of the files is "YYYYYMMDD_OD_routes.csv.gz", where "YYYYMMDD" is the year, month and day of study. The routes matrix presents the following fields:

| Fields | Description |
|---|---|
| **origin** | Identifier of the area of origin. |
| **destination** | Identifier of the destination zone. |
| **route** | Route identifier used to travel from origin to destination. |
| **distance** | Average distance of the stages between origin and destination measured in km along the route followed.<br>The "distance" field can take the following values:<br>- 0-0.5: trips of less than 0.5 km.<br>- 0.5-2: trips between 0.5 km and 2 km.<br>- 2-5: trips between 2 km and 5 km.<br>- 5-10: trips between 5 km and 10 km.<br>- 10-50: trips between 10 km and 50 km.<br>- >50: trips of more than 50 km. |
| **median_t** | Median duration (in seconds) of trips under the indicated segmentation. |
| **travel** | Expanded number of travelers, excluding professional carriers. |

### 4.4.2 Leg-route relationship

This product contains all the sections contained in each route, and their relationship, which are the keys to relate the route matrix with the spatial representation.

These are files in "CSV" format with the route composition information, indicating the sections that make up each route.

The nomenclature of the files is "YYYYYMMDD_relacion_relacion_tramo_ruta.csv.gz", where "YYYYYMMDD" is the year, month and day of study. The relational file presents the following fields:

| Fields | Description |
|---|---|
| **route** | Route identifier. |
| **section** | Leg identifier. |

### 4.4.3 Geometry of sections

Here you can find the geometry and spatial location of the sections. The file contains a field to be able to relate it to the route matrix through the section-route relationship file.

These are files in "SHP" format with the georeferenced information of the sections used in the route study.

The file nomenclature is "geometria_tramos.shp", where all the stretches used in the study are collected. The file has the following fields:

| Fields | Description |
|---|---|
| **section** | Leg identifier. |
| **geometry** | Geometry of the section. |

### 4.4.4 Segment information

These files contain information on the number of trips passing through each section. Stages with the same origin and destination whose average travel distance is less than 500 meters are excluded.

These are files in "CSV" format with information about the nature of the trips that use each section (short, medium or long, and intra-provincial, inter-provincial and intra-CAA). The nomenclature of the files is "YYYYYMMDD_tramos_info.csv.gz", where "YYYYMMDD" is the year, month and day of study. The tranche matrix presents the following fields:

| Fields | Description |
|---|---|
| **section** | Leg identifier. |
| **total** | Total number of trips passing through the section. |
| **short** | Total number of trips of less than 50 km passing through the section. |
| **medium** | Total number of trips of more than 50km and less than 250km passing through the section. |
| **length** | Total number of trips over 250 km passing through the section. |
| **intra_provincial** | Total number of trips made with origin and destination in the same province that pass through the section. |
| **inter_provincial** | Total number of trips made with origin and destination in different provinces that pass through the section. |
| **intra_ccaa** | Total number of trips made with origin and destination in the same autonomous community that pass through the section. |
| **inter_ccaa** | Total number of trips made with origin and destination in a different autonomous community that pass through the section. |

# Annex I. Methodological considerations regarding mobility analysis during the state of alarm.

The daily mobility studies of the present project are intended to give continuity to the continuous mobility monitoring studies carried out in the framework of the project "Analysis of mobility in Spain with big data technology during the State of Alarm for COVID-19 crisis management".

The methodology used for the calculation of indicators in this project is very similar to that used in the previous project, incorporating a set of relevant improvements. Among the improvements implemented, it is worth mentioning the change of the main data source used, being in this project data from network probes. These data provide a significantly higher temporal granularity than CDR data, potentially allowing a better characterization of mobility (for example, detecting some trips that with the previous data source could not be reported due to lack of records) and a substantial increase in the useful sample of users, which reduces the sampling error in the sample expansion processes. These methodological modifications may lead to changes in the estimates of mobility indicators between the two studies. Therefore, in the case that both sources of information are to be analyzed jointly, it is important to take into consideration that the variations in the observed mobility values may be motivated by: (i) the change in mobility patterns since 2021 and (ii) the methodological modifications incorporated in this project. It may be useful to analyze trends in mobility variation within each data series, with the objective of minimizing differences caused by the different methodological approaches used (e.g., variation in mobility between January 2021 and February 2021 versus variation between January 2022 and February 2022).