INSTITUTO NACIONAL DE ESTADISTICA

EM-1 study of mobility from mobile telephony

Technical Project

Sub-Directorate General of Socio-Demographic Statistics

December 2020

Experimental

Average

1 Introduction

Carried out every 10 years, the Population and Housing Censuses are a statistical operation with a long tradition and wide scope. Until 2001 they were carried out by sending questionnaires to all households in Spain. The population and housing census was carried out in 2011, with a sample of approximately 10% of the population. The 2021 census -which will be prepared with a reference date of January 1st- will be, for the first time, based on the use of records. Spain will join the small number of countries worldwide that are in a position to undertake the census using this type of methodology.

The 2021 census will be a major methodological challenge. The use of quality records, such as those currently existing in Spain, will allow the dissemination of information with a periodicity of less than 10 years and will increase information quality.

But while almost all the variables investigated in the census can be obtained through administrative records, there is a series of variables that are in great demand by users, and which are not available through these administrative sources. The continuity of the census series of these variables can only be guaranteed by carrying out specific operations.

The 2011 census, as was the case in previous editions, introduced several questions that made it possible to measure both daily mobility, and a certain very specific approach to measuring tourist movements (the study of what is known as "linked population"[1]).

Information related to the mobility questions was among the most requested in the 2011 Census. In it, people were asked about the location of their place of work/study (question 19 of the census questionnaire) and also if they spent 14 nights or more (for work or summer reasons) in a municipality other than the one of residence (question 4). With these two questions, the so-called linked population was calculated -included for the first time in the 2001 Census- which provides a better estimate of the real population burden that a municipality supports.

Additionally, three other questions related to the number of daily trips, the means of transport used and the time elapsed to get to the place of work/study were included.

Obtaining all this information from administrative sources is not possible. But there is a new and promising source that could provide an answer to much of the information provided in 2011: the use of **space-time information from mobile phones**.

The study proposed here is an analysis of the suitability of the statistical information obtained from the positioning of mobile phones. The objective is for this information to replace that traditionally offered in the Population and Housing Censuses.

Specifically, it is about building, with information from the three main mobile phone operators (MPO), three mobility matrices that will be detailed later.

[1]For more details see https://www.ine.es/censos2011/censos2011_meto_pobla_vinculada.pdf

For the pilot study, those mobile phone companies with a sufficient level of representation throughout the Spanish territory have been chosen.

According to official data published by the National Commission of Markets and Competition (July 2019), the number of mobile phones per operator is as follows:

Operator	Terminals	Market share (%of total)
Movistar:	16,270,132	30.3%
Orange:	13,705,972	25.5%
Vodafone:	12,293,129	22.9%
Masmóvil Group:	7,018,220	13.1%
Virtual mobile operators:	4,471,348	8.3%
Total:	53,758,801	100%

Specifically, the first three companies have a mobile telephony share that widely exceeds 20% and can be considered acceptable for the objectives pursued in this project.

It seems logical to assume that MPOs with market shares greater than 20% will be able to offer data for all geographic areas determined by the INE (about 3,200, as specified below). According to data by Autonomous Communities published by the NCMC[2], the first three companies are distributed quite evenly in the territory. In the worst case, these companies offer market shares of around 17% (in the case of Orange in País Vasco). This is not the case with the fourth company in terms of market share, the Masmóvil Group, which barely reaches 5% in some Autonomous Communities, according to this source and which could yield non-significant information in a large number of the areas under study.

2 Objectives

The main objective of this pilot statistic is to evaluate the suitability of mobile phone positioning information for use as an alternative to the traditional population census method in providing mobility matrices.

Three mobility matrices were thus constructed, which were published on the INE website within the experimental statistics section in order to obtain feedback from users and include this data source in the future within the plan for the regular use of demographic statistics.

In addition, based on the daily mobility information collected in the 2011 censuses, a new geographical unit can be defined, which will soon have an official character: the ***functional urban area*** (FUA)[3]. The information collected in this research allows us to analyse the possible delimitation of the AUF from this new source.

[2] http://data.cnmc.es/datagraph/jsp/graph/grafico-cuota-mercado.jsp updated to 2017 data
[3] https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Functional_urban_area

Although the definition is complex, essentially an FUA is a large city (normally greater than 100,000 inhabitants) and a set of municipalities that surround it, in which more than 15% of the workforce travels daily to that main municipality. This concept of building urban areas based on daily mobility, measured in one way or another and using different thresholds, is the one present in this study, although the delimitation of possible functional urban areas is not part of the work presented here. This will be the subject of a subsequent specific analysis, based on these same data.

All information is constructed with the sum of the results of the three operators, so that it will be impossible to determine any data for any specific operator.

3 Research Scopes

3.1 POPULATION SCOPE

The population scope consists of the mobile phones of the resident population in Spain with service from one of three above-mentioned operators; that is, mobile phones with national dialling codes. Foreign-numbered telephones on roaming -normally used by tourists- are excluded.

Therefore, we can implicitly affirm that the population scope is the population resident in Spain.

The information requested refers only to mobile phones, not to all devices that may have a SIM card.

3.2 GEOGRAPHICAL SCOPE

The geographical scope covers the entire national territory. The territory is divided into 3,214 "mobility areas" and information is provided for each of them.

3.3 TEMPORAL SCOPE

Depending on the specific matrix, there are several areas, as detailed in the subsequent section.

The collection period foreseen in matrices 1 and 2 is four consecutive working days. The choice of these dates is not very relevant. It is a matter of choosing a "normal" week in terms of mobility, that is, with normal educational and work activity, without holidays. The chosen week was the one composed of the days **18 to 21 November 2019 (Monday to Thursday)**

For matrix 3, four specific days are taken: July 20, August 15, November 24 and December 25, all of them from 2019. It is intended with them to have two representative days of summer - a time of maximum seasonal population such as

August 15 and another not as intense as a Saturday in July - Christmas day, and a "valley" weekend that it would coincide with the week of the working days chosen above, to have an almost complete picture of a "normal" week.

4 Characteristics of the investigation

Three types of matrices are built which are described below, always built as the sum of the data of the different operators.

The minimum unit of geographic observation is the **"INE mobility area"**. To determine the areas, the population registered in them is considered, so that each area must exceed the threshold of 5,000 registered persons (some few areas are very slightly below that threshold). This is to ensure that the number of users of any MPO in each area is sufficient for the analysis of mobility between areas. The areas would be built like this:

- Within each province, municipalities with less than 5,000 inhabitants are grouped geographically to reach at least that population threshold. The name given to each area is that of the largest municipality that makes it up.
- Each of the municipalities with between 5,000 and 50,000 inhabitants constitutes an area.
- Municipalities with more than 50,000 inhabitants can be broken down into districts or neighbourhoods, always with a threshold higher than those 5,000 inhabitants (about 1,150 areas for all of Spain). In the most extreme case, Madrid, the municipality is divided into the 128 neighborhoods that compose it since all of them exceed the threshold of 5,000 inhabitants contemplated. In cities with more than 250,000 inhabitants, the official districts are not used, but the neighbourhoods specifically for the INE's Urban Indicators project (SCD, sub-city districts)[4] because they are more homogeneous in population, although in many cases these DCS are the districts themselves.

The exact number of areas finally formed has been 3,214. The average size of each area is about 15,000 people, which means an estimated average number of 12,000 telephones adding the three operators. The complete list of areas is published on www.ine.es, as well as their geographical boundaries and the municipalities that comprise them.

With these areas delimited, the three products built are:

TABLE Nº 1: DAILY MOBILITY MATRIX

This is the main matrix of the project, as it is what allows defining the FUAs. It is an origin-destination matrix formed by N (3,214) areas x N areas. In the box "row

[4] See Urban Indicators project

A-column B" of it would appear the number of terminals that, on average, move daily from the residence area (A) to the daily destination area (B).

For the construction of this matrix, the area of residence of each of the **mobile phones of each MPO** must first be determined, using this method:

The four consecutive working days of November 2019 (Monday 18 to Thursday 21) are taken. The area of residence of each mobile phone is the area where the mobile phone is found for the longest time between 00:01 and 06:00 of the four days observed. After a joint analysis between the INE and the OTMs, that area of residence has actually been assigned using historical information of a minimum of 60 days. Thus, the area of residence of a telephone is the one where that telephone is most frequently found during the previous period considered (between two and three months, depending on the operator).

Assigned for each mobile phone the area of residence, an attempt is made to determine the area of daily destination. This recurring destination is built taking as reference only the same four days chosen above.

The method is similar to that of assigning the residence area but it will not always give a value: all the areas (including the residence area) in which the terminal is located must be taken from 10:00 to 18:00 as long as when that terminal is found in that area for at least four hours a day and at least two of the four days observed. If there is more than one area, the one that is most frequent (in which it is the longest), other than the residence area, will be taken. If only the residence area meets this condition, then the residence area will also be the destination area.

As mentioned above, this method will not always be able to obtain a destination area, because in some cases it will not be possible to determine that there is an area that is visited on a recurring basis. Therefore, it is not necessary to assign a destination area to all mobile phones, but it is nevertheless necessary that all mobile phones have an associated residence area.

TABLE Nº 2. MATRIX DAY AND NIGHT POPULATION

No source or destination information is used for this matrix. It is simply a matter of knowing how many mobile phones there are in each area of daily mobility, at M different times of the day. It is therefore a matrix of size (N x M). It allows to have information about floating population. 12 moments of the day have been taken with intervals of two hours (00:00 to 22:00 hours) to classify the zones according to the hours in which they are active.

It is important to bear in mind that matrix No. 2 is built taking the average of the same four days in November 2019 mentioned above.

Each mobile phone must be accounted for in a single area at all times. To do this, when it is not possible to determine which area it is in at the given instant and there is information on a previous or subsequent instant, the area in which it was at the closest instant in time to the chosen instant is assigned. Thus, if a telephone

is located in area A at 5:35 p.m. and in area B at 6:10 p.m., it is considered that at 6:00 p.m. it is in area B.

TABLE Nº 3. SEASONAL POPULATION MATRIX

It is about building an N x N x 4 matrix (or four N x N matrices). Each of the four corresponds to a specific day. The days of July 20, August 15, November 24 and December 25 have been considered, all of them in 2019.

For each day chosen, the row A-column B box shows the number of people who, residing in area A, spend the night in area B during that day, according to the following specifications:

For the chosen day, the area where it stays overnight is determined for each terminal (where that mobile phone is most frequently found during the period from 10:00 p.m. the day before to 06:00 on that date). Therefore, more precisely, the data for July 20 correspond to the period between 10:00 p.m. on July 19 and 6:00 a.m. on July 20, and thus respectively with the other three dates.

This method provides an overnight area for all telephones present in Spanish territory on those dates.

For each of the telephones, their area of residence is found using the same method described for matrix No. 1, taking as reference dates prior to the date under study in each case. In other words, to determine the area of residence necessary to prepare the December 25 matrix, the most frequent area of residence is taken during the previous two or three months.

5 Data collection and information processing

The INE receives the already aggregated matrices prepared by each operator. Therefore, it does not have individual information of any kind at any time, but only matrices of final results prepared by each of the OTMs that offer counts at the mobility area level.

It is important to note that the source used for the estimated location of mobile phones is from the telephone antennas, so the location of a phone cannot be determined with total precision. The margin of error is from tens of meters in urban areas to hundreds of meters, even kilometers, in rural areas

In addition, movements observed between bordering areas should be taken with caution because they may be caused by an error made when setting each terminal's position. This can result in a terminal being counted that is immobile in two different adjacent areas at different times.

This is an extract with simulated data of the information of the files type "matrix 1" sent by one of the operators:

```
54Mu;51MU;26
54Mu;52MU;30
54Mu;53MU;118
54Mu;54Mu;1797
54Mu;55MU;48
54Mu;56MU;#
54Mu;57MU;27
54Mu;58MU;130
54Mu;59MU;33
54Mu;60MU;56
54Mu;61MU;21
54Mu;62MU;#
54Mu;63MU;52
54Mu;64MU;20
```

That is, it would indicate that from area 54MU 26 telephones move to area 51MU, or that from area 54MU a number less than 10 or 15 terminals move to area 56MU (depending on the operator).

The nine matrixes of results received (three for each OTM) are integrated into three by adding the data from the OTMs.

Whenever the number of observations is less than 10 or 15 in an area, for a given operator, the result will be censored at origin. Therefore, origin-destination mobility data could only be offered for areas in which at least one of the three OTMs counts more than that threshold.

But in addition, the data is subjected to a second level of confidentiality: data from one cell of any matrix is only offered when the sum of the three operators, raised to population as explained below, is greater than 15 people. These limitations are irrelevant for the project that only aims to know the daily mobility of a certain entity. Flows of less than 15 people are grouped into categories other destinations and other origins. For this, the OTMs also offer the total flows for each origin and destination area, that is, they do not identify flows smaller than 10-15 but they do provide the data of the total number of telephones that leave or enter an area, provided that respect those same thresholds.

The three complete matrices are published in numerical form and also in the form of maps. As for the numerical information, the important thing about the project is to know the people who move between areas, so it has been chosen to offer information, not on the number of telephones that move but elevated to population figures.

5.1 PREPARATION OF MATRIX 1 (DAILY POPULATION)

The first section of the publication consists of six tables:

1.1 Summary of daily mobility by Autonomous Communities, provinces, islands and main cities (over 50,000 inhabitants or provincial capitals)

1.2 Summary of daily mobility by areas of residence

1.3 Origin-destination flows (more than15 people)

1.4 Destination-origin flows (more than15 people)

1.5 Origin-destination flows between main cities with more than 50,000 inhabitants or provincial capitals (more than15 people)

1.6 Destination-origin flows between main cities with more than 50,000 inhabitants or provincial capitals (more than15 people).

The flows of the three OTMs are added, provided that you have numerical data for at least one of them. These values are raised to the resident population aged 12-89 years for each area of residence. The choice of the age range is somewhat arbitrary. However, the user could undo this elevation and adjust it to their own since the population data 12-89 for each area is also published.

This age interval has been chosen since to measure daily mobility it is assumed that for these age ranges the probability that the person carries a mobile phone is much higher.

The following data is provided for each mobility area:

- Resident population as of January 1, 2019 (according to Register) (A)
- Resident population that is located during the day in your area of residence. It is about counting the population whose area of residence coincides with the destination area (B).
- Resident population found during the day in another area (C)
- Non-resident population that is located during the day in the area (D)
- Total population located during the day in the area (E = B + D)
- Population balance entering and leaving the area (F = DC)
- Flows of more than 15 people between the area and others, originating from or destined for the area under consideration.
- Flows of more than 15 people between the main cities (greater than 50,000 inhabitants or provincial capitals) that have their origin or destination in the city considered.

Thus, the resident population in an area (A) can therefore be divided into three groups: those found in the same area (B), those found in different areas (C) and the rest, which may or may not be found. They have a destination area in which they are located most of the time (A- (B + C)).

In addition to publishing the details of all flows over 15 inhabitants, data on the number of different origins or destinations is also provided which, based on the data received from the OTMs, exceed the threshold of 10-15.

In addition, a map is provided that provides the following information:

- Population summary by areas
- Origin-destination flows: for each mobility area, the flows of residents in it are analysed.
- Destination-origin flows: for each mobility area, the arrival flows to it during the day from other areas are analysed.

5.2 PREPARATION OF MATRIX 2 (DAY/NIGHT POPULATION)

The mobile telephony positioning data received from the OTMs offer some difficulty when it comes to analysis (this is common to all three operators). The universe of mobile phones of any operator located at a time of day is not constant, as might be expected. Significant fluctuations occur, especially at night, which can have their origin in different causes that are difficult to separate: on the one hand, telephones can be switched off at different times, especially at night; on the other hand, when a phone is connected to a Wi-Fi network, it is no longer visible to the antennas, which makes many telephones "disappear" when arriving at home or when connecting to the Wi-Fi at the work or study center; the population can leave the field of visibility of antennas for different reasons (borders, airplanes, ...).

Furthermore, the very technology with which the operator detects the phone may vary. Thus, you do not have the same amount of information if the positioning is done by CDR (*call detail records* or active events) or by probes (or passive events).

In summary, it has been more operative to offer the information for matrix 2 (population at different times of the day) taking as a reference the average number of mobile phones found at 8:00 p.m., which is equal to 100 in all areas, and offering global indicators of increase and decrease, instead of offering numerical data.

Therefore, a single table is offered that for each area offers the 12 average values of the number of telephones observed during the four days (November 18 to 21) that the observation lasted.

5.3 PREPARATION OF MATRIX 3 (SEASONAL POPULATION)

The problems mentioned in the previous point are also evident when dealing with the analysis of the seasonal population. Every day, the mobile universe of each operator is different. To the above reasons must be added the possible more pronounced departure of the population abroad during the summer or even the variation in the market shares of the operators, given that every month there are significant movements of change of operator.

Determining, therefore, what is the total number of mobile phones resident in a given area requires arbitrary decisions. It has been chosen to take as the number of mobile phones residing in an area the maximum number of mobile phones observed in that area throughout the four days observed, taking for this the sum of the three operators together (not the maximum of each operator per separated).

With this, the total number of mobile phones observed on a given day will logically always be less than the maximum. Subsequently, the data are raised to the total population residing in Spain at 1-1-2019 according to the standard, unlike what is done in matrix 1, in which the elevation is to the population between 12 and 89 years old. The explanation for raising matrix 1 to a subset of the total population lies in the fact that matrix 1 tries to measure the daily trips that are made mainly to go to the workplace or study during working hours, so it seems logical to limit the population. It is not possible to take other reference dates for the registered

population since only the number of inhabitants per mobility area is available as of January 1. Thus, the total population is going to be, in the worst case, 96% (August 15) of the registered population and in the best, 98% (Christmas Day).

This method results in a large population appearing spending the night in areas bordering each area of residence. To the general reason mentioned above about the lack of precision of the antennas when it comes to accurately locating a telephone, it is necessary to add a specific one of the method chosen to elaborate this matrix. A person who resides in a certain area and moves at night (let's say he goes out to dinner) to a nearby area, if he is in that destination area for more than four hours in the analyzed time interval (which goes from 22:00 to 6:00) will be counted as "staying overnight" in the destination area. Therefore, in general, movements between neighboring or very close areas should not be considered as true movements.

Again, it is important to note that these tables only offer data on where the resident population is located, not tourists, who probably in the main tourist destinations add very relevant amounts to those. Therefore, it is not a question of estimating in any case the total population that is found in each area at any given time, but only knowing how the population residing in Spain is distributed by areas of mobility.

The following data is provided for each mobility area:

- Number of different overnight areas in which the population residing in the area is spending the night
- Number of different areas of residence in which the population that is spending the night in the area resides
- Resident population (A)
- Total resident population found on selected date (B = C + D))
    - Resident population that spends the night in their area of residence (C)
    - Resident population found that stays overnight in another area (D)
- Non-resident population that stays overnight in the area (E)
- Total population that stays overnight in the area (E)
- Population balance of the area (G = F - A)

The tables offered are:

3.1 Summary of seasonal population by dates, Autonomous Community, provinces, islands and main cities (over 50,000 inhabitants or provincial capitals).

3.2 Summary of seasonal population by dates and areas of mobility

3.3 Flows of residence/overnight stay by dates and areas (more than15 people)

3.4 Overnight/residence flows by dates and areas (more than15 people)

3.5 Flows of residence / overnight stays by dates and main cities, greater than 50,000 inhabitants or provincial capitals (more than15 people)

3.6 Overnight / residence flows by dates and main cities, over 50,000 inhabitants or provincial capitals (more than15 people)

3.7 Population staying overnight in a different Autonomous Community than the one of residence, by dates

3.8 Population staying overnight in a different province than the one of residence, by dates

Likewise, a map is offered for each day, for the analysis of the seasonal population that offers the following information:

- Summary of seasonal population by areas of mobility
- Residence-overnight flows: for each area, it is analyzed where residents are spending the night in it.
- Overnight-residence flows: for each mobility area, the population that stays overnight is taken and the areas of residence of all of them are analyzed.