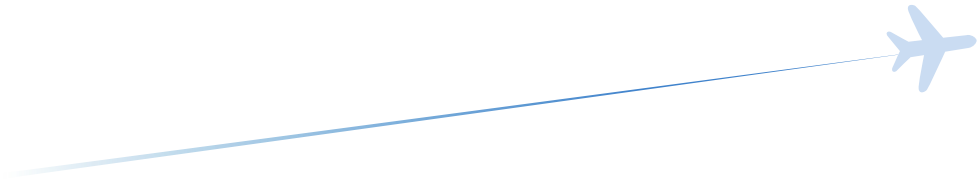

Analysis of Passenger Behaviour from ICT-based Geolocation Data

D 3.1

BigData4ATM

Grant:	**699260**
Call:	**H2020-SESAR-2015-1**
Topic:	**Data Science in ATM**
Consortium coordinator:	**Nommon**
Edition date:	**2 February 2018**
Edition:	**01.01.00**

Founding Members

EUROPEAN UNION EUROCONTROL

SESAR JOINT UNDERTAKING

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Pedro García / Nommon	Project Contributor	15/11/2017
Riccardo Gallotti / IFISC	Project Contributor	15/11/2017
José Javier Ramasco / IFISC	Project Contributor	15/11/2017
Gennady Andrienko / Fraunhofer	Project Contributor	15/11/2017

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
José Javier Ramasco / IFISC	Project Contributor	21/11/2017
Oliva García-Cantú / Nommon	Project Contributor	21/11/2017

Approved for submission to the SJU By — Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
José Javier Ramasco / IFISC	Project Contributor	24/11/2017
Ricardo Herranz / Nommon	Project Coordinator	24/11/2017

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
-	-	

Document History

Edition	Date	Status	Author	Justification
00.00.01	11/11/2017	Draft	José Javier Ramasco	Initial draft
00.01.00	24/11/2017	Approved for submission to the SJU	José Javier Ramasco	Review and approval by BigData4ATM Consortium
01.00.00	01/02/2018	Approved by SJU	José Javier Ramasco	Approval by SJU
01.01.00	01/02/2018	Approved by SJU	José Javier Ramasco	Version for public dissemination

BigData4ATM

PASSENGER-CENTRIC BIG DATA SOURCES FOR SOCIO-ECONOMIC AND BEHAVIOURAL RESEARCH IN ATM

This document is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No 699260 under European Union's Horizon 2020 research and innovation programme.

Abstract

This document presents the methodology and the algorithms developed to extract relevant information about passenger behaviour from the data sources described in deliverable D2.1. The results obtained with the proposed methodology are validated by comparing them with those obtained through alternative, more traditional data sources. The advantages and limitations of the information extracted from new, big data sources are discussed with a view to inform their use in the case studies to be developed in WP4.

Table of Contents

Founding Members
EUROPEAN UNION EUROCONTROL

Executive summary

The aim of BigData4ATM is to explore the potential use of the big data coming from personal mobile devices to obtain information that is useful for the management of ATM system. The different sources of data collected by the project are described in deliverable D2.1 'Inventory and Quality Assessment of Data Sources for ATM Socioeconomic and Behavioural Studies'. The present document describes the methodology and the algorithms developed to extract relevant information about passenger behaviour from these data sources, as well as the validation experiments conducted to ensure the validity of such information and assess its potential limitations.

The document encompasses the following analyses:

- analysis of door-to-door mobility patterns and travel times;
- analysis of passenger mobility and expenditure patterns at airports;
- analysis of people's opinions, sentiments and attitudes towards air transport and ATM.

First, the methods to extract mobility information from mobile phone records and geolocated information coming from Twitter are thoroughly described. The results are validated in the cases where information from other sources such as surveys or census is available. The deviations and limitations of the methods are studied in detail, and the capacity of these methods to study door-to-door mobility is discussed. We then analyse passenger mobility and expenditure patterns at airports, and search for differences between a normal situation and days with a high delay impact. Finally, a semantic analysis is run on the content of Twitter messages in days with low performance of the air traffic networks. In this last case geolocated tweets are not statistically representative to produce a satisfactory result, so a new method with a wider information search is proposed.

1 Introduction

1.1 Scope and objectives

The goal of BigData4ATM is to investigate how different passenger-centric geolocated data can be analysed and combined with more traditional demographic, economic and air transport databases to extract relevant information about passengers' behaviour, and to study how this information can be used to inform air transport and ATM decision making processes. The specific objectives of the project are the following:

1. to develop a set of methodologies and algorithms to acquire, integrate and analyse multiple distributed sources of non-conventional spatio-temporal data coming from Information and Communications Technologies (ICT) — including mobile phone records, data from indoor geolocation technologies, credit card records and data from Internet social networks, among others — with the aim of characterising passengers' behavioural patterns;
2. to develop new theoretical models translating these behavioural patterns into relevant and actionable indicators for the planning and management of the ATM system;
3. to evaluate the potential applications of the new data sources, data analytics techniques and theoretical models through a number of case studies relevant for the European air transport system.

During the initial stage of the project, different datasets potentially relevant for the project, including both traditional data sources and 'Big Data' sources, were collected and characterised. Once these data were collected, it was necessary to develop methods and algorithms able to extract passenger behavioural patterns out of the data. The aim of this document is to describe these methods as well as the tests and checks done to ensure their correct performance.

1.2 Reference and applicable documents

- Grant Agreement No 699260 BigData4ATM – Annex 1 Description of the Action
- BigData4ATM D1.1 Project Management Plan, v01.00.00, November 2016.
- BigData4ATM D1.2 Data Management Plan, v00.01.00, November 2016.
- BigData4ATM D2.1 Inventory and Quality Assessment of Data Sources for ATM Socioeconomic and Behavioural Studies, v00.01.00, November 2016.

1.3 List of acronyms

Acronym	Definition
ADS-B	Automatic Dependent Surveillance - Broadcast
ALDT	Actual Landing Time
ANSP	Air Navigation Services Provider
ATFM	Air Traffic Flow Management
ATM	Air Traffic Management
ATOT	Actual Take-Off Time
API	Application Programming Interface
ASQ	ACI's Airport Service Quality
IATA	International Air Transport Association
CDR	Call Detail Record
CODA	Central Office for Delay Analysis
ECAC	European Civil Aviation Conference
EMMA	Estudios de Movilidad del Modo Aéreo (Aerial Mode Mobility Studies)
EU	European Union
GDS	Global Distribution System
GPS	Global Positioning System
ICT	Information and Communications Technology
INE	Instituto Nacional de Estadística (Spanish National Statistical Office)
MCC	Mobile Country Code
MIDT	Marketing Information Data Tapes
MNO	Mobile Network Operator
NUTS	Nomenclature of Units for Territorial Statistics
POS	Point Of Sale
SIBT	Scheduled In-Block Time
SID	STATFOR Interactive Dashboard
SOBT	Scheduled Off-Block Time
STATFOR	Statistics and Forecast Service
PaxIS	Passenger Intelligence Services

Table 1: List of acronyms

1.4 Structure of the document

The deliverable is organised as follows:

- Section 2 presents the analysis focused on extracting mobility information from mobile phone records and Twitter data. For both data sources, the chapter is structured in a parallel way: section 2.1 presents the data characteristics and a first data exploration; section 2.2 describes the methodology for the extraction of mobility patterns; section 2.3 discusses the upscaling methods used to expand the sample to the full population; finally, section 2.4 we present the validation of the results against other data sources. Section 2.5 presents a visualisation tool developed in the context of the project that facilitates the direct exploration of mobility flows.
- Section 3 presents the analysis of passenger behaviour inside and around airports using Twitter data (section 3.1) and credit card records (section 3.2).
- Finally, section 4 details the results of a first semantic and sentiment analysis using the contents of the tweets on days with air traffic performance problems.

2 Door-to-door mobility analysis

2.1 Mobility data characterisation and exploration

Two main data sources have been used for the analysis of passenger door-to-door mobility: mobile phone records and Twitter data. A brief summary of the main characteristics of these data sources is included here to facilitate the understanding of the analysis presented throughout the document. A more detailed characterisation of the data can be found in D2.1 'Inventory and Quality Assessment of Data Sources for ATM Socio-economic and Behavioural Studies'.

2.1.1 Mobile phone data

In this document, the term 'mobile phone data' is used to refer to Call Detail Records (CDRs). A CDR is a register that is generated every time a user interacts with the mobile phone network. It records the antenna to which the user was connected and the timestamp of the interaction. This interaction may be directly motivated by the user (e.g., sending/receiving a call, using the Internet), by apps running in the background (e.g., e-mail apps looking for new messages) or by the network (e.g., changes of coverage area). .Since the temporal granularity of the data is not solely dependent on the user behaviour, CDRs present higher temporal resolution than data coming from other geolocated devices that require an interaction of the user with the device to generate a record. Figure 2.1.1.1 shows the probability density, obtained from the CDR data, of having two consecutive registers at less than a given time difference. It can be seen that there are peaks of activity each multiple of 30 minutes, due to automatic network updates.

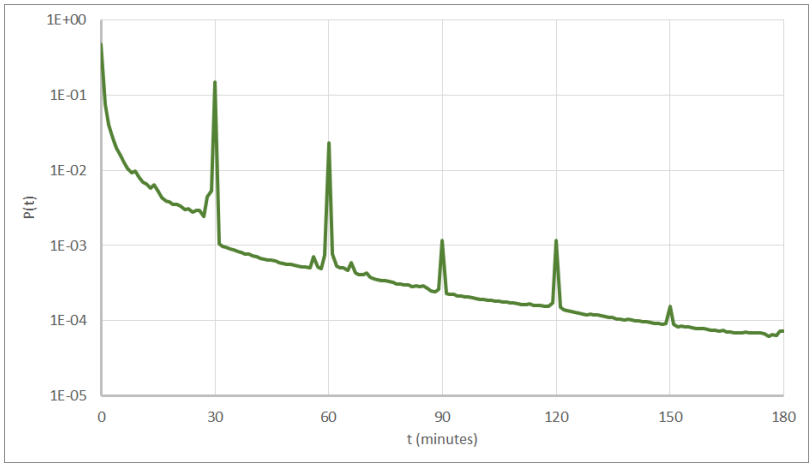

Figure 2.1.1.1 Cumulative distribution of time between consecutive internet CDRs

In terms of spatial granularity, the geolocation information of the CDRs is expressed at antenna level. This results in an accuracy of 100-200m in urban environments (where the mobile network is quite dense) and up to several kilometres in rural areas.

The BigData4ATM project has access to the CDR database of Orange Spain, which grants a continuous flow of information for around 30% of the Spanish population. This database also contains the CDRs associated to roamers that connect to the Orange Spain network, which allows us to study the relationship between the ATM system and the mobility of foreign visitors in Spain. In Figure 2.1.1.2a, we present the number of records per user generated during one day of July 2016. It can be seen that while 50% of national mobile phone users have more than 50 registers during a day, only 10% of roamers present this number of registers per user. Recent changes in EU legislation removed roaming charges leading to a higher mobile phone usage by EU roamers. This can be observed in Figure 2.1.1.2b and Figure 2.1.1.2c, where the number of records per user generated during one day of July 2017 is presented. There it can be observed that the number of records per user for roamers increased with respect to 2016. In Figure 2.1.1.2c, it can be seen that the number of registers belonging to EU roamers is halfway between the number of registers generated by national users and non-EU roamers.

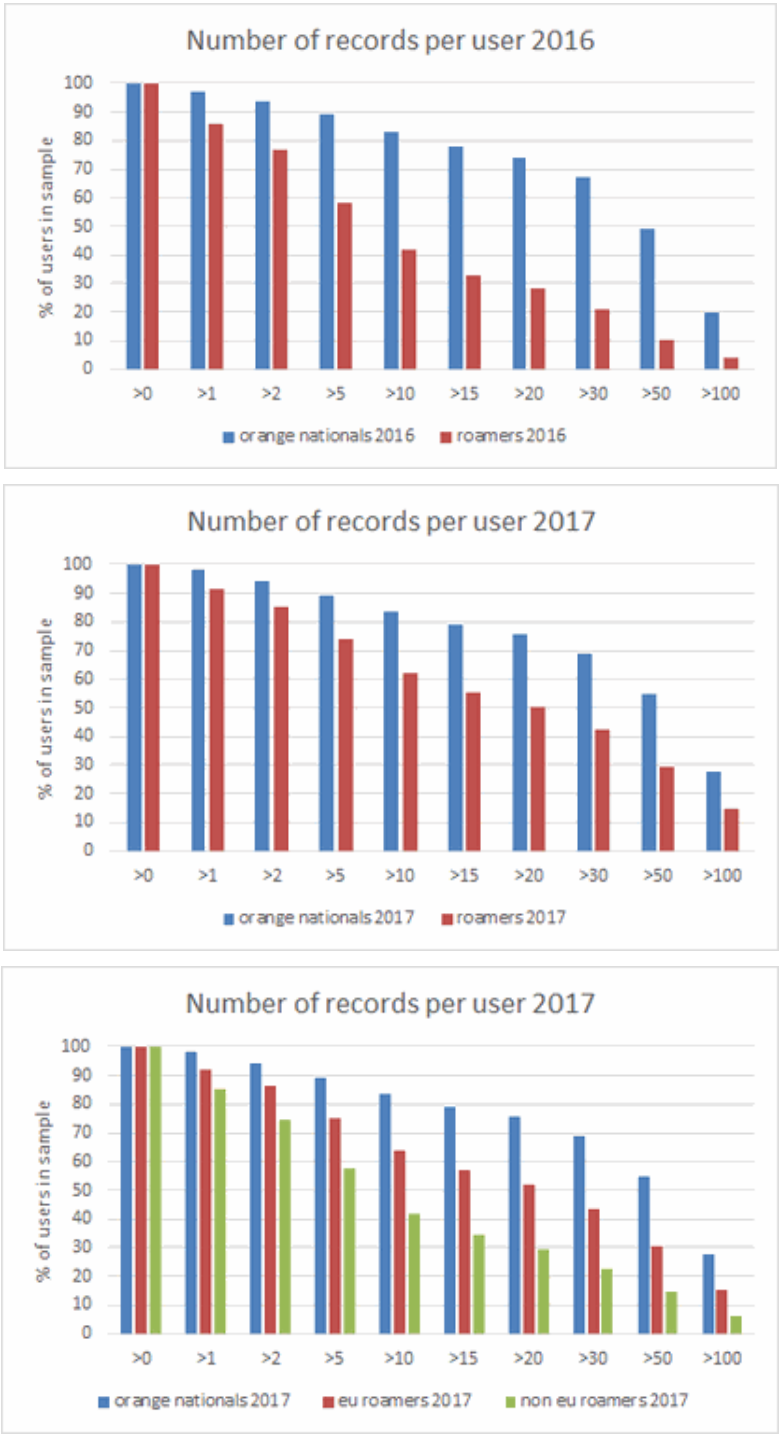

Figure 2.1.1.2 Number of registers per a) (top) national and roaming users in July 2016, b) (middle) national and roaming users in July 2017 c) (bottom) national and roaming users in July 2017 differentiating EU and non-EU users

2.1.2 Twitter data

2.1.2.1 Information available in each tweet

A hardware and software infrastructure to continuously query the Twitter API for geolocated tweets has been implemented at IFISC. The system allows capturing the almost entirety of tweets containing geolocated information in any part of the world. A tweet is saved as a dictionary, in which the different keys capture different pieces of information. The most important keys for the current project are:

- ***text:*** the content of the message, which is a string limited to 140 characters (280 starting from November 2017), possibly comprising links, hashtags (#) or mentions (@);
- ***created at****:* it contains the UTC time at which the message has been written;
- ***user****:* this is a unique id per user. The API provides the Twitter user id, which is encrypted to ensure privacy protection;
- ***place***: querying for geolocated information ensures, with the exception of a limited number of database errors, the existence of information concerning the place. The place field associates each tweet with a location that can be of different types and scale: poi (point-of-interest), neighbourhood, city, admin (administrative area), and country. A bounding box for the area is provided (corresponding to a point for the poi), together with name of the place, name of the country, country code and a univocal place id. Names are provided in different languages. Therefore, the same place id and bounding box can be associated to different names.
- ***coordinates***: if the users allow it, also the exact latitude and longitude of the tweet are provided.

2.1.2.2 Temporal span and changes in the user interface

The recording of the database discussed in this deliverable started in October 2014. Another database is available with data for the preceding years (2012-2014), but it has different characteristics and will not be treated here. For the period November 2014 - December 2016, considered in our analysis, over one billion tweets were recorded.

Between April 27 and 28, 2015, Twitter's user interface underwent a major change. The new interface allowed users to decide more easily for each tweet whether to share location information or not. As a consequence of this, there was a remarkable drop in the availability of tweets with high precision location (coordinates). This can be seen in Figure 2.1.2.2.1, in which the evolution of the number of users detected daily in the database is plotted as a function of time. After May 2015, only 25% of the users remain with the tweet's coordinates option on. However, the use of the place field became generalised, providing an alternative for tweets' location. Indeed, in Figure 2.1.2.2.2 it can be also appreciated that the number of users with geolocated information has not decreased but it remains mostly constant or even increasing, in many cases substituting place by coordinates. The resolution of the field 'place' is enough to analyse the long-range displacements targeted in most of this project.

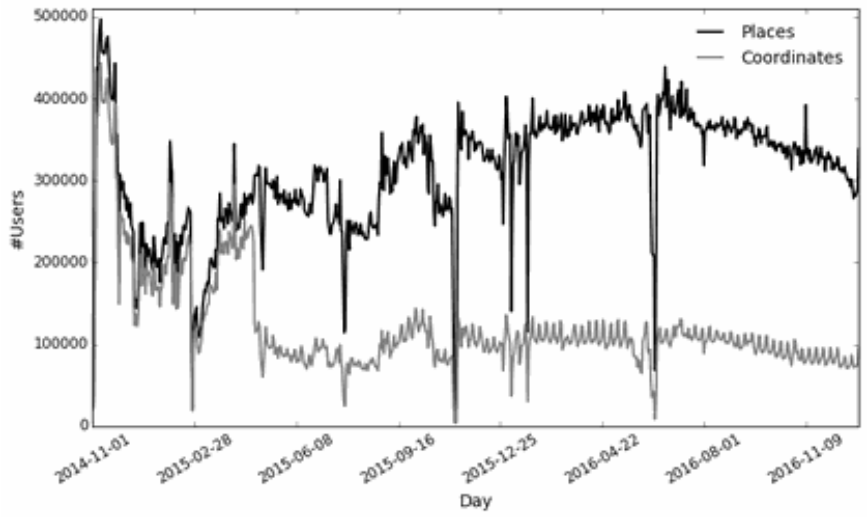

Figure 2.1.2.2.1 Number of users detected by day in the database posting geolocated tweets in Europe (ECAC area + Russia and Belarus)

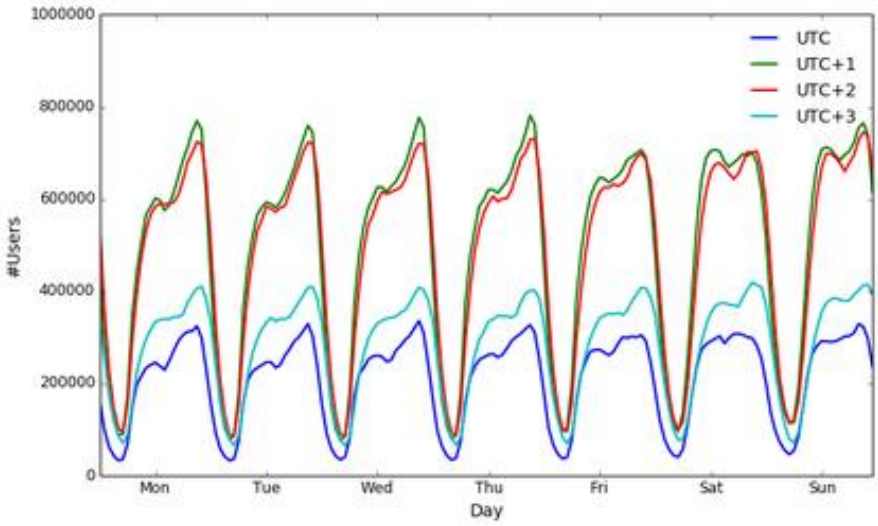

Figure 2.1.2.2.2 Number of users detected at least once at each hour of the week (local time) in the database posting geolocated tweets in Europe (ECAC area + Russia and Belarus). The tick in the x-axis represents midday of the associated day in local time.

2.1.2.3 Area of analysis

Given the wide spatio-temporal extension of the data available the focus is set on the European Civil Aviation Conference area (represented by shades of blue in the image below). For the sake of simplicity, from now on we will refer to the ECAC area as 'Europe' (see Figure 2.1.2.3.1).

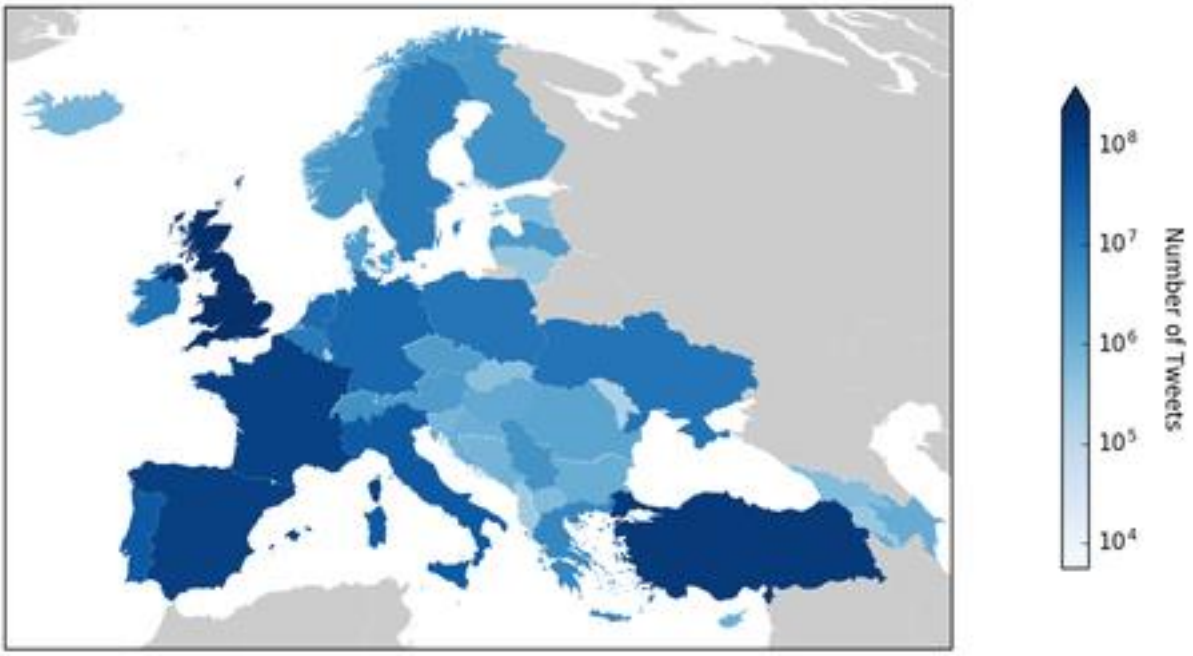

Figure 2.1.2.3.1 Map of the countries selected for the Twitter analysis and the associated number of tweets recorded.

Europe represents a very interesting laboratory for the proposed analysis. All countries included are technologically developed, thus allowing us to have a relatively large user base. At the same time, the socioeconomic landscape is relatively inhomogeneous. As a consequence of the particular local economic conditions, it is expected to observe significant differences not only in the use of Twitter as a communication device, but also in the modal split between air and ground transportation. Naturally, the modal split will also vary strongly because of the different geographical constraints and typical distance of the trips for each country. This inhomogeneity represents a challenge, but also an opportunity for BigData4ATM, as any methodology developed is expected to be easily applicable to any other area of the globe.

2.2 Mobility information extraction

2.2.1 Mobile phone data

The analysis of mobile phone data for mobility information extraction differentiates between Orange Spain clients and roamers, as they do not only present different register generation rates, but they also have different sampling frameworks.

- For national mobile phone users, there exist in the literature robust solutions to extract mobility information for ground transport. However, several extensions were required to be able to adequately capture door-to-door trips involving one or more air transport legs.
- For the analysis of roamers' data, there were two objectives: studying how the different register generation rates of roamers impacted the valid sample criteria and finding an adequate upscaling method able to correct potential biases in the roamers data and produce representative results.

2.2.1.1 Mobility information extraction for Spanish residents

2.2.1.1.1 Data pre-processing

Raw CDR data cannot be directly used to determine the mobility of the Mobile Network Operator (MNO) clients, as this kind of data is very noisy. There are two main sources of data errors that need to be cleaned to accurately replicate the mobility of mobile phone users: mislocated antennas and the so-called 'ping-pong effect'.

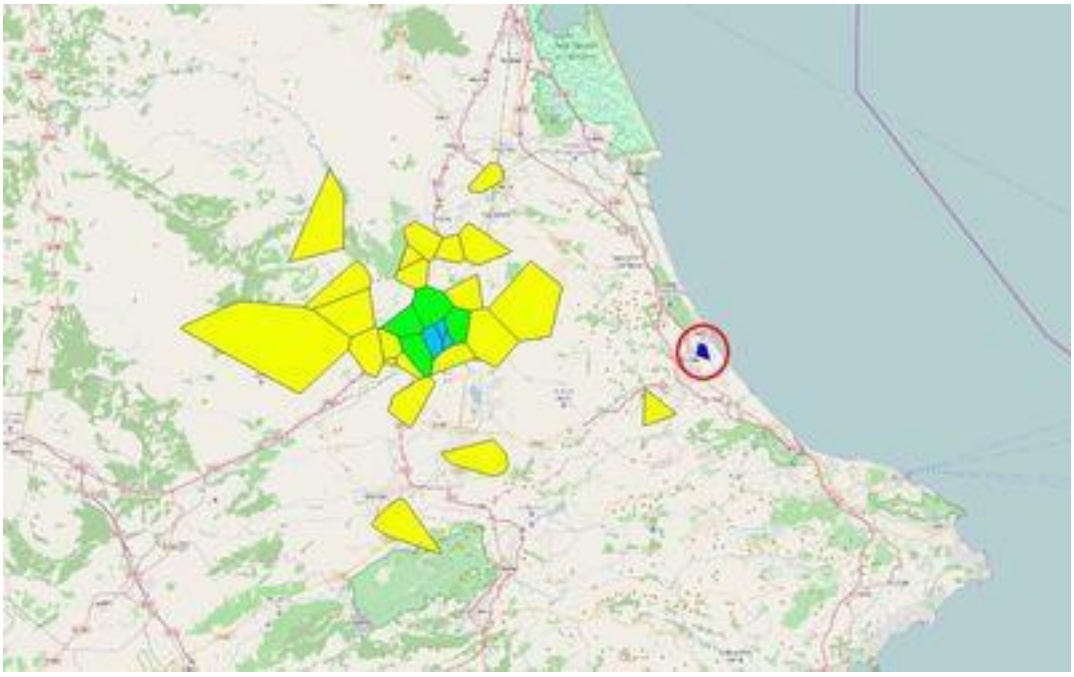

Figure 2.2.1.1.1.1 Example of the detection of a misplaced antenna.

The mobile phone network is constituted by a large number of antennas: for a country like Spain there are more than 150,000 antennas for a single MNO. This infrastructure is also characterised by great dynamism. Tower replacement and changes in the environment (new buildings, activation of support antennas for special events, etc.), performed to provide coverage in the best possible way, may lead to changes in the structure of the antenna distribution. Sometimes, these changes are not reflected in the antenna location databases of the MNO. This results in mislocated antennas that need to be filtered out. The methodology used to correct wrong antenna locations is based on the assumption that a user connecting to an antenna has a higher probability of connecting to other antennas spatially close to that antenna than to other antennas located further away, and that temporally close antennas should also be spatially close. A map of relationships between antennas, which relates each antenna with the other antennas users typically connect before/after connecting to it, can be derived from CDRs. By comparing this map with the coverage maps derived from the antenna location databases, it can be determined which antennas are out of place. An example is shown in Figure 2.2.1.1.1, where the distribution of antennas with a significant number of connections made before/after connecting to the one highlighted with the red circle is represented. It can be seen that these antennas are spatially clustered in another location. This indicates that the

antenna under analysis is mislocated. Once these mislocated antennas have been detected, their location can be corrected and assigned to the centroid of the cluster of temporally close antennas.

The ping-pong effect occurs when a user is located in the edge between two adjacent coverage areas, and rapidly (in matter of few seconds) connects repeatedly between the two of them, giving a false feeling of back and forth movement. These jumps between consecutive locations must be filtered out if we want to avoid overestimating of the number of short distance trips made by mobile phone users. In the literature, there are several approaches to solve this problem:

- speed-based filters, which consider as false positions those produced by a movement of a speed higher than a certain threshold;
- oscillation-based filters, which identify connection sequences produced by this ping-pong phenomenon;
- cluster-based approaches, which group connections made to spatially close antennas, filtering out the observed jumps.

The problem with speed-based filters is how to set the speed threshold, as it is affected by the network density: a 1 second jump made between adjacent towers will produce different speeds depending on the distance between those adjacent towers. These distances may range from few hundred metres in urban environments to a couple of kilometres in rural areas, leading to highly unrealistic speed values. On the other hand, setting a too low speed threshold could result in filtering out trips made by plane, which in the BigData4ATM context are the object of study. The main drawback of the oscillation-based filters is that they need to predefine the sequences that will be classified as a jump, which may be very diverse and difficult to capture since the area of study is a whole country and the size of the sample exceeds 8 million mobile phone users. The problem of the cluster-based approach is that it may underestimate the number of short distance trips, as it filters out connections made to spatially close antennas. Since BigData4ATM focuses on long distance trips, a cluster-based approach has been followed to eliminate the ping-pong effect.

2.2.1.1.2 Extraction of a valid sample

Another important aspect of working with CDR data is identifying those mobile phone users for whom mobility information can be extracted (i.e., determining the valid sample). This is usually carried out based on mobile phone activity criteria. Therefore, for a given day, only those users that do not have long time intervals without generating mobile phone registers are considered part of the valid sample. However, this needs to be done carefully when trying to obtain indicators relevant for the air transport field:

- It is very rare that mobile phone users produce registers while they are travelling on a plane. For mobile phone users performing national air trips, this can lead to periods of time where the user “vanishes” from the mobile phone network. If this period is long enough, the user may be removed from sample, therefore losing air trips.
- Also, passengers arriving to the country in the afternoon or departing during the morning usually present a big gap without registers. Removing them from the valid sample will introduce important biases, as ground trips to/from the airport will not be calculated.

Hence, a ‘time without registers’ criterion must be coupled with a clause that allows taking into account travel times by plane.

2.2.1.1.3 Identification of recurrent locations

Once the raw CDRs have been cleaned from possible errors, an analysis of locations frequently visited by the mobile phone users is carried out. The aim of this analysis is to determine locations that correspond to meaningful places in the user's daily life: home and work mainly, but also other recurrent places that the user may visit, such as second residences visited during weekends or other frequent activities. The residence place is determined as the most visited location during sleep time, and work place as the most visited location during working hours. A criterion based on mobile phone activity can be added under the assumption that, during the period of time a user is at home, his mobile phone activity should decrease (due to the use of Wi-Fi or because he is sleeping). These criteria help to identify night jobs, which would not be detected by simply using criteria based on time of the day. Also, a minimum activity threshold is applied in order to avoid assigning home location to a place with few connections, mainly because the user has a small amount of registers. In this case, the user is removed from the effective sample.

2.2.1.1.4 Reconstruction of activity-travel diaries

To obtain the mobility of the mobile phone users, it necessary to determine whether the different registers correspond to a stay at a place or an intermediate register in a trip context. Additionally, stays in a given place may be due to activities (work, going to the gym, etc.) or part of a trip chain (staying at the airport, or having a half an hour rest while driving). These subtleties can make the difference between obtaining good mobility indicators or poor estimates. In order to identify the complete door-to-door trip chain of an air transport user, the stays at the airport must be detected and classified as intermediate stops of the trip, allowing us to identify the real origin and destinations of the passenger and the times spent in each phase.

The approach followed in this project is based on three steps:

- First, from the mobile phone registers of a user, stays and trips between stays are determined.
- Then, with the aid of land use information, stays are classified as either activities or stops. Registers that have a strong semantic meaning (e.g., registers close to an airport or corresponding to a change in the mode of transport) are also classified as a stop.
- Finally, mode and route is assigned to the trip chains identified in previous steps with the aid of external information, such as Google Maps APIs and air transport connections and timetables.

The distinction between stays and intermediate trip points is carried out through a spatio-temporal clustering algorithm. This approach cleans the noisy locations caused by the ping-pong effect, without losing information about the long-distance trips. The clustering algorithm consider two thresholds: a spatial threshold to decide whether consecutive stays belong to the cluster (i.e., the same stay) and a temporal threshold representing the minimum time a user must spend in a location in order for it to belong to the cluster (i.e., for the location to be considered as stay). This clustering algorithm is applied in two steps, with two different sets of thresholds, with the aim of identifying first long distance trips (considering larger spatial and longer temporal thresholds) and then, during the second step, the mobility at urban level (considering shorter spatial temporal thresholds). This way, both the door-to-door trip and the airport access/egress modes can be obtained. An example of the application of the clustering method for both short and long-distance trips is depicted in Figure

2.2.1.1.4.1. There, a trip from Algeciras to Badalona that involves a Malaga-Barcelona flight is represented. The purple polygons represent the signals left by the mobile phone user during his trip. The black circles represent the places where the user has been detected long enough to assign a stay there. These places correspond to two types of stays: activities in Algeciras and Badalona, and stops at the Malaga and Barcelona airports.

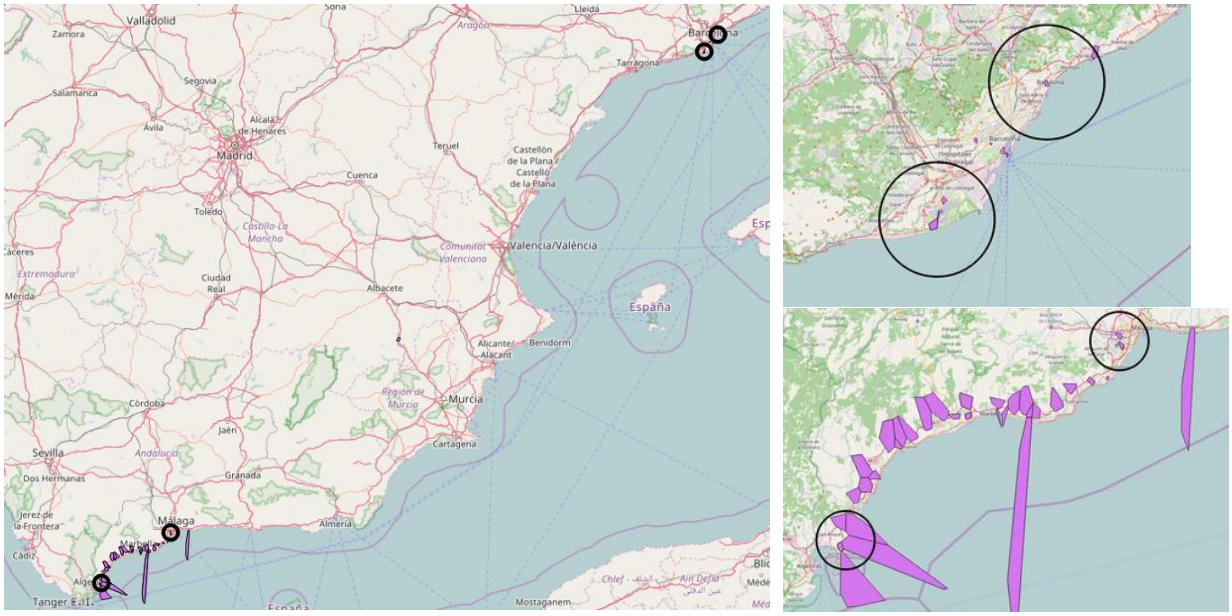

Figure 2.2.1.1.4.1 Maps with the towers involved in a trip Algeciras – Badalona.

The second step classifies the stays resulting from the clustering algorithm as activities or stops. Stays are classified as stops when they take place near an airport/port/train station and they are the origin/destination of a trip with destination/origin and speeds compatible with the mode of transport represented by the location (for example, a stay in an airport will only be considered a stop if before/after there is a trip characterised by a long period without registers and a typical flight speed profile). These locations may imply long stays, particularly at airports, but they are not an activity. Stays are classified as an activity when they do not happen near these significant transport locations. Additionally, for every trip that is susceptible of being made by plane or train, the registers that compound the trip are analysed in order to identify possible stops not detected with the clustering algorithm at the significant transport locations (for example, business travellers may not spend as much time inside the airport as leisure passengers, and they may not have been assigned a stay at the airport).

The third step uses external data sources (e.g., Google Maps APIs data, public transport smart card data) to complete the information and identify the route followed by the different mobile phone users during their trips. The use of complementary data sources is necessary when working with mobile phone data to solve issues arising from the characteristic time and space uncertainty of these data. Such issues are listed below:

- There is uncertainty about the true hour of departure/arrival of a trip. The information provided by mobile phone data consists of the last register at the departure location and the first register at the arrival location, which provides an upper bound of the duration of the trip. In order to correctly determine the true departure and arrival times, first it is needed to obtain an estimation of the true trip duration with an external data source (See Figure 2.2.1.1.4.2). The estimation of the trip duration will vary depending on the mode of transport used and the route followed. We propose an approach that obtains these estimations from Google Maps Directions API for the cases of road, train and public transport, and DDR2 in the case of air transport.

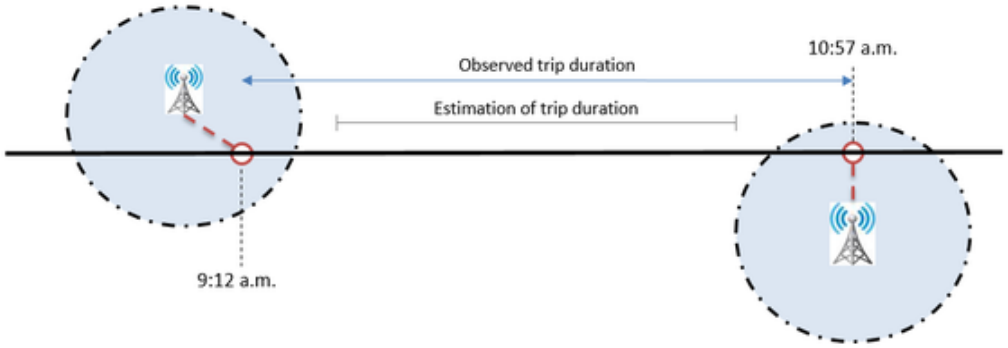

Figure 2.2.1.1.4.2 Uncertainty associated to the departure/arrival times of a trip detected through mobile phone data.

- The observed speeds from the mobile phone data do not correspond directly to the expected speeds for the different modes. Travelled distance calculated as the crow flight distance between the origin and destination towers underestimates the real travelled distance, which highly depends on the followed route. This, together with the longer trip times caused by the uncertainty in the trip departure/arrival time mentioned above, leads to much lower observed trip speed. However, these observed speeds can be pre-processed with machine learning techniques to derive the relationship between the observed speed and some preliminary estimates of the mode, as it can be seen in the Figure 2.2.1.1.4.3. Here transport modes have been estimated from the histogram of speeds for the trips occurring in the Madrid-Alicante corridor.

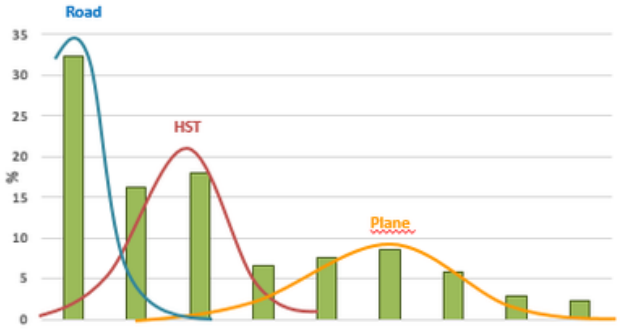

Figure 2.2.1.1.4.3 Estimation of transport mode from a speed histogram. The blue line corresponds to road trips, the red line corresponds to high-speed train, and the yellow line to trips performed by plane.

The process of merging mobile phone data with external data sources is as follows. We use Google Maps Directions API to obtain several route alternatives for each trip detected with mobile phone records. Route alternatives are obtained for the available modes for that trip. The registers left by the user during the trip are compared with the different route options provided by the external data sources. The route best fitting those registers is assigned as the route followed. As routes for different transport modes are typically different, the mode can be directly detected once a route has been assigned. This process is illustrated for the case of the Madrid-Torrevieja (typical touristic location in the South-East of Spain) corridor in Figure 2.2.1.1.4.4. The main routes available between these two locations are presented. The red routes correspond to roads and the green one is the High-Speed Train (HST). In Figure 2.2.1.1.4.4b, the antenna coverage areas associated to each one of these routes are presented. By looking at the registers of a user, and comparing them with the towers providing service to each of the routes, it is possible to distinguish the route followed by the user. Figure 2.2.1.1.4.4c shows an example of the registers left by a HST user.

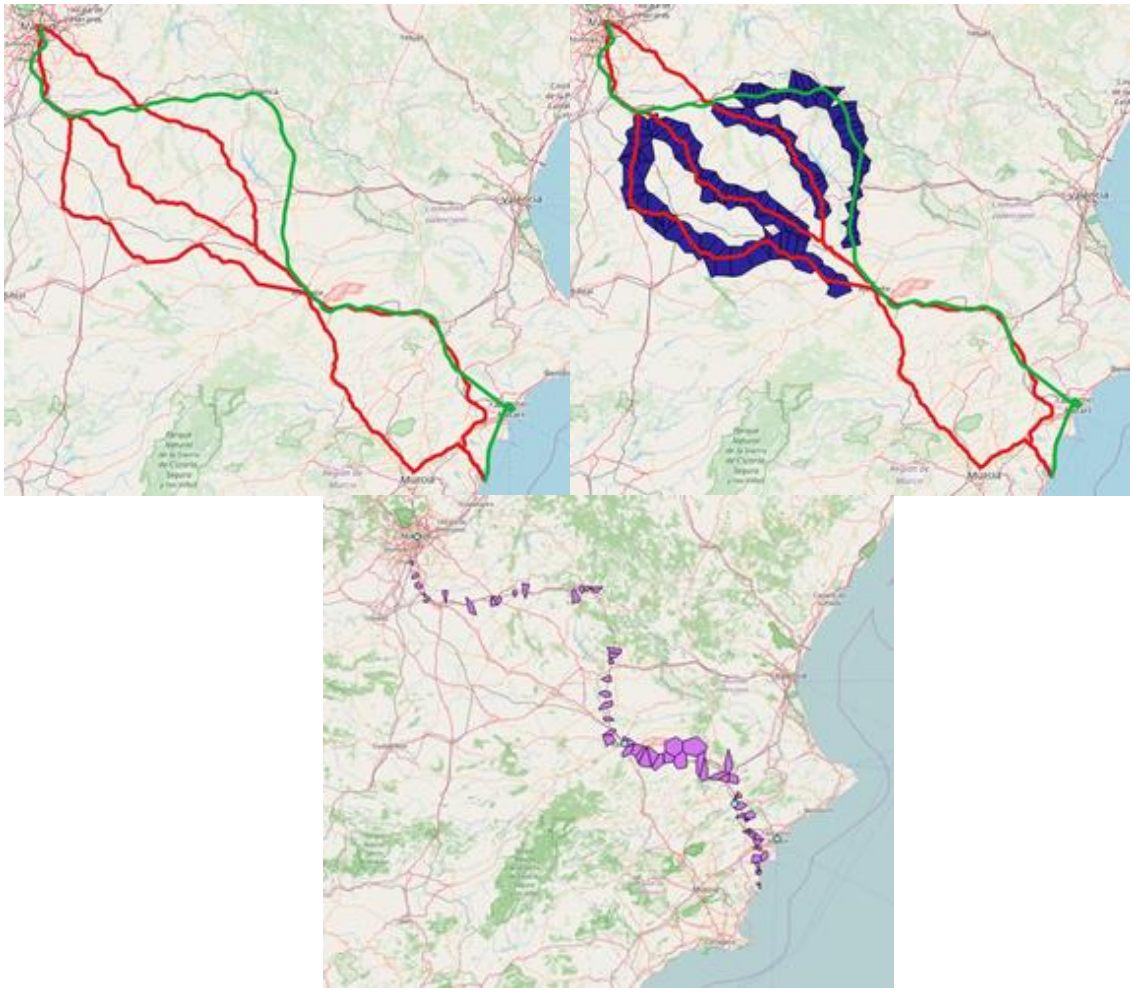

Figure 2.2.1.1.4 Example of identification of transport mode and route by matching registers and existing routes for Madrid-Torrevieja.

2.2.1.2 Mobility information extraction for roamers

Roamers data available from CDRs have much lower temporal granularity. These data do not allow us to capture users at any moment or location during the day. As a consequence, the mobility analysis to be performed for these users should be adapted to the data characteristics. Some examples of the questions that can be addressed are: main entrance port and mode, distribution of tourists according to nationality across the territory (main destination at the level of big cities), presence of tourists in different Points of Interest (PoI), link between different PoI (i.e., list of PoI visited in a single visit to Spain), and length of stay. The questions that cannot be addressed for all roamers include the detailed reconstruction of door-to-door trips, in particular those starting outside Spain.

When roamers enter the country, they have a very different behaviour than nationals in terms of mobile phone use. Some users switch off they phone as soon as they land or few hours later, or they reduce their use, particularly regarding the use of data. Internet access is often switched off due to the its high cost outside the country of residence, drastically reducing the temporal resolution of records. This improved for EU residents' data obtained after the 15th of July 2017, when mobile roaming fees were abolished in the EU. Another issue encountered with CDRs coming from roamers is that, while abroad, roamers are not always connected to the same network. This may create the false impression of multiple visits to the country performed by the same user, when actually it is a user that disappears from the network (by getting connected to another) and then reappears a few days later. To overcome the aforementioned issues, users are characterised according to their activity and only the most active users are kept for the analysis.

2.2.1.2.1 Data pre-processing and users' characterisation

Users are classified into: i) tourists: visitors spending at least one night in the country; and ii) excursionists: visitors not spending a night in the country. The classification is not as straightforward as it may seem, as some excursionists visiting regularly the country may be misclassified as tourists, while some tourists not having much mobile phone activity may be misclassified as excursionists. To solve this, the following criteria are used:

- Those visitors having registers in more than one consecutive day and at least one register at night time, i.e., in the interval 23:00-07:00, are considered to be tourists.
- Those visitors having registers in more than one consecutive day but without any registers observed during the night time are considered as tourists if they are found at a distance X from the frontier. Different values of X have been tried (10km, 20Km and 50km), being 50 km the one providing the best results. Visitors found at a shorter distance than 50km appearing on consecutive days but without night registers are considered excursionists.

Another important issue to consider is to distinguish between single trips and multiple trips. Multiple trips may be confused with long stay trips. To solve this issue, the following criteria are taken into account:

- If the maximum time between two consecutive calls is greater than 3 days, the visit is considered as two (or as many 3-day absences are found for a single user) different trips, with each trip ending the last day before the 3-day gap and the next trip starting the first day after the 3-day gap.
- If the user shows activity with a gap of less than 3 days but there are not any consecutive days with activity in the period of the visit, some extra criteria are considered: if no night registers

are found in the whole period, the visit is split into different trips. For example, let's consider a user with events registered on Monday, Wednesday and Friday. If no night registers are found in the period from Monday to Friday, then the visit is split into 3 different trips of one day. However, if at least one night register is found, then the whole period is considered as a single visit. If, for instance, a user shows activity on Monday, Wednesday, Thursday and Friday but no register is found at night, the visit is still considered a single trip because in the whole period there is not a gap of 3 days and there are two consecutive days with activity.

2.2.1.2.2 Sample selection

A maximum time X between events is required during the whole day for a user to be considered valid. This criterion is applied differently to tourists and excursionists. For excursionists, the maximum time criterion is applied to the whole period of the visit, i.e., from the first register to the last register. For tourist, the maximum time between events is required for the whole day, where the definition of 'whole day' depends on the day of the trip. For the first day of the trip, the day is considered to start with the first register (arrival time). For the last day of the trip, the day is considered to end at departing time (time of the last event registered). For the intermediate days of the trip, a minimum time between events is required during the whole day, with the day "starting" at 10:00 and "ending" at 22:00. The value of X is the same for both tourists and excursionists. Tests have been performed for different values of X, and X= 4hrs is the one giving better results when comparing with official statistics (see section 2.4). After keeping only valid users, with the restriction of having at least one register every 4hrs, the sample represents a 24% of the total number of visitors entering the country in the period considered when compared with official statistics.

2.2.1.2.3 Estimation of length of stay

Once multiple trips are distinguished from single trips, the estimation of length of stay is straightforward. The length of stay is calculated as the time (in days) elapsed between the first and the last register of a single trip. To compare the results obtained from mobile phone data with official statistics, the length of stay is grouped into five categories: i) 0 nights (excursions); ii) 1 night; iii) 2-3 nights; iv) 4-7 nights; v) 8-15 nights; and vi) more than 15 days.

2.2.1.2.4 Main destination

To estimate the main destination for each user, we count the total number of days spent at each destination at the level of Spanish Autonomous Communities (CCAA), to be able to compare and validate the results with official statistics. The CCAA having the larger number of days with associated night time registers is considered the main CCAA of destination. If two or more CCAAs have the same number of days with night registers, the day time registers are considered and the CCAA with more days with associated daytime registers is considered the main CCAA.

2.2.1.2.5 Tourists presence in PoI

The presence of tourists in different areas of a city/region can be estimated by counting tourists in predefined zones of interests at different times of the day. If one user is found to be in more than one zone of interest in the same time interval, his/her presence is accounted for as a fraction of the number of zones in which it has been registered.

2.2.1.2.6 Entrance port

By registering the first position in Spain where a non-resident is detected, the entrance port can be inferred. To detect the entrance port, we compute the distance between the position of the first register of the trip and all frontiers, ports, and airports. The alternative with the shortest distance to the first registered event is taken as candidate entrance via. If such distance is shorter than a given threshold X, the port is considered as the entrance via, else the entrance via is set as unknown. Different values of the distance threshold were tested, being 5 km the one providing the best results.

In some cases, the entrance points differ considerably from the final/main destination, especially for intercontinental trips were different airports serve the same final destinations. The identification of entrance points for a given final destination may help to point out possible competing airports. Figure 2.2.1.2.6.1 shows the entrance points for Andalucia, Castilla y Leon and Valencia Regions. Location of the different Airports is shown in Figure 2.2.1.2.6.2.

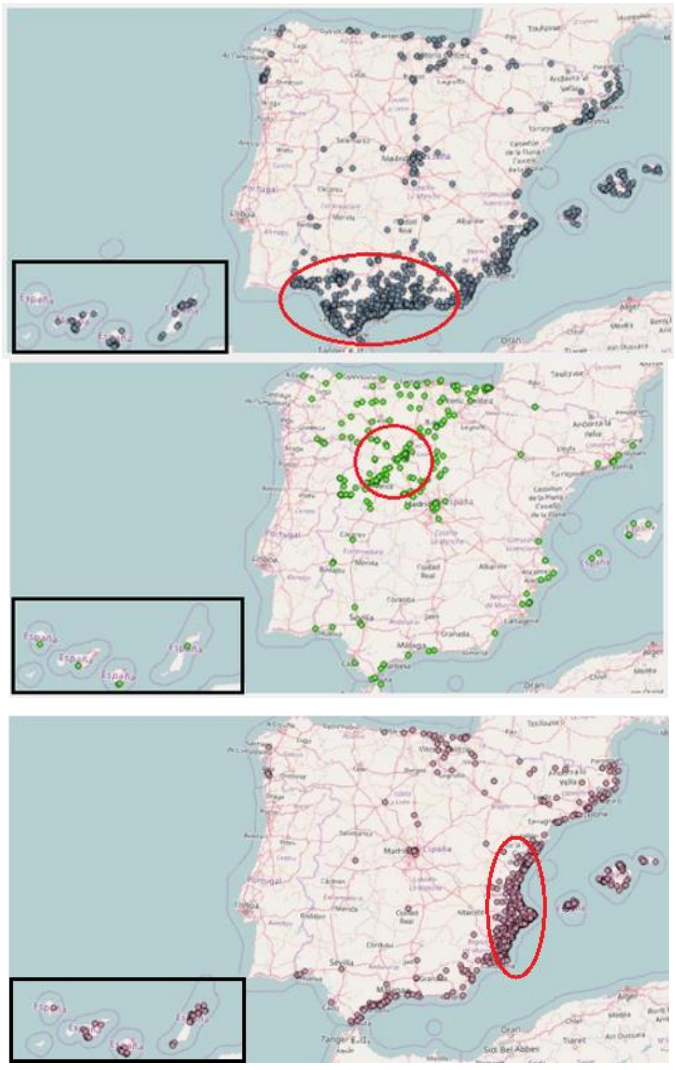

Figure 2.2.1.2.6.1 Entrance port for visitors with final destination at a) Andalucia, b) Castilla y Leon and c) Valencia. Relevant Autonomous Communities are highlighted in red circles. Canary Islands (in black rectangle) have been moved for visualisation purposes.

Figure 2.2.1.2.6.2 Location of main Spanish airports.

2.2.2 Twitter data

2.2.2.1 Places

The geographical information of Twitter data is provided essentially at two levels. The first one includes the coordinates of the user at the moment of posting the tweet with the usual geolocation uncertainty of a few tens of meters. The second level is the so-called place field. As already explained, this second convention is more abundant since the change in the Twitter policy regarding location self-reporting. The place information can be associated to different geographical levels. These types of places are not defined and used homogeneously in all countries (see Figure 2.2.2.1.1). In some cases, only information at the level of country or administrative area is available.

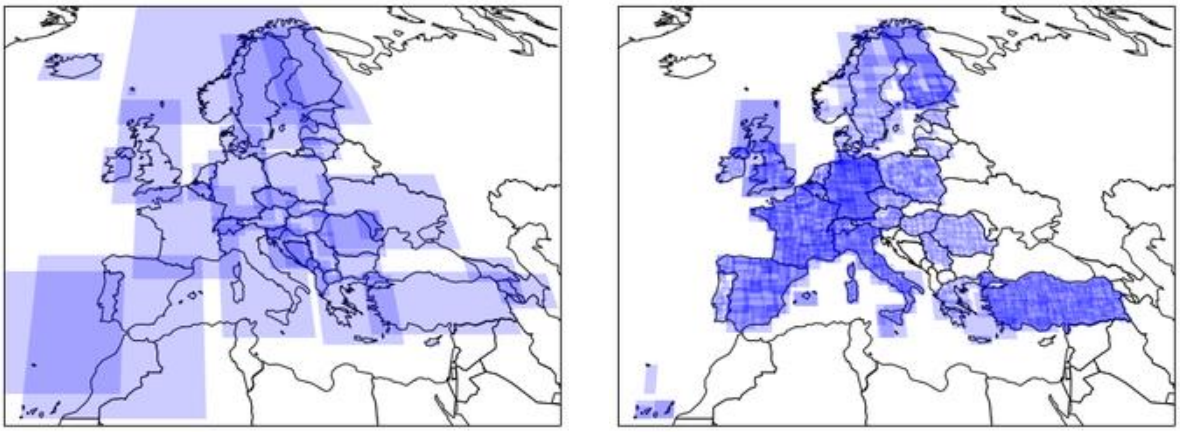

Figure 2.2.2.1.1 Map of the bounding box of Countries (left) and Administrative areas (right) used as 'places' in the Twitter database

In general, we will use the geographical information of the place when the coordinates are not present and if the type provided is sufficiently precise. This restricts to the three cases (cities, neighbourhoods and POIs) pictured below, plus a few metropolitan cities that appear as Administrative areas. The place type called 'city' is reasonably well spread in Europe. However, it is not present in some countries in the Balkans and in Eastern Europe. Neighbourhoods and points of

interests are defined in a limited number of countries: noteworthy neighbourhoods are defined across all the Netherlands and Turkey, and POIs are widespread in the UK and Germany. Besides this, in some cases neighbourhoods and POIs are defined for a limited set of cities.

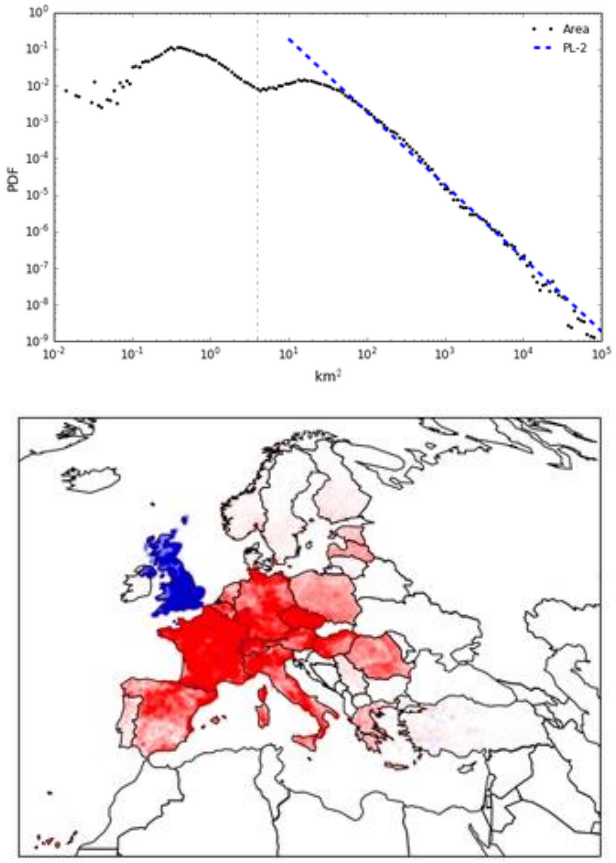

Figure 2.2.2.1.2 (Up) The Distribution of the areas of the cities defined as Twitter 'places' in Europe manifestly follows a bi-model. The two peaks identify two types of cities: small (< 4km²) and large (> 4 km²). The threshold is represented here by the vertical dot-dash grey line. The scaling of the tail is consistent with a space filling tiling. (Bottom) we represent in red the location of larger cities and in blue the smaller ones, which appear to be concentrated in the UK.

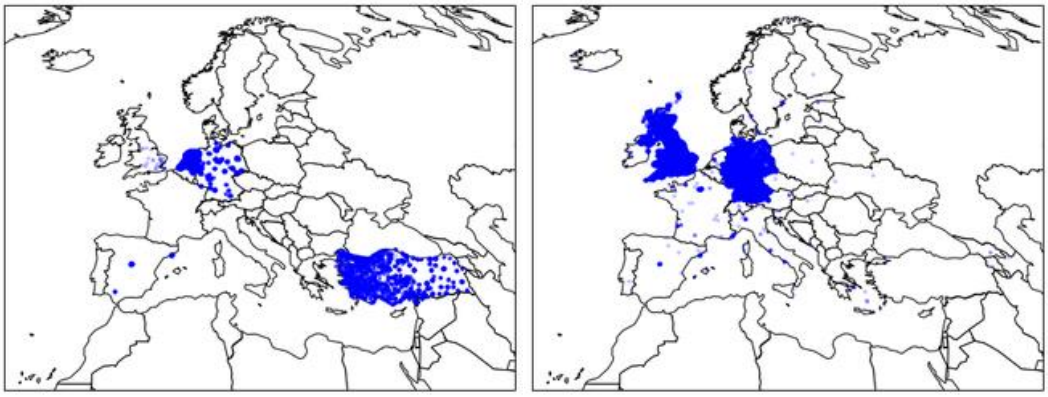

Figure 2.2.2.3 Map of the places belonging to the neighbourhood category (left) or to the POIs category (right).

2.2.2.2 Filtering the user database

Not all the 9.8 million users we observe in the selected area provide valid information for our purposes. In this section, we illustrate the series of filters needed to reduce biases that might be introduced by tracking trajectories too short to be correctly characterised, as well as by Twitter accounts that should not be associated to individual movements.

Number of Tweets. There is a large fraction of users that do not use Twitter frequently enough for us to identify their home location. Understanding a user's residence area is a necessary passage in order to upscale all the statistics we derive from Twitter to the total population (see section 2.3). Varying local socio-economic characteristics make the upscaling constants significantly different among different countries as well as among different areas of the same country. We therefore select only users with a number of tweets allowing us to reasonably associate them with an area identified as *home*. This selection allows us to limit the noise that would otherwise be introduced by misplaced users, and at the same time reduces the memory and computational time required for the analysis. In Figure 2.2.2.2.1, we show the Cumulative Density Function (CDF) of the number N of tweets per user. We decided to include in our analysis only the 45% of the users that wrote more than 10 tweets.

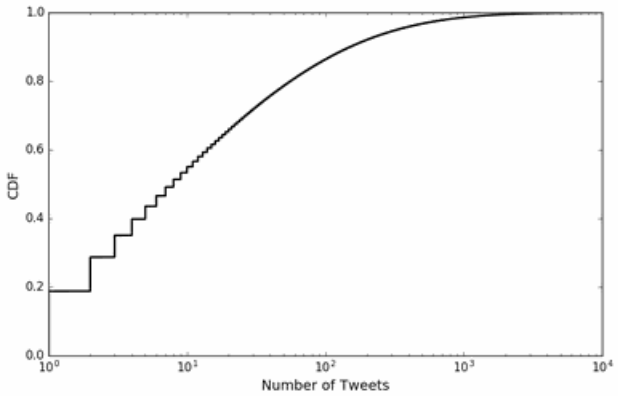

Figure 2.2.2.2.1 Cumulative density function of the number of tweets per users

Interval of observed presence. After selecting the users who wrote a sufficient number of tweets, we control for the time span in which these messages have been made. If the first and last tweets are too close to each other, it could be that we are only tracking the movement of an intercontinental tourist during a trip. In that case, our procedure for assigning him/her to a specific home area of Europe for the upscaling would, again, introduce errors. For this reason, users observed for a time interval shorter than three months have been filtered out. As we can see in Figure 2.2.2.2.2, this further excluded 20% of the users with more than 10 tweets (corresponding to 9% of the total).

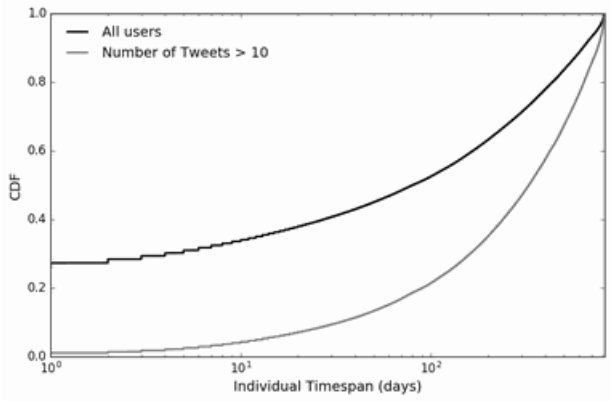

Figure 2.2.2.2.2 Cumulative density function of the number of days between the first and last tweet recorded.

Bots and multi-accounts. The Twitter user database also includes accounts that must be excluded because the locations registered are not associated to the movement of a (single) person, such as accounts shared by more than a user, or accounts automatically handled by bots. We identify all these by analysing the inter-event time between subsequent tweets (*dt*) and the associated observed displacement (*dr*), which are displayed in Figure 2.2.2.2.3.

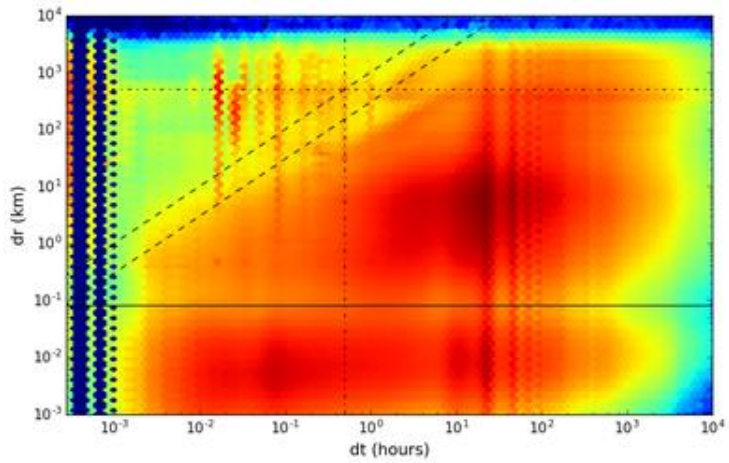

Figure 2.2.2.2.3 Distribution of the inter-event posting time in seconds (x axis) and displacement between recorded coordinate positions in kilometres (y axis). (Left) before bots filtering. (Right) after bots filtering.

The inter-event time distribution shows remarkable peaks at regular times, in particular multiples of 60, 90 and 300 minutes. This is caused by bots posting messages with regularity because they follow the instruction of a script, as illustrated in Figure 2.2.2.2.4:

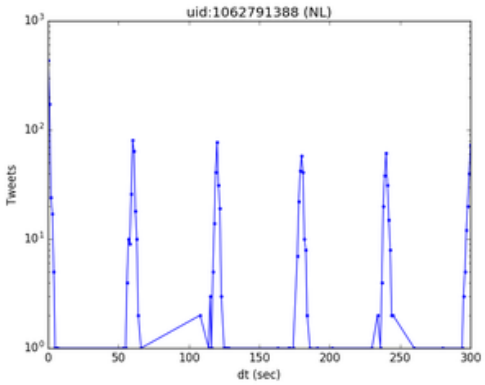

Figure 2.2.2.2.4 Number of tweets as a function of time for a bot geographically based in the Netherlands, with a periodicity of 1 minute.

It is straightforward to automatically and consistently isolate this kind of users. The procedure developed to do this consists in counting the number n_1 of inter-event times that is a multiple of 60 seconds, with a tolerance of 2 seconds. We notice that the total number of tweets N observed for a fixed n_1 is distributed approximatively as a lognormal (Figure 2.2.2.2.5). This allows us to use a consistent statistical criterion for excluding all users that have a value of N too small to have produced the observed number of inter event times at multiples of minutes.

More precisely, it starts by defining a threshold probability $p = 1/U(n_1)$, where $U(n_1)$ is the number of users with a particular value of n_1. Then the lognormal distribution $P(n|n_1)$ is fitted for each value of n_1, identifying the parameters μ and σ of the lognormal. The lognormal model allows us to associate to p a number s of log-standard deviations σ at which it is expected not to observe any data point. Any value of N such that $(\log(n) - \mu)/\sigma < s$ is thus unlikely to appear in the sample at hand and identified as bot.

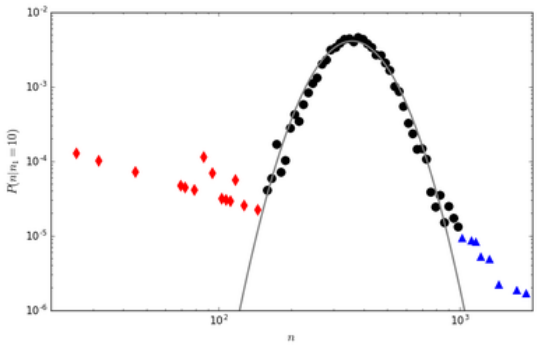

Figure 2.2.2.2.5 Criterion used to identify bots. In black the values of P(n|n1=10) comprised between s log standard deviations. The values in blue are users who tweet less than what would be expected with dt = 1 minute, possibly because of a very bursty behaviour where typically dt < 1 minute. The values in red are those who have more events than expected with dt = 1 minute, and are therefore excluded from our analysis.

This method works well for n1 < 100 (red in the figure below). For n1>100 the number $U(n_1)$ is not sufficiently large to correctly fit the lognormal in the data biased by the bots. Fortunately, the lognormal criterion highlighted a clear visual pattern in the joint distribution $P(n_1,n_{All})$, so we take advantage of it to proceed with the bot identification for large values of n_1 (green in Figure 2.2.2.2.6). The same criterion is used also for periodicities of 90 seconds and 5 minutes.

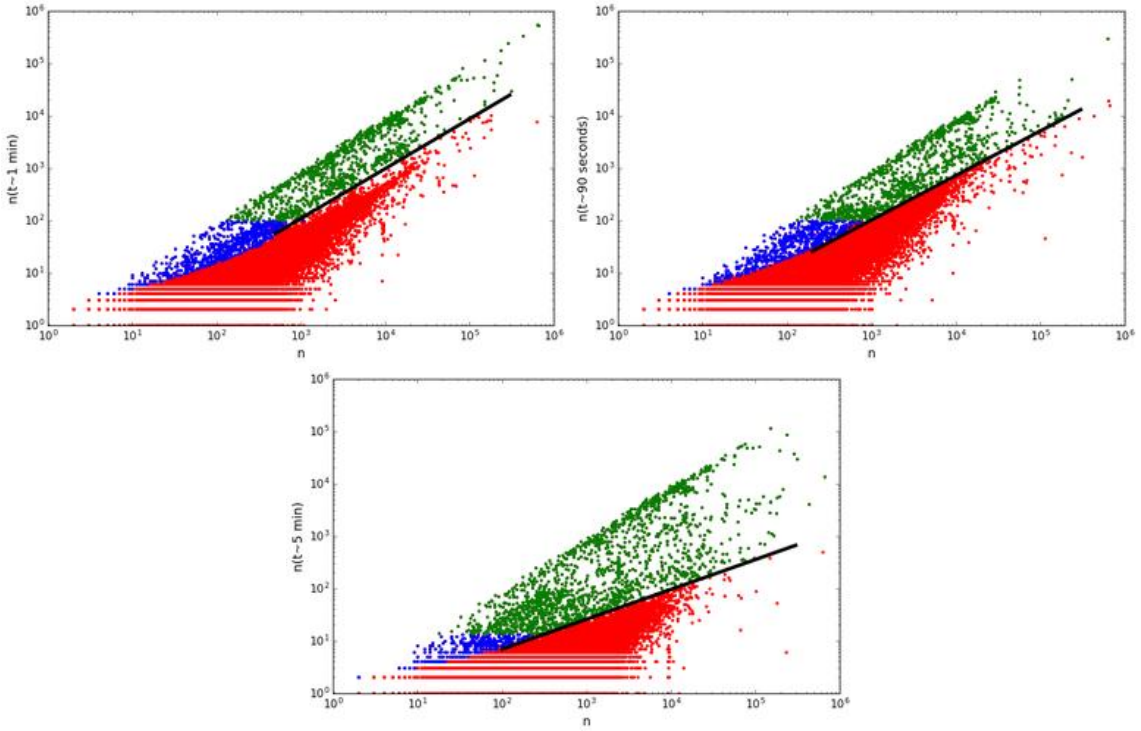

Figure 2.2.2.2.6 Criterion used to detect bots. In red the users kept, in blue those excluded with the lognormal criterion, in green those isolated following the visual pattern suggested by the lognormal criterion. (Up-left) bots set for tweeting after multiples of minute. (Up-right) bots set to tweet after multiples of 5 minutes. (Bottom) bots set to tweet after multiples of 90 seconds.

With this filter, we can identify a small number (~2500) of bots that are, however, responsible for a large fraction of tweets. A visual inspection shows that the distribution of inter-event times after filtering does not present the aforementioned peaks.

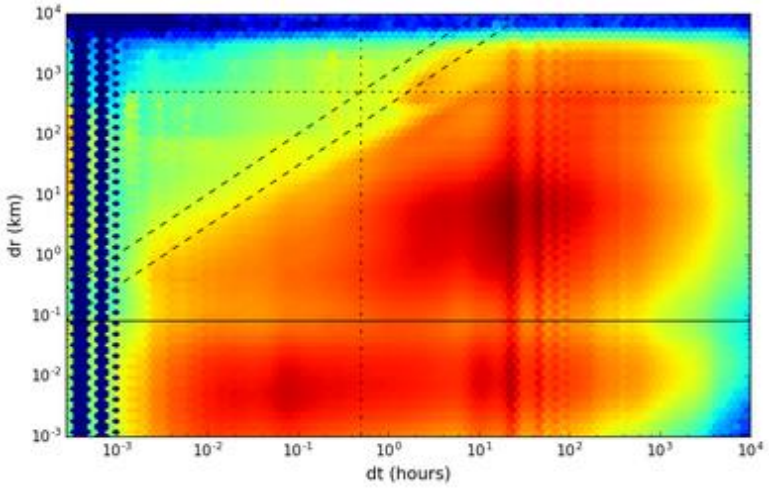

Figure 2.2.2.2.7 Inter-event time / displacement joint probability distribution for consecutive tweets with coordinate information after filtering potential robots.

Bots can be also identified simply by the excessively high tweet frequency. This can be spotted in two ways: i) by evaluating the number of tweets day by day and excluding accounts that in a particular day wrote tweets with an average frequency higher than 5/minute; and ii) by identifying accounts that produced at least a single couple of tweets in the same second. More than a single tweet at the same time might also happen when the same account is handled by different (human) users. We want to exclude also this case, which would clearly introduce in our data trips that never occurred. We do this by identifying accounts that produced at least a single couple of consecutive tweets that are separated by a distance dx > (dt × vmax). A maximal speed vmax = 1000km/h is fixed as an upper bound that cannot be reached even in the case of a flight. If two tweets by the same account are separated by a time dt and a distance dx such that the observed speed is higher than 1000km/h, it is impossible for the two communications to have been made by the same person travelling and so it is excluded.

2.2.2.3 Identifying users home approximate location

We can easily associate a user to a single country where he/she tweeted the largest fraction of times. The number of users observed is naturally expected to depend on a country's population size.

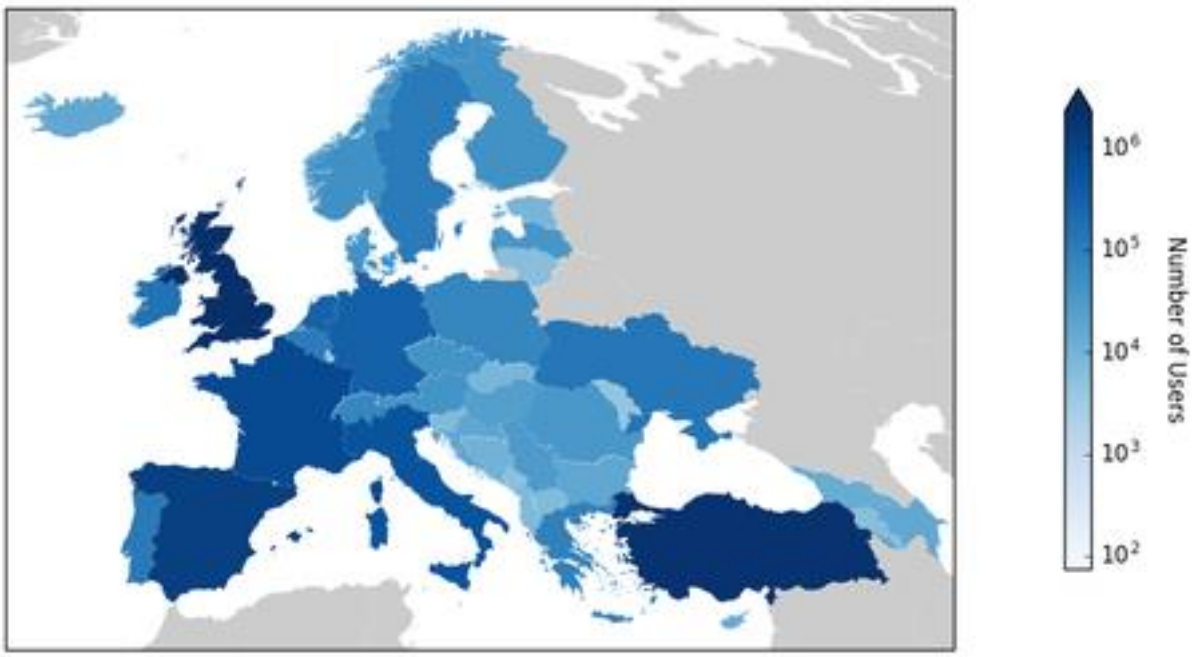

Figure 2.2.2.3.1 Heat-map with the number of users left after filtering per country.

However, it is not a simple proportionality, since Twitter has a different market penetration depending on the area. We define this market penetration M = users/population for each country (see Figure 2.2.2.3.2). We will use the information on market penetration at a NUTS3 level to upscale the flows reconstructed from Twitter data to population level (see Section 2.3.2).

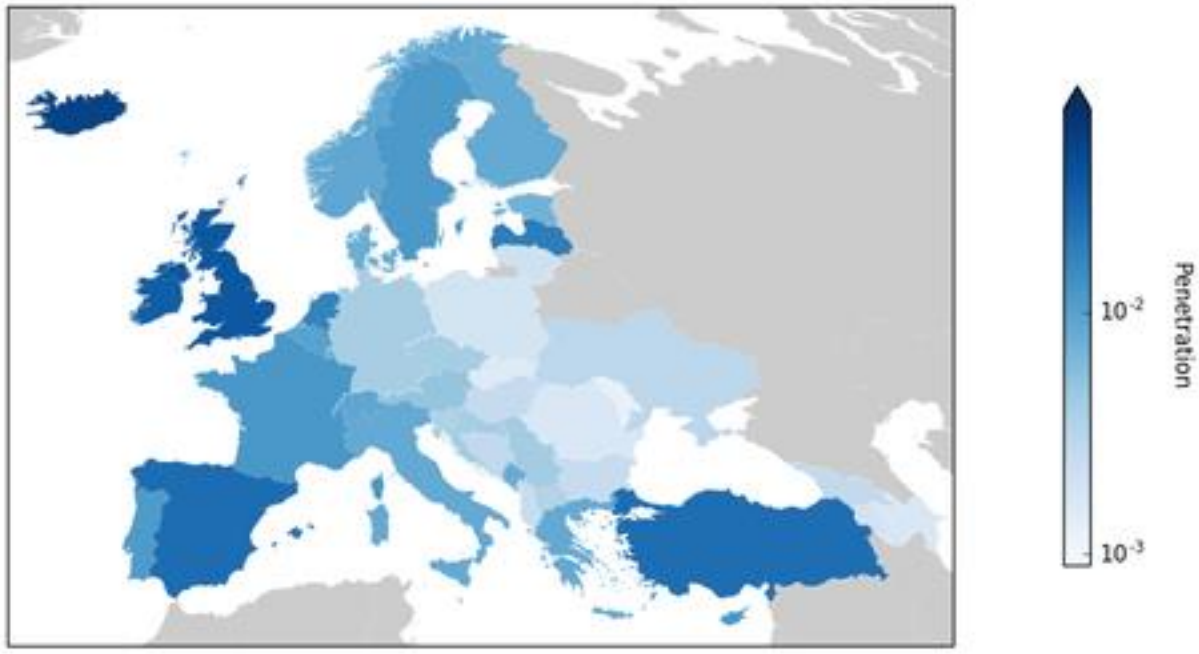

Figure 2.2.2.3.2 Heat-map with the geolocated Twitter user penetration per country.

2.2.2.4 Domestic users, international travellers, and migrants

About 80% of users tweet only from a single country. We call these domestic users. Naturally, the fact that we observe a user tweeting from different locations or countries can be associated to international travellers of different nature. We distinguish in particular on the basis of the length of stay into two categories: short term movements (tourism, business trips) and long-term movements. We consider as long-term movements those associated to stays longer than three months, and will include both temporary stays (3-12 months), which can be typical e.g. for students in exchange programmes or seasonal jobs, and prolonged stays (> 12 months) that will be identified as international migration. We detect these long-term movements by identifying the country where a user tweeted the most month by month. If we observe users in the same country for 3 consecutive months (eventually interrupted by lack of records), we exclude the user since we cannot univocally associate his/her to a single home country. About 10,000 expats have been isolated in this manner from the dataset.

2.2.2.5 Movements associated to air travel

As illustrated above, we used the displacement (dr) and inter-event time (dt) to characterise multi-user accounts for which we observe “fake trips” faster than an airplane’s maximal commercial speed. This speed is illustrated in Figure 2.2.2.2.7 and Figure 2.2.2.5.1 as the upper of two dashed lines. Figure 2.2.2.2.7 represents displacements observed with the information collected in the coordinate field of the tweets, while Figure 2.2.2.5.1 here below shows displacements observed with the place information.

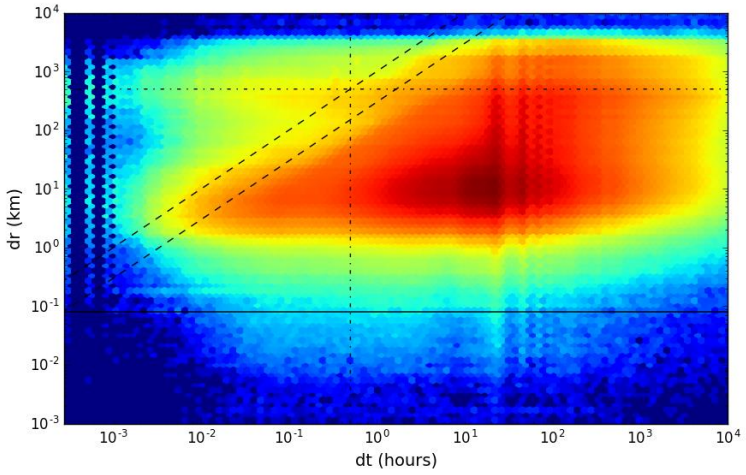

Figure 2.2.2.5.1 Inter-event time / displacement joint probability distribution for consecutive tweets with only place information in one or both tweets, after filtering potential robots.

The lower dashed line represents instead the fastest speed of high-speed trains. The trips comprised between two thresholds, and with dt > 30 minutes, represent movements that are necessarily associated to air transportation, since no other mode of transport would allow to be faster than 300 km/h. We will use this subsample, which includes approximatively 0.5% of the trips observed, as natural control sample for any result obtained from the whole dataset. Since people do not necessarily tweet exactly at the moment when the trip ends, we can eventually try to expand this subsample by progressively lowering the threshold, while tracking an aggregated quality parameter of comparison with ground truth flight statistic to see at which point we start introducing too many errors.

2.3 Upscaling of mobility information

2.3.1 Mobile phone data

As the project has only access to mobile phone data for Spain, only a sample of the Spanish population and non-Spanish visitors is covered by this dataset. The mobility information extracted from CDRs needs to be upscaled to the total population in order to produce meaningful results. Different methodologies are applied depending on whether we are working with Spanish residents or with roamers, as the information available for these two groups is different.

2.3.1.1 Spanish residents

In this case, we have plenty of information available for the scaling process:

- Through mobile phone data, as explained in the previous section, the home area of the user can be determined.
- The Orange client database contains information about age and gender for each mobile phone user.
- As presented in D2.1, the Spanish Statistical Institute (INE) publishes a population census every 10 years.

A fast (but naive) approach would be to simply apply an expansion factor derived from the market share of Orange. However, this approach does not take into account the non-homogeneity of Orange market share across the Spanish territory. We follow a home-based scaling approach at a census tract level. The home area of the users is obtained at the level of mobile phone network antenna coverage areas. These coverage areas are approximated through a Voronoi tessellation, following the assumption that a mobile phone will always connect to the closest tower. Therefore, as the home area for each mobile phone user is one of these Voronoi Polygons, it is needed to transform these home areas from the Voronoi tessellation to the zoning used in the census data. Once each mobile phone user home area has been assigned to a census tract, we can use the information from the census to apply an expansion factor for each mobile phone user.

Figure 2.3.1.1.1 Sample size across the Spanish territory.

2.3.1.1.1 Transformation of Home area from Voronoi to Census

In Figure 2.3.1.1.1.1, it can be seen an example of both Voronoi areas and census tracts in Madrid's city centre. From this picture, it can be seen that two zonings do not match at all. Typically, the same Voronoi will contain several census tracts, leading to some uncertainty about which one is where the mobile phone user really lives. Additionally, the same census tract may be fed with mobile phone users from two or more Voronoi areas, which, if the transformation is not carefully done, may lead to erroneous results (e.g., census tracts with more mobile phone users than people officially living there, or empty census tracts).

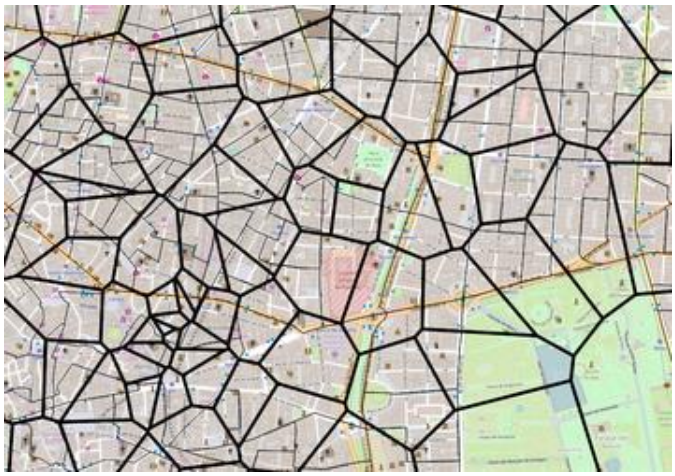

Figure 2.3.1.1.1.1 Comparison of Voronoi areas with census tracts

Some assumptions are made during this zoning transformation step:

- The market share for the mobile phone operator is similar between adjacent or close census tracts.
- As the real coverage areas of any tower may overlap with the towers close to it (phenomenon that is not captured with the Voronoi approach), the census tracts to be considered as home candidates will be the ones covered by the home Voronoi and adjacent Voronoi cells.

The process of transformation is done probabilistically, by taking into account the total population of each candidate census tract and the already assigned mobile phone users to each one of these census tracts. Different probability functions and assignment methods were tested and the method resulting in the most homogeneous distribution was selected. To measure the homogeneity of the sample resulting from each method, the following metric was derived:

$$d = \frac{\sum_{i=0}^{n} \frac{\sum_{j=0}^{m} (p_j - p_i)^2}{m}}{n}$$

with n the number of census tracts of the area of study, m the number of census tracts adjacent to tract i, p_i the sample percentage in census tract i, p_j the sample percentage in the adjacent tract j. The assignment method reporting the best results is the one based on a user by user assignment to any of the census tracts intersecting either the 'Home' Voronoi or the adjacent ones, with a probability

$$p_{u \to i} = \left(p_{real_i}\right)^2 / p_{assig_i}$$

where $p_{u \to i}$ is the probability of assigning a mobile phone user u to census tract i, p_{real_i} is the population of the census tract I, and p_{assig_i} is the number of mobile phone users already assigned to that census tract. The objective behind this formula is to homogenise the sample size between close census tracts. The spatial distribution of users after assignment for the different tested methods is presented in Figure 2.3.1.1.1.2.

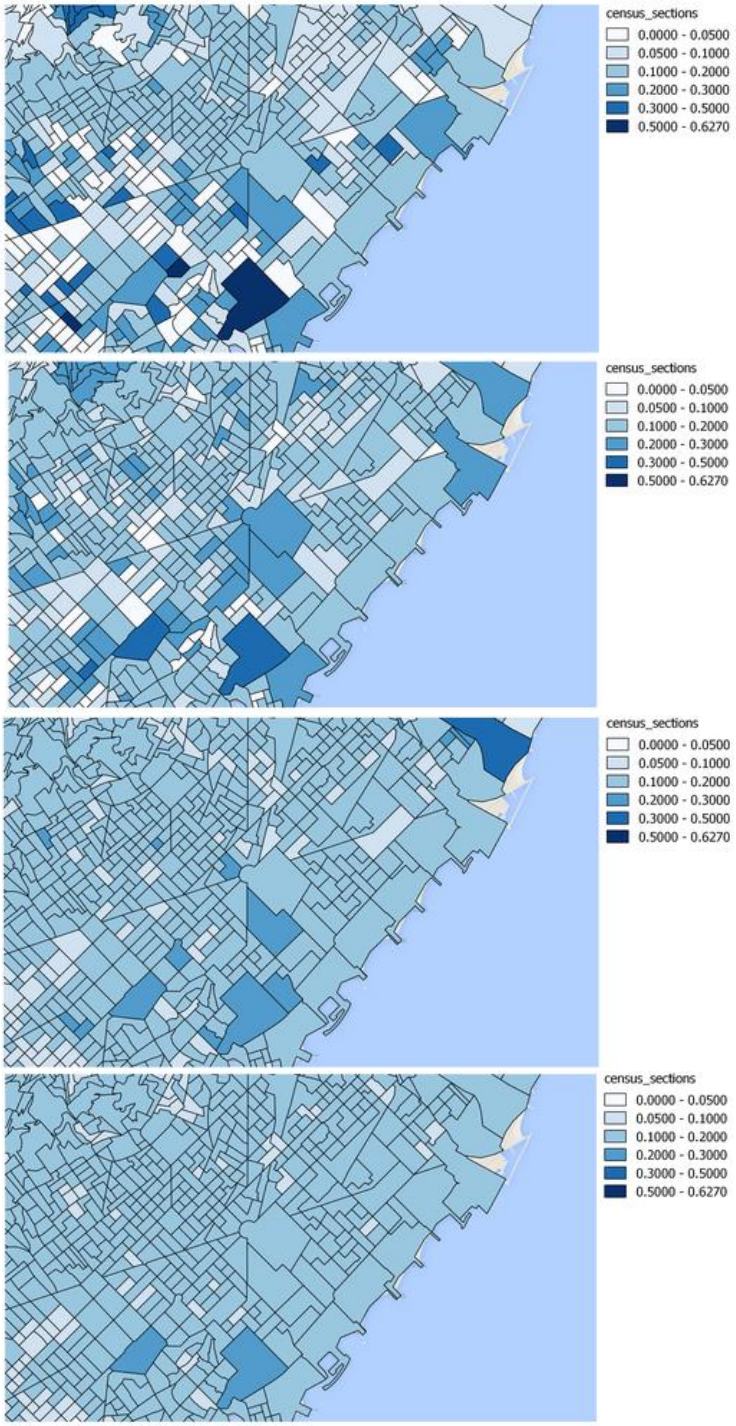

Figure 2.3.1.1.1.1 Example of spatial distribution of users' area around Diagonal Avenue in Barcelona. The colour scale corresponds to sample penetration in the different census tracts. From top to bottom: First two methods assign all users of a given Voronoi u area to census tract $c_i$intersecting it with a probability a) (up) $p_{u \to i} = p_{real_i} - p_{assig_i}$ and b) (second from above) $p_{u \to i}=p_{real_i}/p_{assig_i}$ (for a randomly ordered Voronoi areas). The last two follow a user by user assignment approach, considering the census tracts intersected by adjacent Voronoi areas, with a probability c) (third panel from above) $p_{u \to i}= p_{real_i}/p_{assig_i}$ and d) (bottom panel) $p_{u \to i}= (p_{real_i})^2 /p_{assig_i}$.

2.3.1.1.2 Expansion methodology from census data

Once the home Voronoi area for each user has been assigned to a census tract, the upscaling factor for every user can be calculated. This can be done in two ways: obtaining a factor based on the total population, or using gender and age information from the mobile phone sample and census data. The main problem with the second method is that gender and age information is not available for every mobile phone user, therefore if we want to use this information the sample of mobile phone users gets reduced to approximately half of its original size. Also, for some census tracts there might not be any sample of a certain gender-age group, and therefore it is needed to "borrow" sample users from neighbouring census tracts. This results into higher expansion factors than expanding by totals, and therefore it should be used only in applications where obtaining a highly detailed population profile is more important than trying to capture all the mobility patterns present in the sample.

2.3.1.2 Roamers

In the case of roamers, the only sociodemographic information available is the place of residence, which can be taken as a proxy of the nationality of the user. Since roamers' mobility information is only available for Spain (phone events occurring in the roamer's place of residence are not available), home location cannot be inferred and hence the expansion method based on census data applied for nationals cannot be used. However, as explained in the previous sections, characteristics such as length of stay, main destination, type of visitor and entrance port, commonly reported by tourist statistics, can be obtained from mobile phone activity.

To up scale the sample, microdata of visitors crossing the frontiers (non-nationals entering Spain) provided by the National Institute of Statistics (INE) have been used. The information of the microdata is aggregated at a monthly level and it contains: nationality, type of entrance mode (train, boat, road, airplane), length of stay, main destination (at the level of Autonomous Community CCAA, which corresponds to NUTS 2), type of visitor (tourist spending at least one night in the country, or daily excursion). This allows us to compute the number of visitors sharing the previous characteristics in the given month.

Different tests performed with the sample data showed that expanding by nationality and type of visitor (tourist and excursionist) reports better results, rather than expanding by total number of visitors entering the territory, nationality, main CCAA of destination, length of stay or port of entrance (see section 2.4).

For the expansion process, visitors are first classified into tourists and excursionists. For each nationality, a different expansion factor is calculated for each visitor type. This factor is given by the number of visitors of a given nationality and type reported by the INE divided by the number of those with the same characteristics (nationality and type) in the sample.

2.3.2 Twitter data

2.3.2.1 Home place and penetration at NUTS3 level

Twitter's market penetration shows significant differences even within the same country. For this reason, for having a more precise upscaling procedure, we want to identify the home location at a smaller scale. We decided to use the statistical level NUTS3, corresponding to Provinces (ES and IT), Kreises (DE), Departements (FR), etc. If coordinates are not available, we use the smaller place categories (city, neighbourhood or POIs) as geographical information to identify the NUTS3 area from where the tweet has been written. The 'place' geographical information is a bounding box, which can in principle overlap with more than a single NUTS3 area (see the example below in Figure 2.3.2.1.1: in red the bounding box of a city lying on the east side of a river that represents the border between two areas in UK). In that case, each place is associated to the NUTS3 area with the largest overlap (in the example, it would be the area on the right).

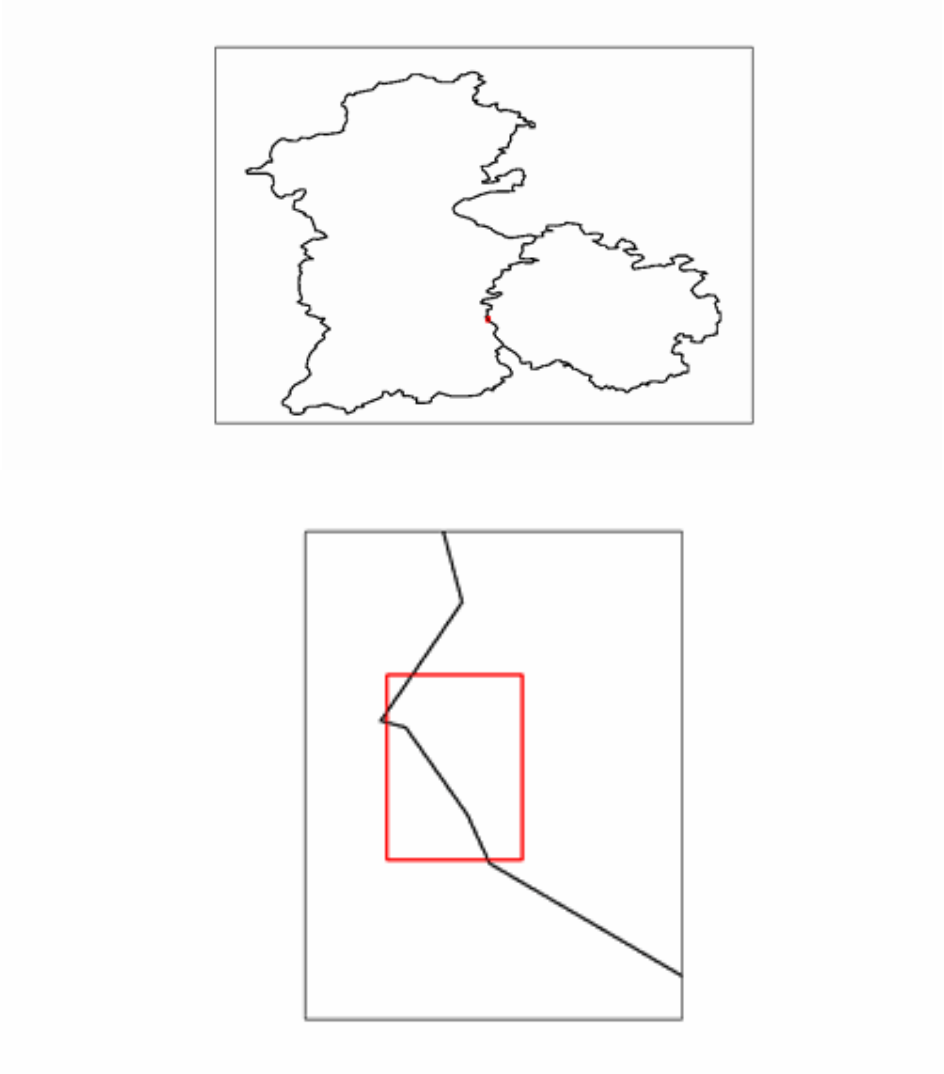

Figure 2.3.2.1.1 The association of a place with a NUTS3 area in case of overlap with the border between the areas.

Moreover, in some countries (e.g., Germany) metropolitan areas are described as NUTS3 areas, while in the Twitter place database they are labelled as 'admin', the same categories containing also regions. Therefore, it is necessary to also consider 'admin' places that are in general significantly larger than NUTS3 areas. In this phase, we proceeded with the association only if the NUTS3 surface represented at least 20% of the 'admin' place bounding box area.

We can compute the Twitter user-base penetration rate at NUTS3 level, with a few exceptions: i) some eastern countries are not described in the NUTS database; ii) some small countries, such as Iceland or Cyprus, are only associated to a single NUTS3 area. In this case, the whole country is necessarily used as the considered area. This reconstruction of the penetration of Twitter users at NUTS3 level allows us to take into account the coupling between different habits in the sharing of geolocated tweets and different travel behaviour between different regions of the same country. The final results are displayed in Figure 2.3.2.1.2 below.

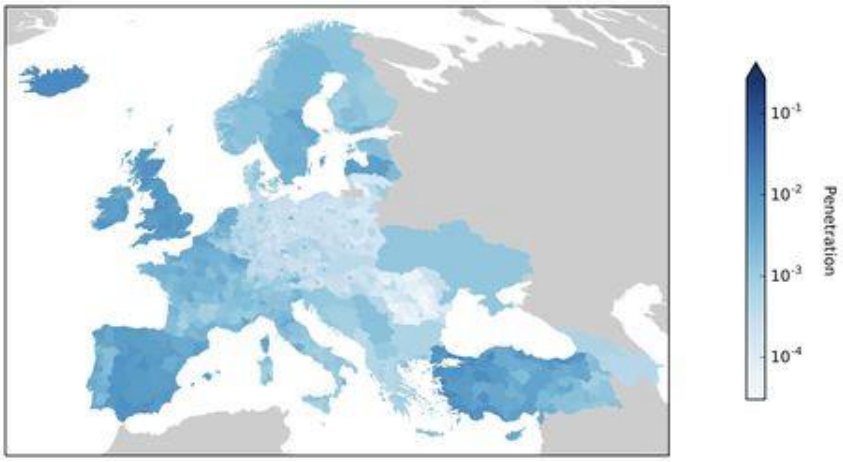

Figure 2.3.2.1.2 Map of the Twitter penetration rate in the ECAC countries at NUTS3 level.

2.3.2.2 Upscaling

All estimates made on the basis of the Twitter data must be correctly upscaled in order to provide statistics that could describe the real characteristic of the air travel in Europe. For doing this, each Twitter user belonging to an area characterised by a population *Pop* and a number *U* of users, will be considered as representing *C(A) = Pop(A)/U(A)* citizens of that same area. For instance, when reconstructing flows observed between the area O and the area D, we sum the contributions from users living in all areas *A*. The contribution of the area *A* (that might be eventually =O or =D) will be the trips *T* between O and D of the users living in A multiplied by the factor *C(A)*

$$F(O \to D) = \sum_{A} T(O \to D, A) \cdot C(A)$$

The upscaling is in principle dependent on two factors: i) how we choose to divide the area of analysis into the set of subareas *A*; and ii) how we estimate the number of users *U(A)*. Here we evaluated different alternative choices:

- **Euro** - we consider Europe as a whole and multiply all flows by the factor C given by the ratio of the total population in the ECAC area divided by the number of Twitter user in our database.
- **Country** - we divide Europe at country level and count for each country the number of users observed separately for each month of analysis.
- **N3** - we divide Europe at NUTS3 level and count for each area the number of users observed separately for each month of analysis.
- **N3 aggr** - we divide Europe at NUTS3 level and count for each area the total number of users observed in all 26 months of analysis.
- **N3 thres** - we divide Europe at NUTS3 level, count for each area the number of users observed separately for each month of analysis, and select only the areas where the penetration rate is above a given threshold, thus avoiding the biases caused by a single user representing a too large fraction of the total population.

The Euro upscaling is clearly the naiver. Nevertheless, a preceding work [EPJ Data Science 5, 30 (2016)] implemented this approach successfully for the reconstruction of US travel from Flickr data. As can be observed in Figure 2.3.2.1.2 above, the use of Twitter in Europe is however extremely heterogeneous, and is therefore necessary to divide in smaller areas. In Figure 2.3.2.2.1 below we show how the division N3 yields smaller errors than the division in Countries when reconstructing Country-to-Country flows, and both are significantly better than the Euro upscaling. We remark that these differences are not noticeable if one considers the correlation coefficient: a reconstructed flow very far from matching the actual magnitude of the real flow can still be highly correlated. As a conclusion, the answer to the question 'how to divide the area' is in this case the smaller, the better.

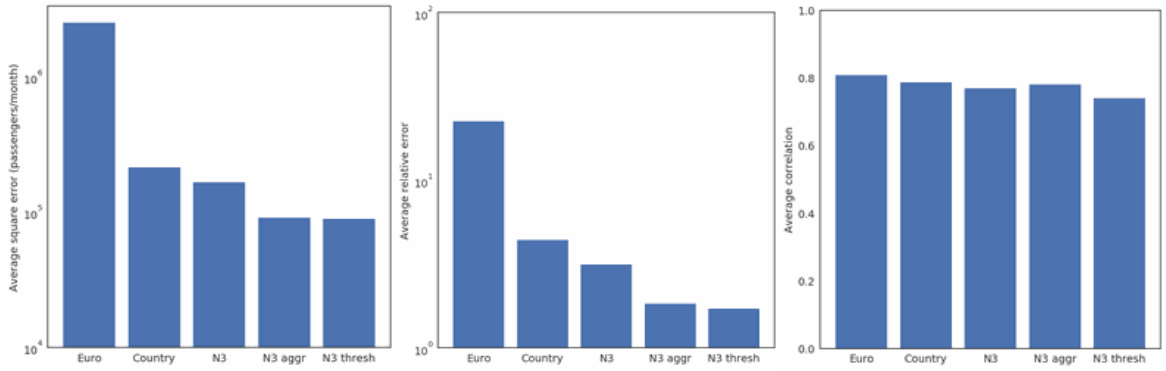

Figure 2.3.2.2.1 Quality of the Country-to-Country reconstructed flows with different upscaling. (Left) Average square error. (Middle) Average relative error (|reconstructed - ground_truth|/ground_truth). (Right) Pearson correlation coefficient.

Building on the NUTS3 level upscaling, we explored three different options on how to estimate the user-base U(A).

In principle, the user-base may change every month, because new users join the micro-blogging platform and old users quit. Moreover, the Twitter API might also provide different number of tweets users in different periods. Therefore, performing different upscaling for the reconstruction of the flows in different months is in principle correct. However, we find that estimating the user-base over the whole temporal extension of the dataset (N3 aggr) allows reconstructions to be more in line with

the ground truth. This is probably because the error introduced by the limited statistics in areas of low penetration is more relevant than the error introduced by the variation of the user-base across the 26 months analysed. This would not probably be true if we were using as a source of data a new social network with an exponentially growing user-base.

An alternative way for compensating the error introduced in the areas where, as a consequence of low penetration, the peculiar travel behaviour of few individuals can skew significantly our observations is to exclude these over-representing individuals from the analysis. We can identify an optimal threshold for the maximal upscaling factor accepted by minimising either the square or relative error (see Figure 2.3.2.2.2). For the country-to-country flows the maximum upscaling is of about 1000 (~800 optimising relative error, ~2000 optimising absolute error, and thus focusing on large flows). We remark however that, even if the errors in the estimates done with this 'N3 thresh' method can be compared with those of the 'N3 aggr' reconstruction, as a consequence of the exclusion of a fraction of the users, this methodology tends to systematically under-estimate flows. This effect will be further discussed in section 2.4.2.

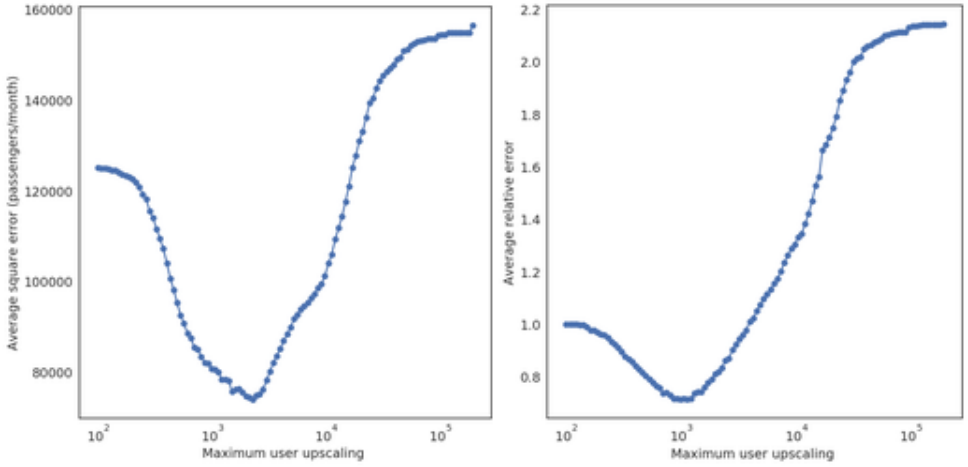

Figure 2.3.2.2.2 Finding the optimal upscaling threshold above which to exclude NUTS3 areas from the N3 tresh reconstruction. (Left) Optimising the average square error. (Right) Optimising the relative error.

2.4 Mobility information validation

2.4.1 Mobile phone data

2.4.1.1 Spanish residents

In this section, we present some exercises that validate mobile phone data as a source of mobility information.

In Figure 2.4.1.1.1, three estimations of the traffic flows obtained from mobile phone data for a highway (red, green and purple lines) in Spain are compared against traffic counts for that highway (shadowed area). The three estimations correspond to different values of car occupancy (as mobile phone data represents people, while traffic flows represent cars). It can be seen that the curve for mobile phone data represents almost perfectly the curves of the traffic counts.

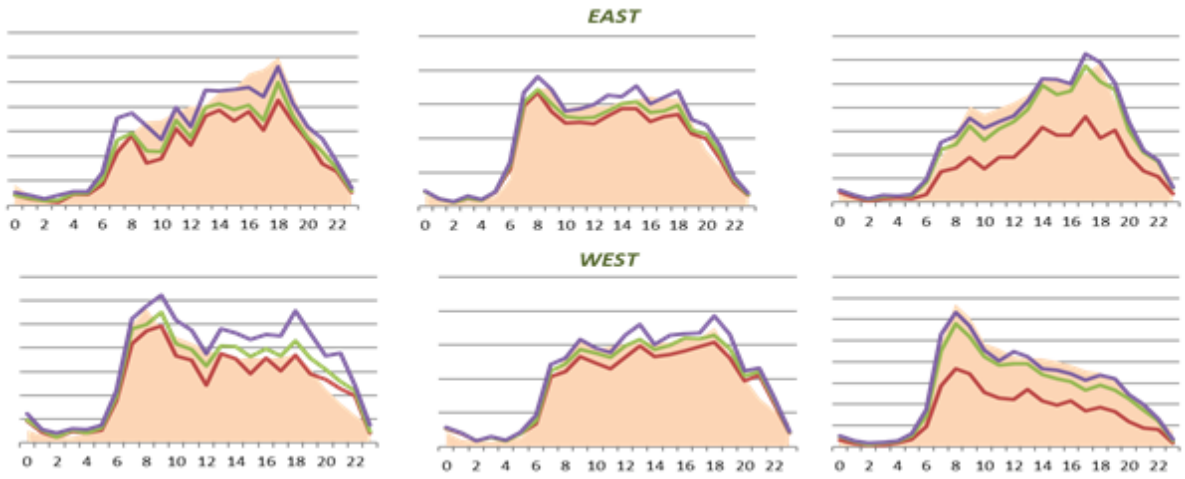

Figure 2.4.1.1.1 Comparison between traffic estimations from mobile phone data and traffic counts

Figure 2.4.1.1.2 shows a comparison of the visitors to a National Park obtained from mobile phone data with the visitors obtained from pedestrian counts for different days. There it can be seen that both methodologies present almost identical values.

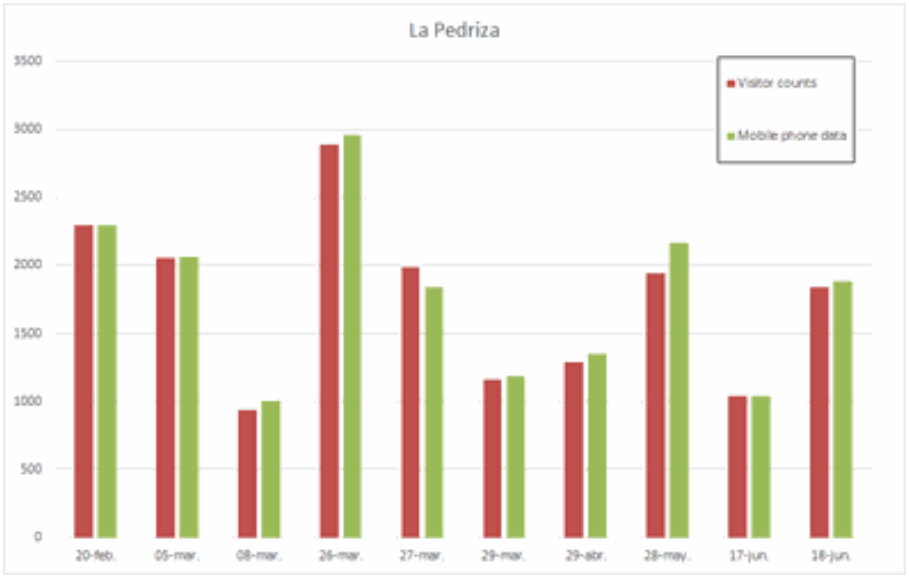

Figure 2.4.1.1.2 Comparison of visitors to a National Park obtained through mobile phone data and traditional methodologies

2.4.1.2 Roamers

Unfortunately, there is not much information about tourist mobility which could be used for validation exercises. In that sense, the effort of investigating new ways to assess tourist mobility is well justified. The available information for visitors is that mentioned in Section 2.3.1.2. For the validation exercise, the correlation between official information and that estimated from mobile phone is calculated for those quantities that have not been used for the expansion. Since the type of visitor per nationality has been used for the expansion, the correlation of this quantity is 1 by construction. But the share or number of excursionists and tourists per Autonomous Community (CCAA), the share/number of visitors per number of nights and the share by entrance port could be used for validation. A validation of the distribution of visitors in the territory is an indirect validation of tourist mobility within the territory at an aggregated level.

Table 2 and Figure 2.4.1.2.1 show the correlation for the number of excursionists and tourists per CCAA. In Table 2, it can be observed that Baleares and Cataluña are the two most preferred destinations for tourists. One can also observe a good correlation between the share obtained with mobile phone data and official statistics, obtaining a Pearson correlation factor of 0.99 for tourists and 0.93 for excursionists. The worse value is obtained for excursions in the Basque country (País Vasco), indicating that we are not capturing all excursionists visiting this CCAA.

Main destination	Tourists Expanded Sample	Share of Tourist Expanded Sample	Share of Tourists INE	Excursions Expanded Sample	Share of Excursions Expanded Sample	Share of Excursions INE
Andalucía	30324.03	13.07%	14.18%	5759.03	9.57%	7.48%
Aragón	1402.90	0.60%	0.56%	1149.35	2.33%	1.33%
Baleares	50840.97	21.91%	23.94%	2947.90	3.58%	4.35%
Navarra	656.45	0.28%	0.50%	2873.55	6.33%	7.24%
Canarias	28112.90	12.12%	12.82%	1804.51	2.34%	0.59%
Cantabria	1447.42	0.62%	0.76%	148.87	0.28%	0.00%
Castilla La Mancha	1002.74	0.43%	0.17%	236.61	0.55%	0.01%
Castilla y León	4036.78	1.74%	0.93%	2379.35	6.25%	3.70%
Cataluña	59093.22	25.47%	23.15%	23638.55	38.33%	37.39%
Ceuta	40	0.02%	0.00%	352.10	0.34%	0.12%
Comunidad de Madrid	7193.71	3.10%	4.79%	1551.29	4.78%	0.42%
Comunidad Valenciana	33482.26	14.43%	13.18%	1847.74	2.47%	1.34%

Main destination	Tourists Expanded Sample	Share of Tourist Expanded Sample	Share of Tourists INE	Excursions Expanded Sample	Share of Excursions Expanded Sample	Share of Excursions INE
Extremadura	457.90	0.20%	0.16%	570.64	2.34%	3.86%
Galicia	3357.42	1.45%	1.35%	2608.87	9.15%	9.96%
La Rioja	211.77	0.09%	0.11%	331.61	0.84%	0.02%
Melilla	79.19	0.03%	0.00%	143.06	0.15%	0.00%
País Vasco	3366.94	1.45%	1.41%	4819.19	9.95%	22.17%
Principado de Asturias	1324.19	0.57%	0.62%	101.77	0.18%	0.01%
Región deMurcia	3497.10	1.51%	1.37%	175.16	0.24%	0.00%

Table 2. Tourists and Excursionists share per CCAA.

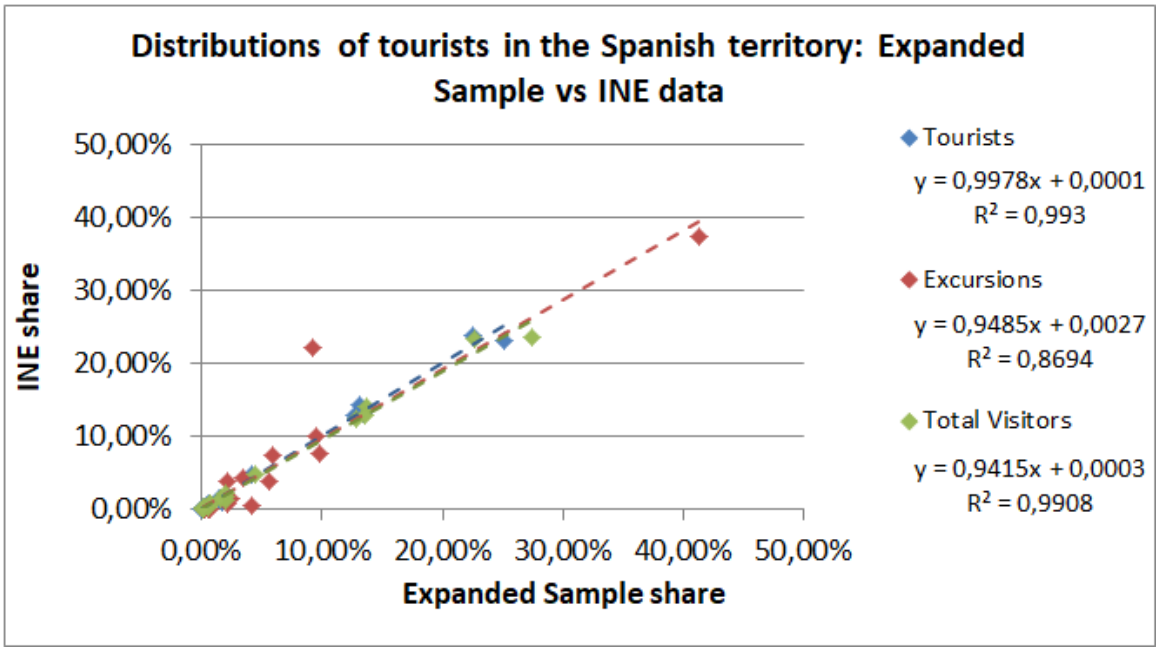

Figure 2.4.1.2.1 Correlation between visitors (tourist and excursionists) distribution in the Spanish official statistics and that obtained from mobile phone records.

Table 3 below shows the comparison of the official and expanded estimated share of visitors by length of stay.

No. of Nights	Expanded Sample	Share of Visitors Expanded Sample	Share of Visitors INE
0	148.277	5,27%	5,18%
1	166.771	5,93%	0,88%
2-3	312.969	11,13%	4,25%
4-7	963.708	34,28%	32,11%
8-15	973.345	34,62%	41,07%
> 15	246.311	8,76%	16,51%

Table 3. Share of visitors per length of stay

The Pearson correlation for the length of stay is of 0.94. This good correlation is due in part to the perfect fit of zero nights, which is a consequence of expanding by type of visitor (i.e., the number of excursions is exactly matched by the expansion).

Table 4 illustrates the share of visitors by entrance point. As it can be seen, the detection of entrance points still needs to be refined. One reason for the large percentage of unknown entrance point is that some of the users are captured for the first time once they are well inside the territory. For the elevated number of maritime port entrance, the reason may be that in some cities like Barcelona and Malaga the port is quite close to the city centre. To overcome the over estimation of visitors entering by port, we intend to explore the use of different distance thresholds according to the cities characteristics and the type of entrance via. As for the number of unknowns, we expect that the number of EU users found for the first time in Spain in the city centre will be reduced for data collected after the change in the roaming charges within the EU. However, keeping only those users for which the entrance via is known still provides a good sample size, representing a 14% of the total visitors registered by official statistics in the period of study. Also, the share of visitors in the different CCAA is still well correlated with those of the official statistics (see Table 5 and Figure 2.4.1.2.2). A research line to be tackled in the future is the characterisation, by using different machine learning techniques, of users entering by the different points and transport modes. With this further knowledge, an entrance port could be probabilistically assigned to those users for which it is unknown according to their characteristics (e.g., nationality, length of stay, mobility behaviour within the country, etc.).

Entrance via	Expanded Sample	Share of Visitors Expanded Sample	Share of Visitors Expanded Sample (considering only known entrance)	Share of Visitors INE
unknown	116616.13	40.71%		
road/train	36450.16	12.72%	21,46%	23.88%
port	27931.29	9.75%	16,44%	1.21%
airport	105484.68	36.82%	62,10%	74.91%

Table 4. Distribution of visitors by entrance port

Main destination	Tourists Expanded Sample (with known entrance port)	Share of Tourists Expanded Sample (with known entrance port)	Share of Tourists INE	Excursions Expanded Sample (with known entrance port)	Share of Excursions Expanded Sample (with known entrance port)	Share of Excursions INE
Andalucía	19098,55	13,31%	14.18%	2474.84	9.78%	7.48%
Aragón	738.87	0.51%	0.56%	623.39	2.46%	1.33%
Principado de Asturias	772.26	0.54%	0.62%	28.71	0.11%	0.01%
Baleares	35628.87	24.83%	23.94%	1380.65	5.46%	4.35%
Canarias	19857.74	13.84%	12.82%	931.61	3.68%	0.59%
Cantabria	886.13	0.62%	0.76%	63.39	0.25%	0.00%
Castilla y León	2062.74	1.44%	0.93%	1108.06	4.38%	3.70%
Castilla La Mancha	494.19	0.34%	0.17%	35.48	0.14%	0.01%
Cataluña	34366.29	23.95%	23.15%	8617.26	34.05%	37.39%
Comunidad Valenciana	18554.03	12.93%	13.18%	568.23	2.25%	1.34%
Extremadura	245.97	0.17%	0.16%	355.16	1.40%	3.86%
Galicia	1950.65	1.36%	1.35%	1457.58	5.76%	9.96%
Comunidad de Madrid	3840.32	2.68%	4.79%	906.13	3.58%	0.42%
Región de Murcia	1995.97	1.39%	1.37%	38.87	0.15%	0.00%
Navarra	443.87	0.31%	0.50%	2352.42	9.30%	7.24%
País Vasco	2367.90	1.65%	1.41%	3669.19	14.50%	22.17%
La Rioja	115.32	0.08%	0.11%	225.81	0.89%	0.02%
Ceuta	29.35	0.02%	0.00%	331.13	1.31%	0.12%
Melilla	63.71	0.04%	0.00%	140.00	0.55%	0.00%

Table 5. Distribution of tourists and excursionists with known entrance port by main destination.

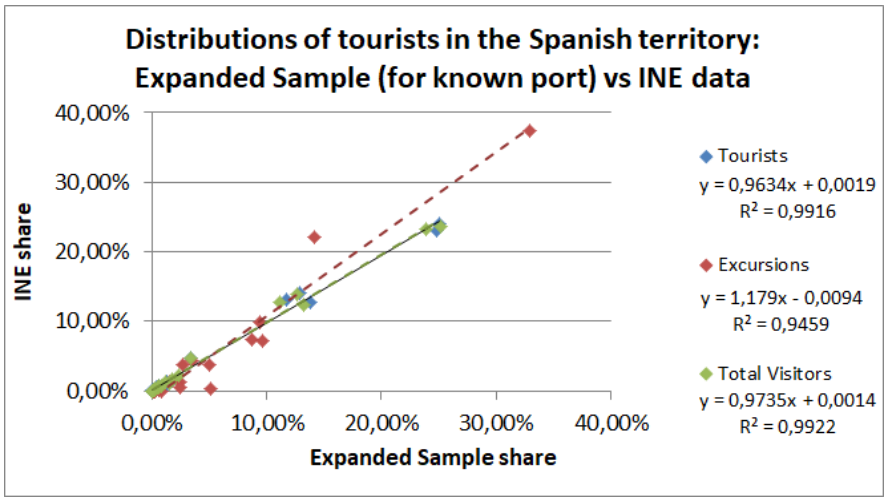

Figure 2.4.1.2.2 Correlation of official and observed visitors (tourists and excursionists) distribution in the Spanish territory, considering only visitors for which the entrance port is detected.

2.4.2 Twitter data

One of the project goals is to verify if, by using the Twitter data, it is possible to reconstruct aggregated statistics about the flows of passengers between countries and between specific airports. Here we intend to validate the use of georeferenced tweets for the description of passengers' trips, since in the following phases we will use the tweets to infer information that is not available in the publicly available statistics, in particular the precise origin and destination area. The deviations observed in will help us highlight the limits inherent to our data, which shall be taken into account and corrected in order to provide an unbiased view on the passengers' travel behaviour.

2.4.2.1 Ground truth data

We validate our estimates by comparing the estimated flows for a period of five months ranging from November 2014 to March 2015 with two independent datasets. The first are the tickets sales provided by Sabre Airline Solutions, where flows between origin and destination of the whole passenger trip are recorded on a monthly based. The second are the statistics provided by Eurostat, again on a monthly basis. This second database differs from the first because here the flows between two airports count the passenger travels between them, and thus passenger who did more than a flight in a trip are counted once for each leg. In Figure 2.4.2.1.1 we can observe that the two datasets are highly correlated, and the correspondence is almost perfect if we extract the passengers per route from the Sabre data.

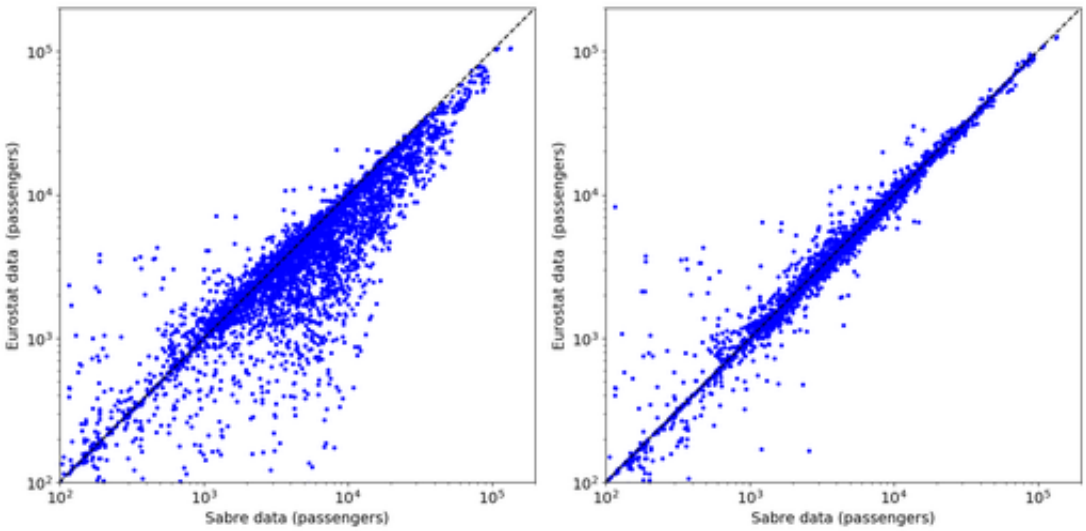

Figure 2.4.2.1.1 Comparison between the flows of passengers at the country level in Sabre and Eurostat data for the month of November 2014: (Left) OD flows from Sabre with Eurostat's passenger per flight; (Right) Passenger per flights from Sabre data vs the same measure for Eurostat data (Pearson-r = 0.996).

The choice of the interval November 2014-March 2015 was made because it is the period of overlap between the Sabre data we own and our Twitter database. It is also worth noting that it represents the best period of our Twitter coordinate data, before the change of interface. The reconstruction of flows from Twitter is however supported by the use of the place field in a tweet's metadata. In the following, we show that the quality of our reconstruction is only slightly influenced by the decreased statistics associated to the change of Twitter's interface.

2.4.2.2 Flows between countries

We start by analysing the result of the reconstruction of flows between countries, as working on larger statistical ensembles allows us to focus more on the quality of the different upscaling methods considered, while the reconstruction of the flows between airports will allow us to focus more on the issues inherent to the Twitter dataset.

Here, we estimate international flows from Twitter simply by tracking the country code of subsequent tweets of the same user, which is information consistently present in each tweet. Any time two subsequent tweets have been made in different countries, it is assumed that a trip has been made by the given user between those countries. The flows are then upscaled using the methods N3, N3-aggr and N3-thresh described in the section 2.3.2. As we can see in Figure 2.4.2.2.1, in all three cases the flows we obtain here are highly correlated (with a correlation coefficient of approximatively 0.77) with ground truth datasets. The relative deviation from the identity has here its minimum for the method N3-aggr, based on the computation of the user-base across the whole time-span covered by the data instead of on a monthly basis as in the method N3.

Moreover, for some couple of countries the Twitter flow largely over-estimates the flight statistics. This is because Twitter data actually represents mobility via any mode of transport. This is exemplified by the trajectory between Belgium and the Netherlands (orange diamond). This Origin-Destination (OD) pair between two neighbouring and relatively small countries, between which it is unlikely to flight, represents here the largest deviation from the identity in the flow reconstruction. Heterogeneity in the modal split is indeed one of the limits we have to face in our analysis.

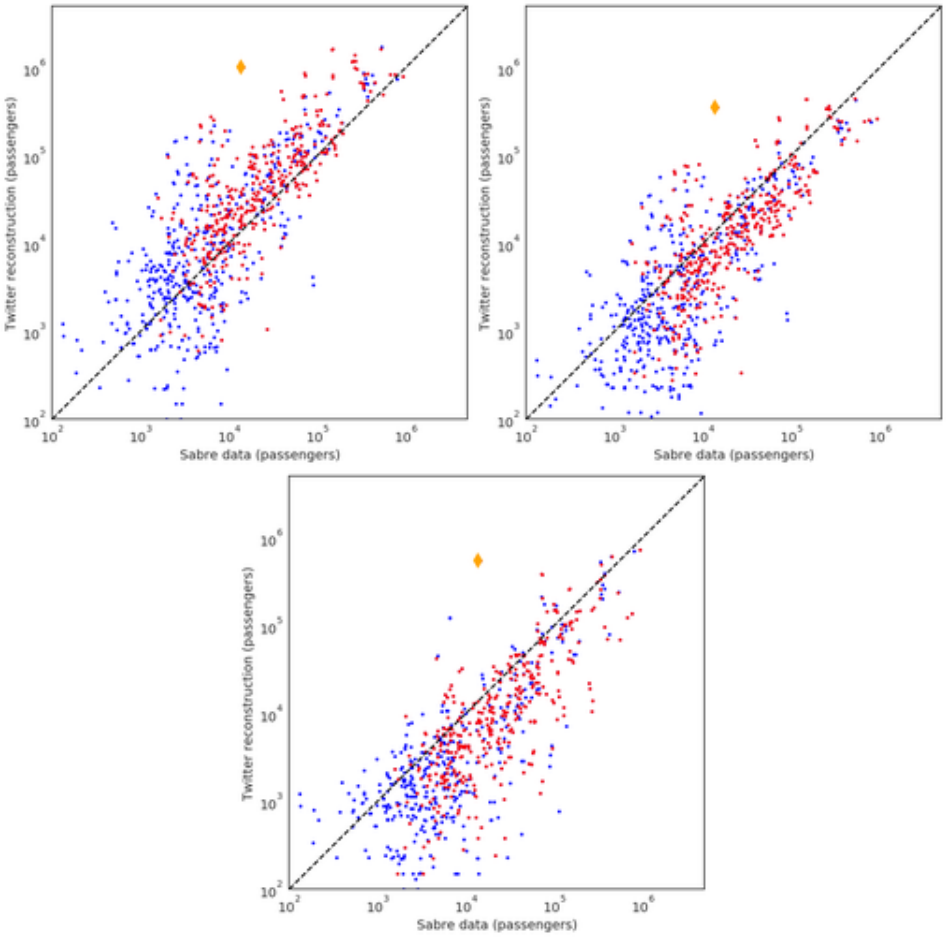

Figure 2.4.2.2.1 Reconstruction of the Country-to-Country international flows in the ECAC area for November 2014. In red, the ensemble of flows included in both the Eurostat and Sabre dataset that have been considered for the comparison between upscaling models. (Up-Left) the reconstruction N3, that systematically over-estimates flow. (Up-Right) the reconstruction N3-aggr, that appears to be the most balanced. (Down) the reconstruction N3-thresh, that systematically under-estimates flows.

Our use of the relative error (evaluated between two quantities x and y as err = |x-y|/x) instead of the average quadratic error is a choice made in order not to bias our choice in the upscaling algorithm towards the optimal reconstruction of largest flows. Nevertheless, as we see in Figure 2.4.2.2.2, regardless of the method chosen, the reconstruction with the N3-aggr method is better when the flow is larger. This is true also true for the airport-to-airport reconstruction and is a natural consequence of the larger number of tweets and users available when tracking movements between large cities or countries.

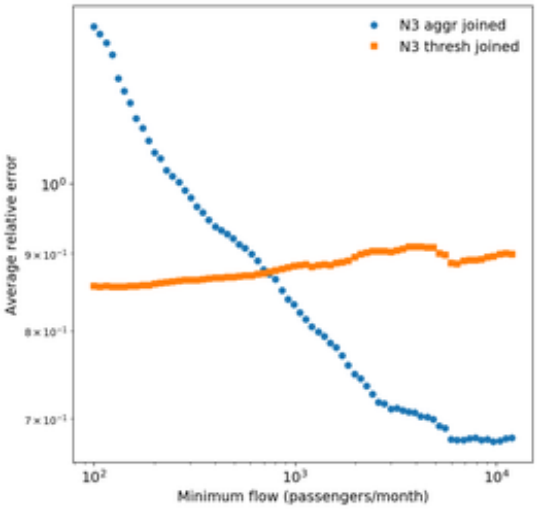

Figure 2.4.2.2.2 Average relative error in the reconstruction of the five months considered, restricted to couple of countries with a flow higher than the value in the x-axis.

As discussed above, we have the opportunity of comparing our reconstruction with two independent datasets. We found that the international flows estimated with Twitter match better inter-airport flight passenger counts rather than the passenger Origin-Destination data describing their whole trip. This is quite surprising, because the inter-event time between tweets is typically quite large. This can be perhaps influenced by the large presence of automatic social network check-ins in the data (see Section 4), and it is also possible that many users use Twitter during their layovers. However, the effect is so remarkable that we cannot exclude the possibility of this being a statistical effect due to the heterogeneous distribution of Twitter users.

In figure 2.4.2.2.3, we use Eurostat data to verify the quality of the reconstruction beyond the five months where Sabre data is available, Eurostat data being here equivalent to Sabre intra-airport flows. The reconstruction maintains relatively good quality even after the change in the Twitter interface of May 2015.

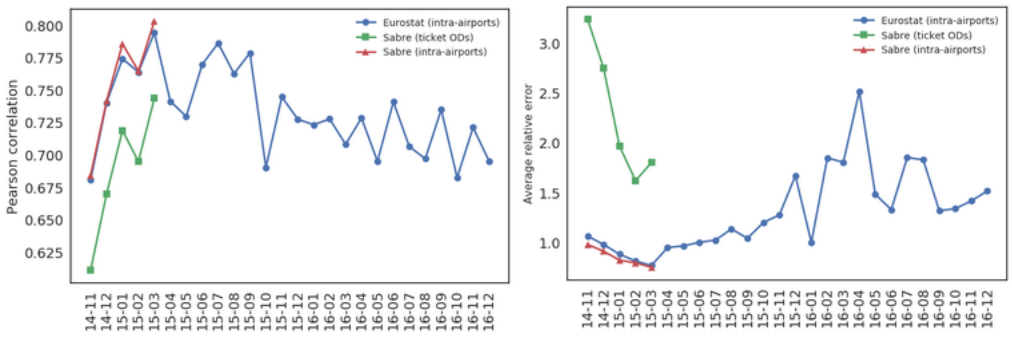

Figure 2.4.2.2.3 Quality of the reconstruction month-by-month. (Left) Correlation coefficient. (Right) Average relative error.

2.4.2.3 Flows between airports

The next challenge is to test the potentiality of our data in reconstructing flows at a smaller spatial scale, the one of the single airport. The starting point to do this is to approximate the airports' catchment areas with a Voronoi tessellation, as represented in Figure 2.4.2.3.1. This assumption follows the rationale that a traveller will use the airport closer to his/her origin and destination.

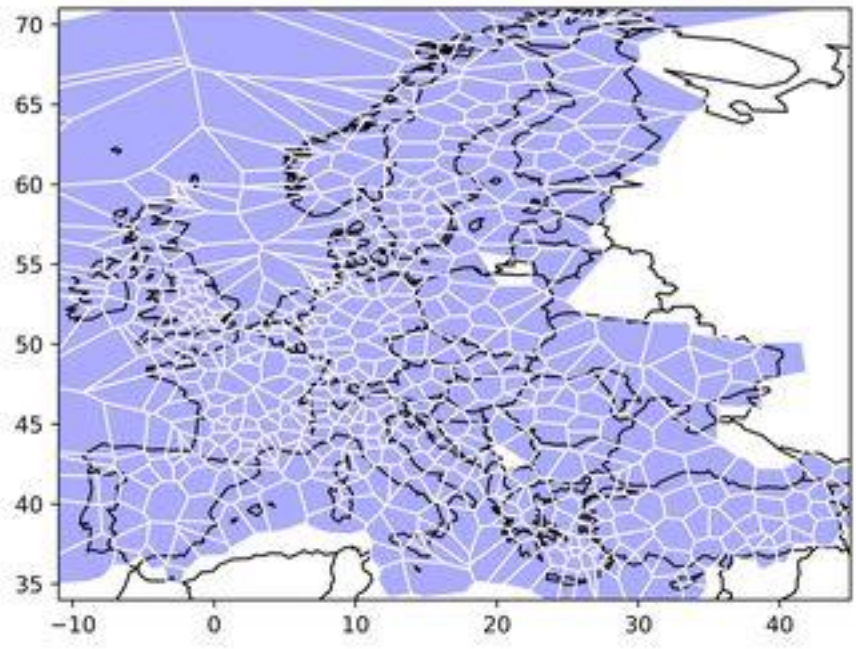

Figure 2.4.2.3.1 Voronoi tessellation approximating the airport's catchment areas.

Naturally, not all pairs of airports are connected by regular flights. So when a couple of consecutive tweets lie in the cells around airports with less than a flight per week, an alternative origin is searched among the nearest geographical neighbouring airports progressively until a viable option is found (see Fig. 2.4.2.3.2). The analysis is limited to pairs of tweets with at least d > 500 km between them, that is, a distance we would expect to be typically covered by flight.

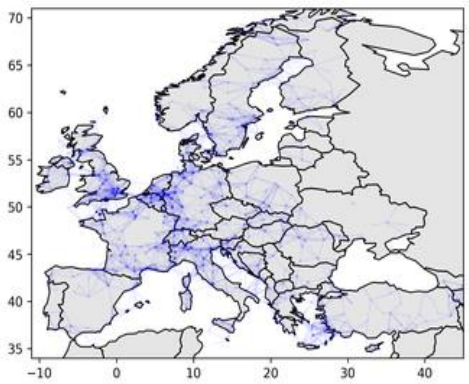

Figure 2.4.2.3.2 Re-assignment of a tweet from a Voronoi cell to the closest origin of a flight to the desired destination.

The results in Figure 2.4.2.3.3 show in general a broader error in the estimate of airport to airport flows as compared to the country to country flows. This is a natural effect of the reduced statistics of Twitter that here is associated to each couple of airports instead of being aggregated at a larger scale.

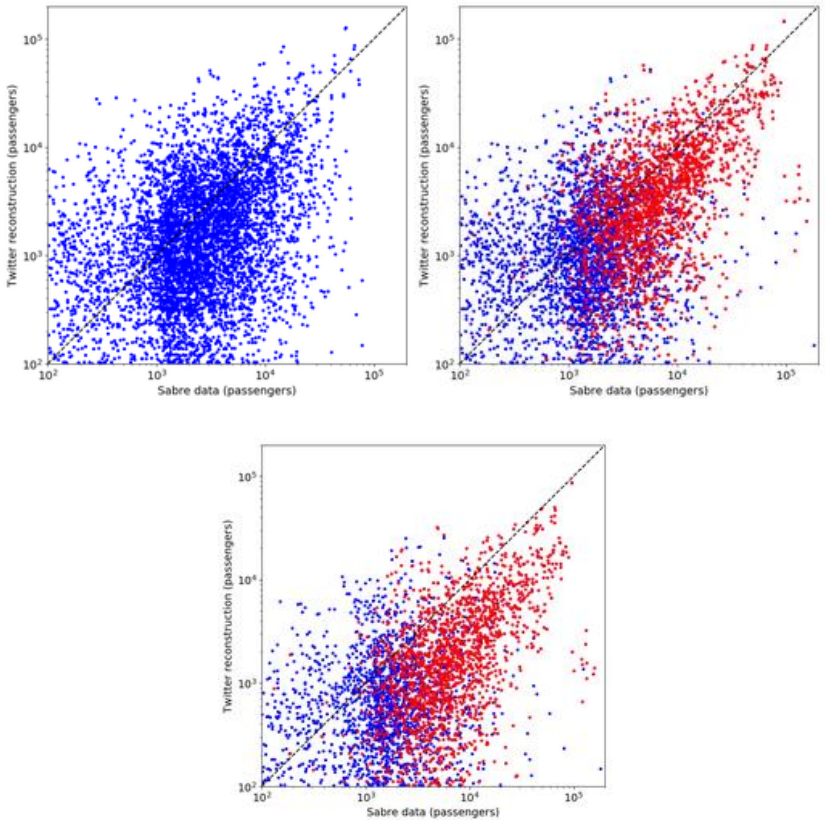

Figure 2.4.2.3.3 Airport-to-airport flows reconstructed for November 2014. All reconstructions are based on the N3-aggr method. (Up-Left) Including all data and with every point representing. (Up-Right) Joining together airports serving the same city. (Down) Filtering out movements reconstructed between tweets written more than a week apart. The red point represents the subset of routes used for measuring the comparison in Figure 2.4.2.3.6

A further source of error is introduced by our approximation to airport catchment areas. This is particularly relevant in cities where more than one airport is available. This particular issue can be avoided by joining together flows toward the same city in what we call here 'joined flows'. The effect of this reduces the relative error and increases correlation with the N3-aggr upscaling (see Figure 2.4.2.3.4), in particular for large flows (see Up-Right panel of Figure 2.4.2.3.3).

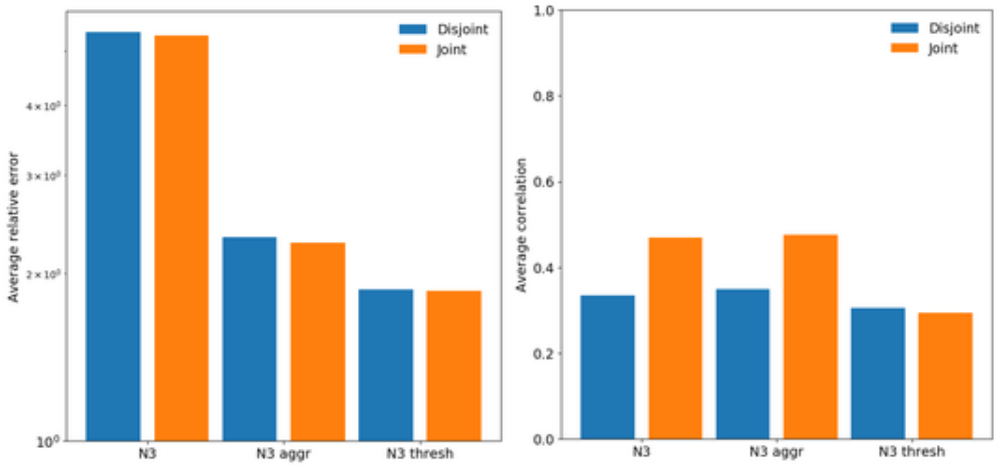

Figure 2.4.2.3.4 Average quality of the reconstruction with different upscaling methods in the five months considered.

Similar to the country-to-country problem, also here the N3-aggr method works significantly better for joined flows than the N3-thresh method based on excluding low representation areas (see Figure 2.4.2.3.5).

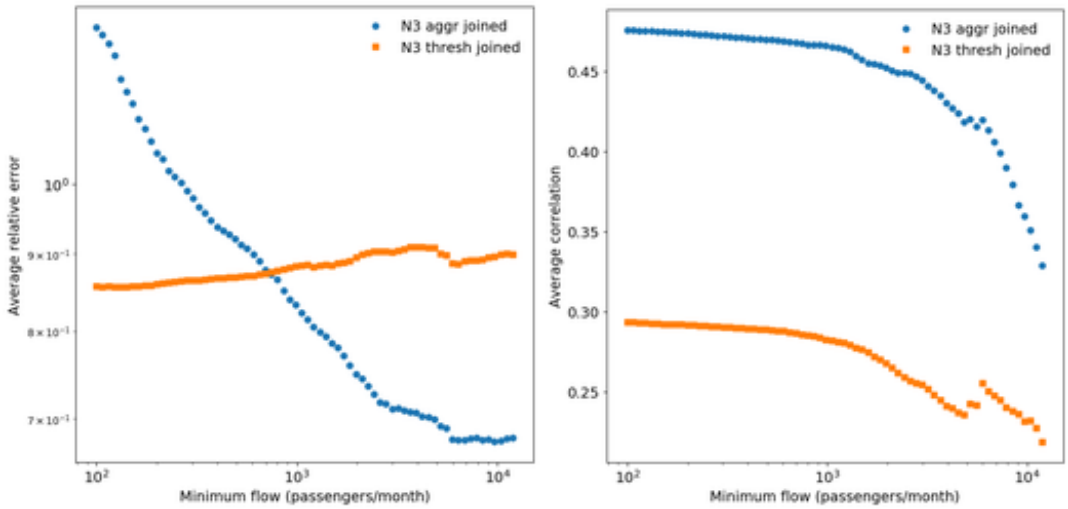

Figure 2.4.2.3.5 Average relative error (left) and correlation (right) in the reconstruction of joined flows of the five months considered, restricted to country pairs with a flow higher than the value in the x-axis. Once again, the correlation is strictly better for N3-aggr, which also progressively yields smaller error for more important flows.

Another condition we have evaluated for reconstructing the flows is selecting only couples of tweets written with less than a threshold dt of time between them. In principle, we could expect that the shorter dt, the more precise the description of the user's movement. However, at the same time reducing the already sparse sampling Twitter data introduces errors. We find (see Figure 2.4.2.3.6) that restricting the analysis to pairs of tweets with dt < 1 week reduces the error, but at the same time reduces the correlation of the reconstruction and introduces a systematic underestimation of the flows. This underestimation could possibly be corrected with a more refined upscaling based on the characterisation of the probability distribution of dt to take into account how many correct trips we expect to systematically cut from our reconstruction as a consequence of this temporal selection.

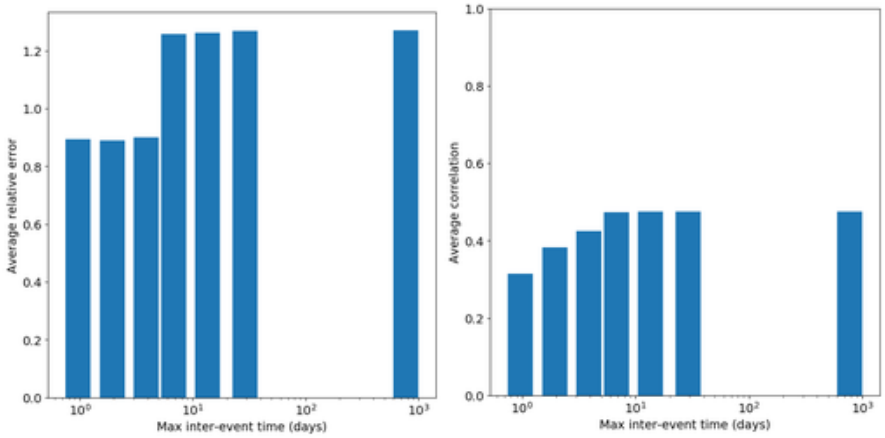

Figure 2.4.2.3.6 Quality of the reconstruction setting a max inter-event time dt. (Left) The average relative error drops about 20% if dt < 1 week. (Right) At the same time, the reduced sample size consequence of the threshold introduces sampling error thus reducing the correlation. The rightmost bar of both plots corresponds to no threshold (dt = time span of the dataset). All data describe joined flows and is averaged over the five months considered.

Finally, in the error observed in the scatterplots of Figure 2.4.2.3.3 we observe both overestimation and underestimation. Overestimation can be partially associated to the fact that our database includes movements that are not necessarily made by air, and actually represents an important feature as it might in principle represent future market opportunities, but is also a consequence of an imprecise matching of a geolocated tweet with the right airport. The same is true for the opposite case of underestimation, that is indeed manifestly reduced by joining flows of neighbouring airports in serving the same city. For instance, a large underestimated flow is the connection between the low-cost airport of Istanbul and Izmir, in Turkey. By checking the relative position of the three airports in Istanbul, it is easy to notice that the other two airports of Istanbul are capturing the almost totality of the urban area with the Voronoi tessellation. This illustrates a problem with cities with more than a single airport, which we can isolate by elaborating those cities separately, taking into account in particular ticket prices and the airport access times. This problem will be tackled in one of our case studies on the study of catchment areas, where we will try to develop a method able to estimate travel demand at an urban level from online social media.

2.5 Visualisation of mobility

As an example of visualisation, a tool has been developed for acquiring timetables of public transport from GTFS feeds (namely, from https://transitfeeds.com/), converting them to trajectories for a target date, and calculating reachability from a given area (a set of stops) to all stops of public transport. The tool supports investigation of reachability for a given time (say, departing from airport at 7:00 or arriving to airport at 19:00), comparing dynamics of reachability (e.g., if departing from airport at 7:00, 8:00, 9:00, etc.), detecting locations which cause many fastest routes to wait, and investigating what-if scenarios. Some examples in Madrid are offered below.

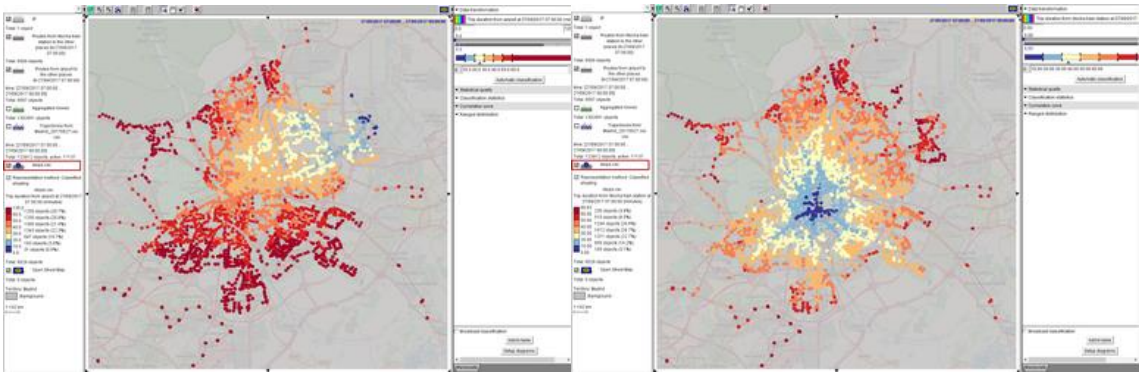

Figure 2.5.1 Travel from the airport to public transport stops in Madrid (left) and from Atocha train station (right).

Figure 2.5.1 shows travel times from the Madrid Barajas airport (MAD) to all stops of public transport in the city of Madrid at 7:00 on 27 September 2017. Similarly, on the right travel times from Atocha train station starting at the same time are shown. It is easy to observe the centrality of Atocha in the South area of Madrid, while the airport in the Northeast has a less central position and more complicated access times.

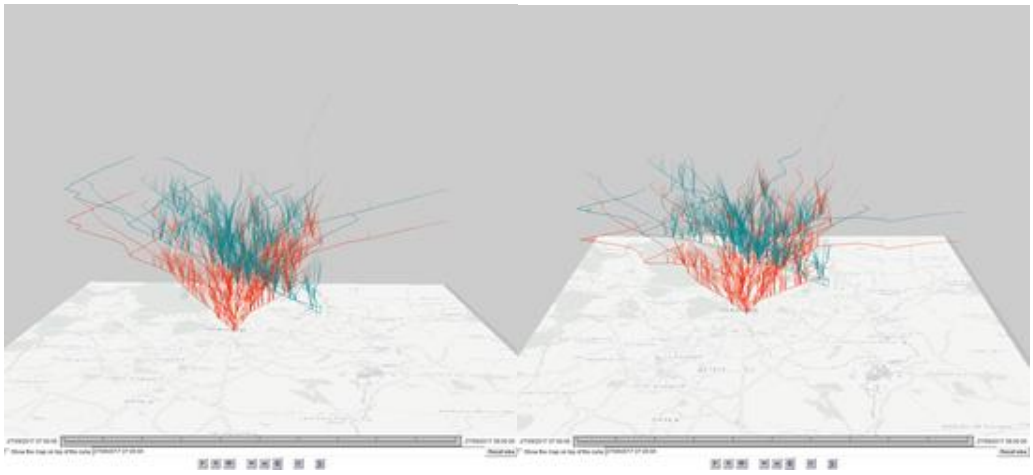

Figure 2.5.2 Spatio temporal trajectories from the airport (left) and Atocha train station (right).

Figure 2.5.2 shows space-time cubes with different viewpoints representing optimal routes from the train station (in red) and airport (in blue). In this representation, the vertical dimension represents time. We can observe that some areas are faster to reach from the airport, while many others are better connected to the Atocha train station. To study this in detail, we have calculated difference in travel times.

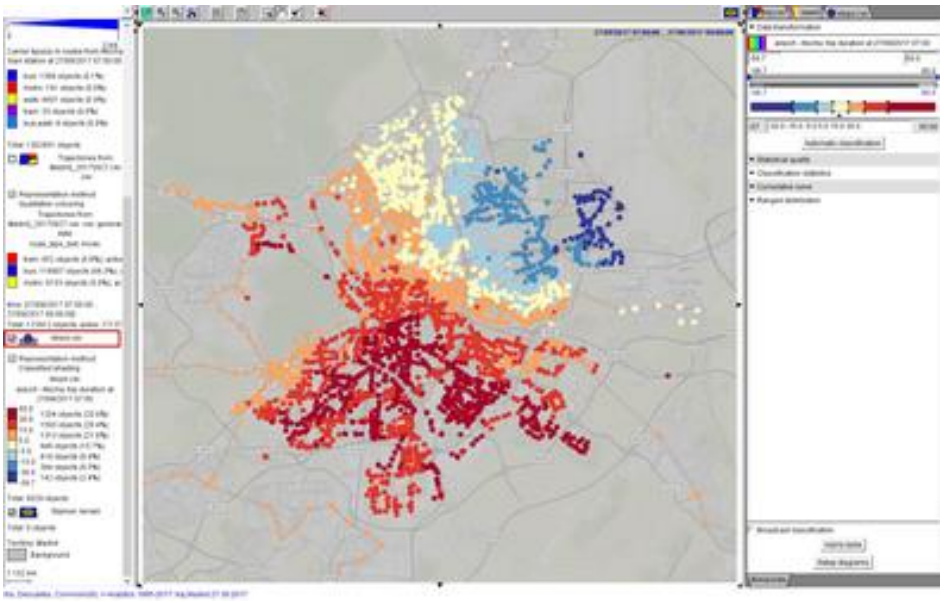

Figure 2.5.3 Time to travel from the airport to the metro stops.

In Figure 2.5.3, the stops with similar travel times for both the train station and the airport are marked in yellow; stops that require more time to travel from the airport are marked in shades of red; stops that are better connected to the airport than to the train station are marked in shades of blue. In the following analysis, we investigated travel times from the airport at 7am in two scenarios. In the first scenario, all public transport modes (bus, metro, tram) operate normally. In the second scenario, the metro line 8 is closed.

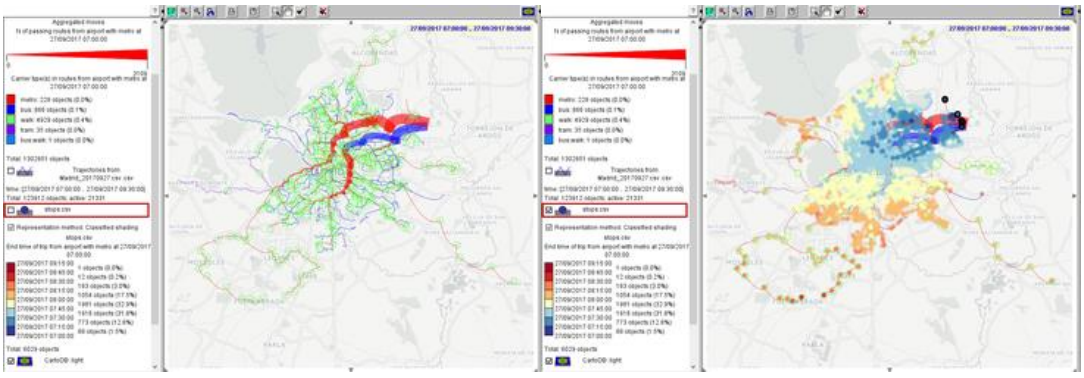

Figure 2.5.4 Fastest lines from the airport (left) and arrival times at the airport (right).

Figure 2.5.4 (left) shows flows of fastest routes from airport to all stops in the city using all possible transportation modes (metro, bus and tram). Colours correspond to transportation modes (red: metro, blue: bus, purple: tram), and line thickness to the number of fastest routes that use each segment. Green lines and circles correspond to moving by foot to a neighbouring stop. In Figure 2.5.4 (right), colours show arrival times for stops in the same scenario. Bluish colours correspond to "before 7:45", yellow to "before 8:00", and orange and red to "arriving after 8:00".

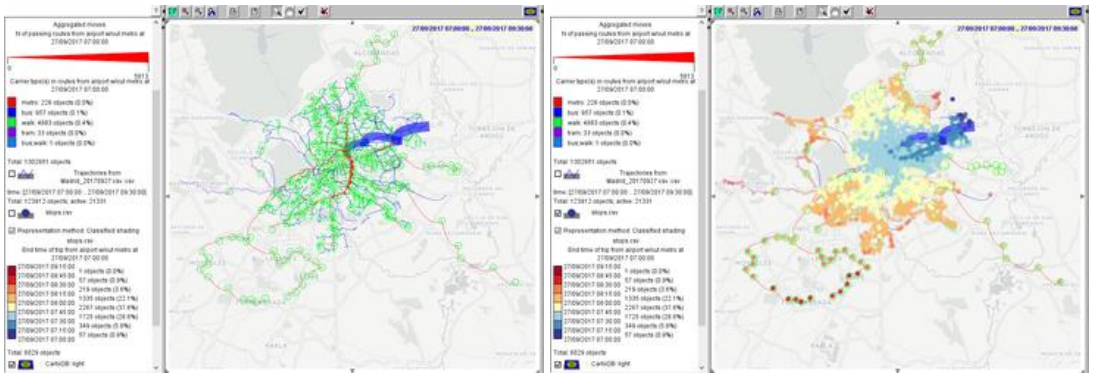

Figure 2.5.5 Fastest lines from Barajas airport (left) and arrival times at the airport (right) in a scenario in which line 8 of metro is closed

Similarly, Figure 2.5.5 shows optimal routes and corresponding arrival times if departing from the airport at 7am in the scenario of metro line 8 not operating.

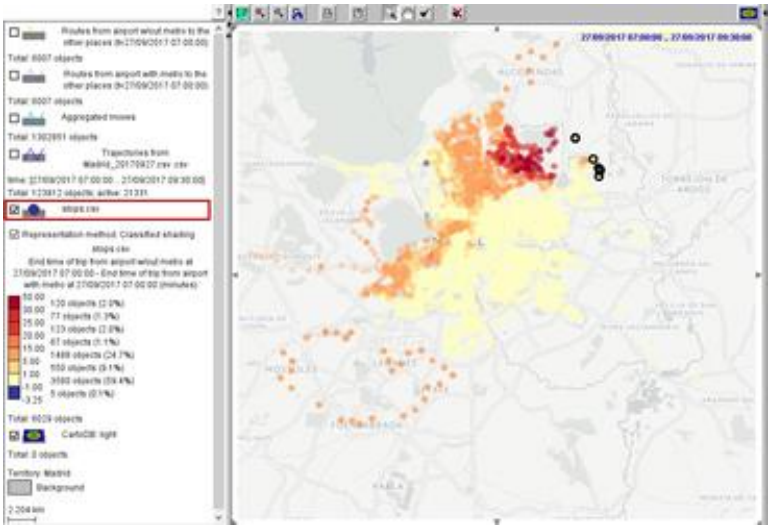

Figure 2.5.6 Delay produced in the accessibility to the airport by the closure of line 8 of metro.

Figure 2.5.6 shows the spatial distribution of the delays in the second scenario (no metro line 8) in comparison to the first scenario. Yellow means no difference (+/- 1 minute), reddish colours mean delays up to 50 minutes. A few stops in blue have arrival times earlier in the second scenario because our method tries to use public transport instead of going by foot, though sometimes the latter is faster.

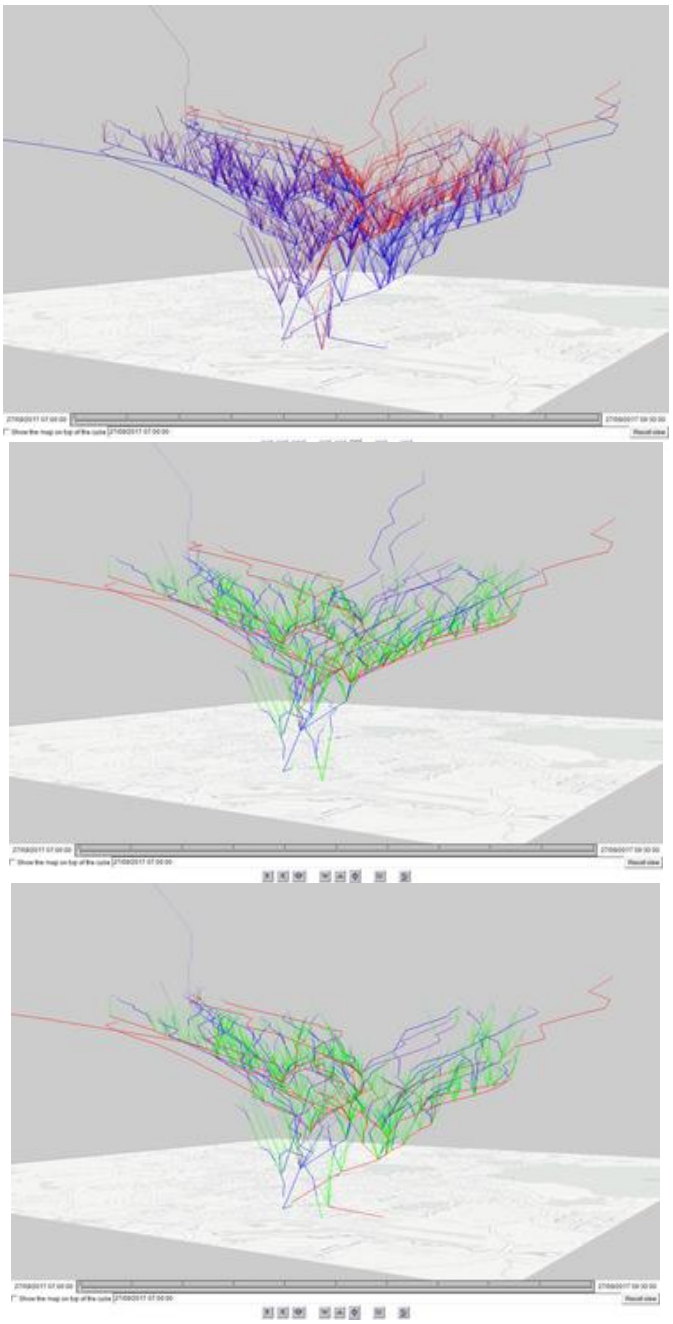

Figure 2.5.7 Changes in the trajectories after the closure of line 8 of metro: (Up) normal situation all modes, (Middle) normal only metro and (Down) line 8 of metro closed.

In Figure 2.5.7, three space-time cubes show the comparison of trees of fastest routes in two scenarios. To identify potential bottlenecks in the transportation network, we have calculated the average waiting time across all optimal routes at all intermediate stops.

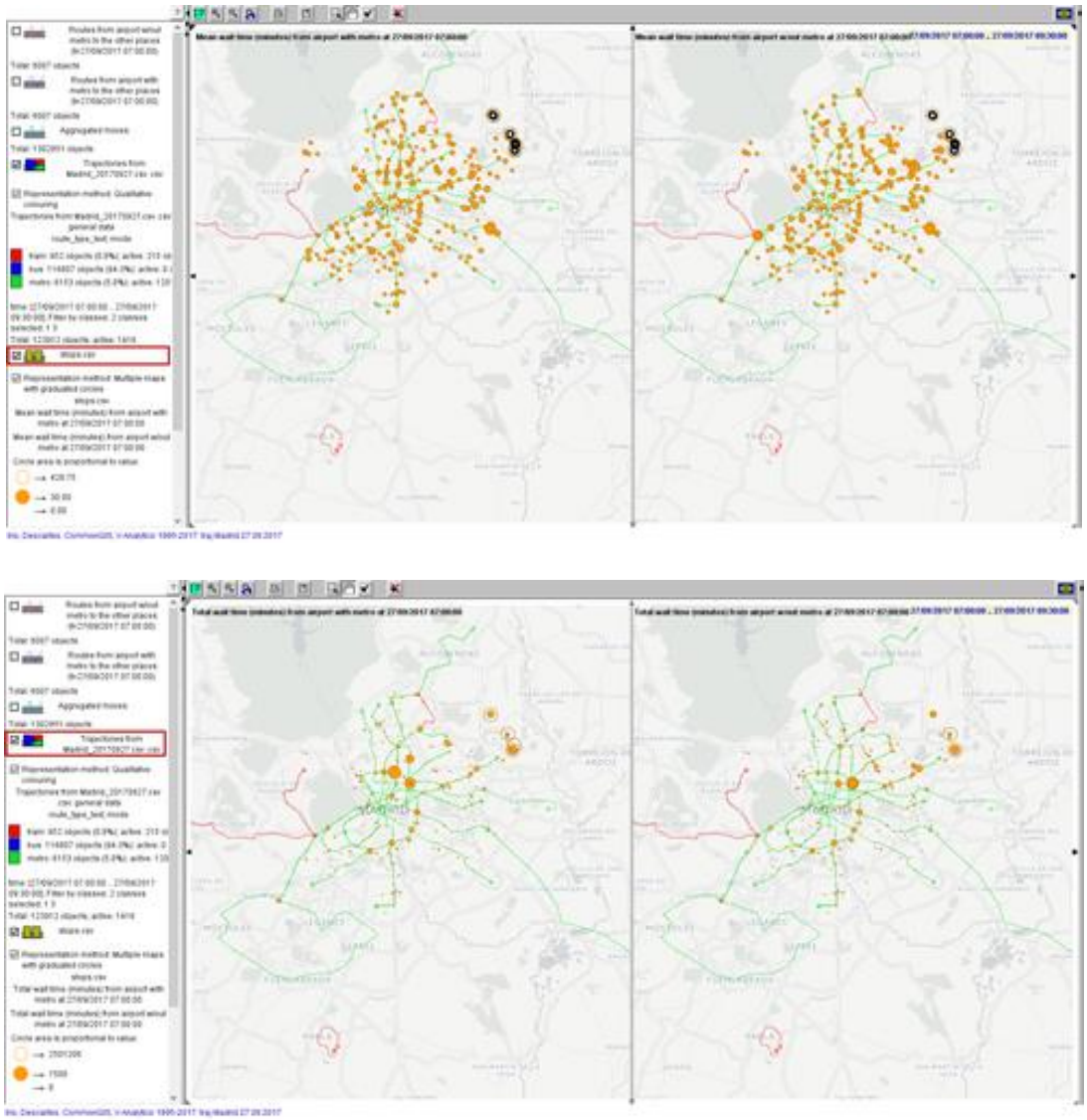

Figure 2.5.8 Waiting times: average (Up) and cumulated (Down).

Figure 2.5.8 (Up) shows the intermediate stops with the longest average waiting time, where the largest circles correspond to 30 minutes. Respectively, Figure 2.5.8 (Down) shows sums of waiting times for all fastest routes going through a stop. Largest circles correspond to about 7500 minutes of cumulative waiting. These stops are very important for the overall network, as any delay in them may affect a large number of optimal routes in the city. Such analysis can be extended for considering trips to selected areas (e.g., airport, bus terminals or train stations) and comparison of reachability at different times of day or days of week.

3 Analysis of passenger behaviour inside and around the airport

3.1 Mobility analysis from Twitter data

Twitter information about the user's coordinates, and eventually the place field if the POIs information is provided at resolution of terminals or higher, can be used to analyse activity and mobility within single airports. Due to statistical issues, this methodology works well in very busy airports like the main international hubs. In addition, it is necessary to look at airports with a structure composed of several terminals to be able to study passenger flows, since the GPS precision falls within a range of tens of meters and in small airports this may hinder the results. We include here an example of a preliminary study for London Heathrow (LHR). A heat map of the activity with all the tweets superimposed on the airport terminals can be seen in Figure 3.1.1.

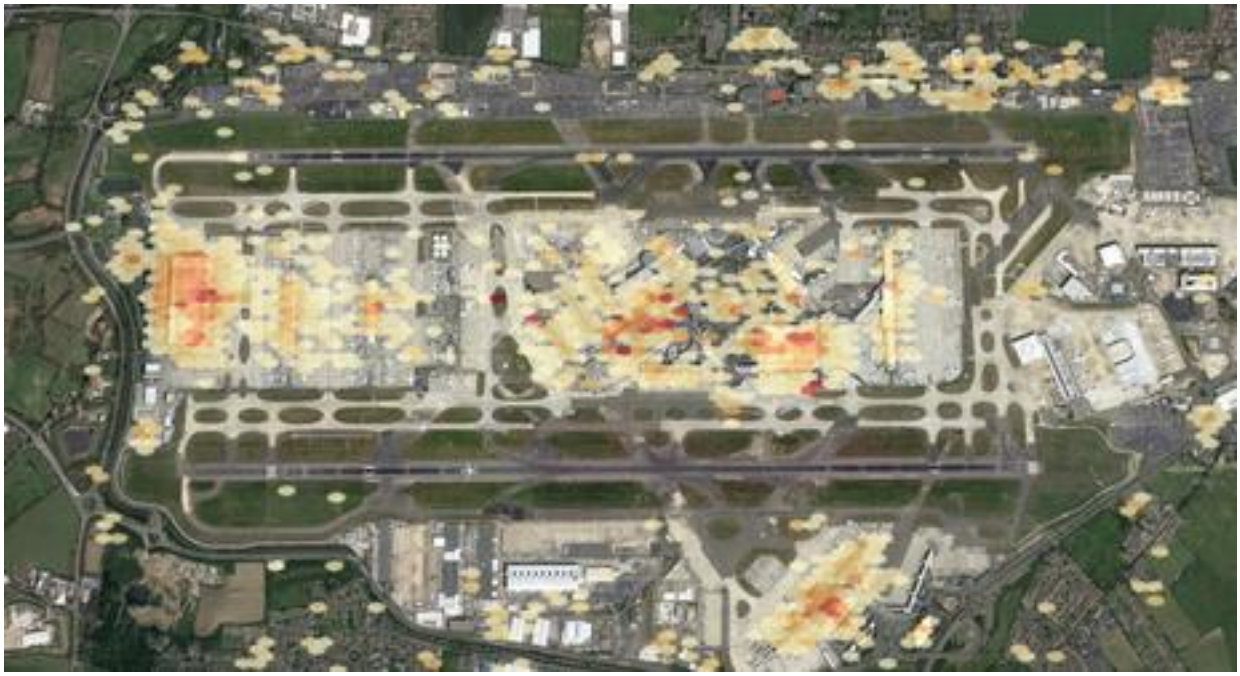

Figure 3.1.1 Heat-map of tweets coordinates recorded in London Heathrow (LHR).

The total number of daily tweets detected in the airport premises as a function of time shows a decrease after the change of policy of Twitter regarding coordinates, but it is now stabilised around 100 tweets/day (see figure below). This information can be aggregated to study hourly behaviours in the same day of the week or disaggregated at the user level (see Figure 3.1.2 below).

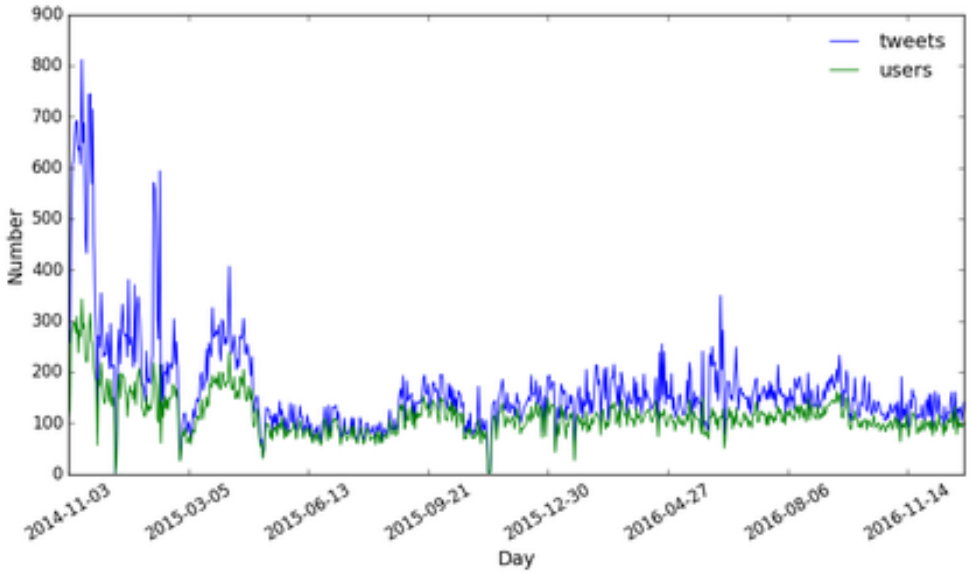

Figure 3.1.2 Number of tweets and active users per day

At the user level, we can obtain the distribution of the number of tweets displayed in Figure 3.1.3.

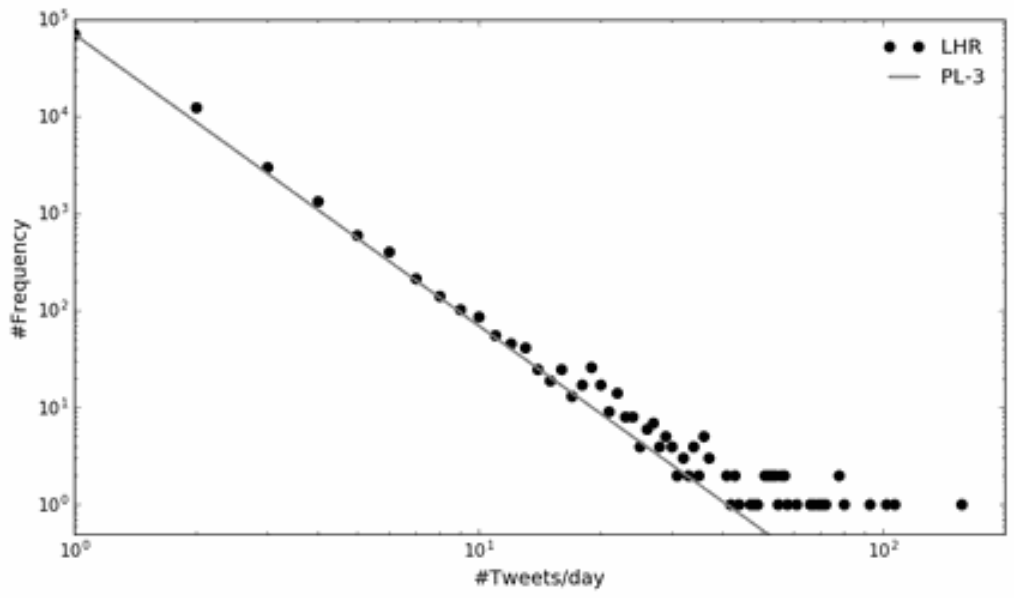

Figure 3.1.3 Distribution of the number of tweets made in any day by any user.

Day by day, the average number of tweets per user produces an irregular curve but in general it is fluctuating around 1.3 (see Figure 3.1.4).

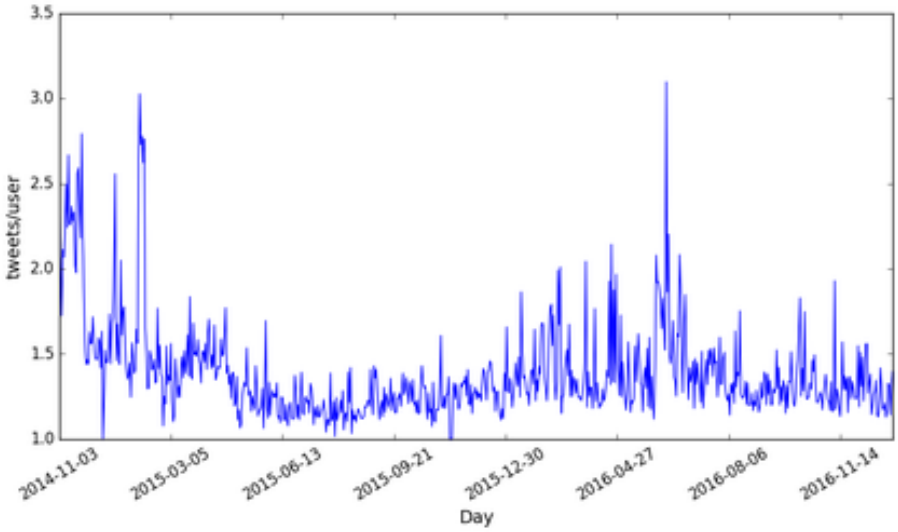

Figure 3.1.4 Average number of tweets per active users each day.

Those users with more than one tweet in the day provide us with information about movements between terminals and allow us to plot the flow map of Figure 3.1.5.

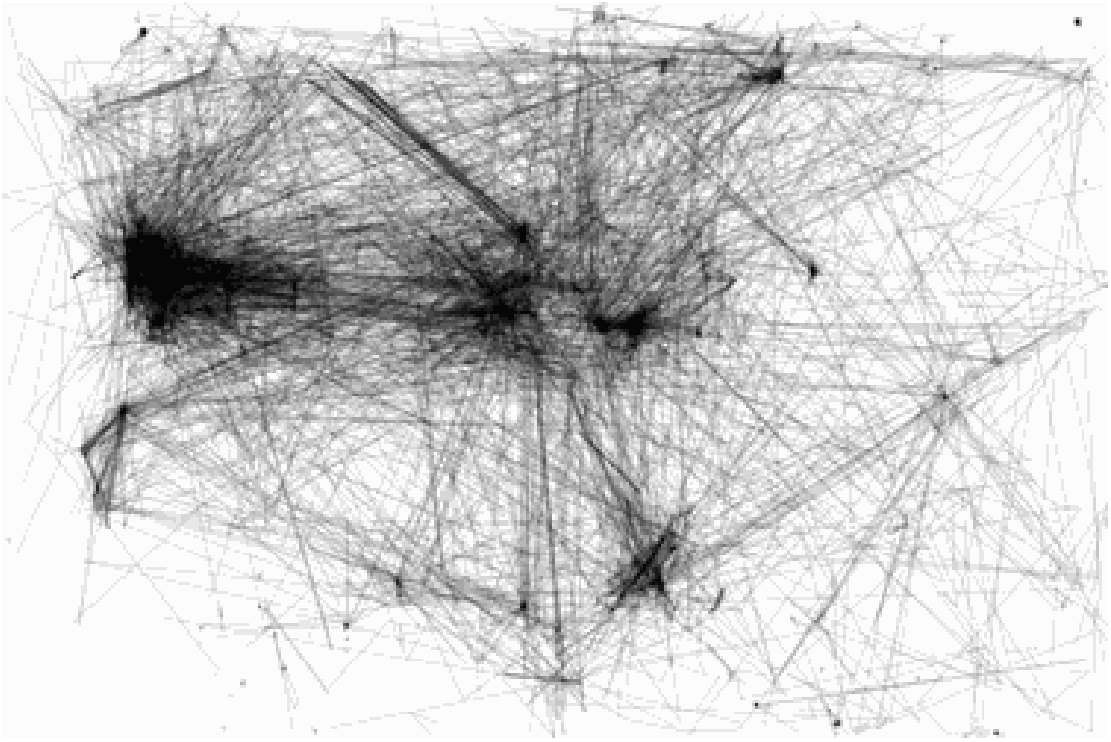

Figure 3.1.5 Real flows observed after filtering out fixed coordinated set by mobile applications.

3.2 Expenditure analysis from credit card records

3.2.1 Geographical information

Public dissemination of these results is pending on receiving explicit consent from the credit card data provider.

3.2.2 Card user information

Public dissemination of these results is pending on receiving explicit consent from the credit card data provider.

3.2.3 Businesses information

Public dissemination of these results is pending on receiving explicit consent from the credit card data provider.

3.2.4 Effect of seasonality

Public dissemination of these results is pending on receiving explicit consent from the credit card data provider.

3.2.5 Consequence of travel disruptions

Public dissemination of these results is pending on receiving explicit consent from the credit card data provider.

3.2.6 Spending behaviour the day of the trip

Public dissemination of these results is pending on receiving explicit consent from the credit card data provider.

4 Opinion and sentiment analysis

One of the objectives of BigData4ATM is to explore the viability of the use of online social networks data to assess the state of satisfactions of air transport travellers, especially in situations of strong traffic disruptions. Here, we will first show how the Twitter database we developed and used for the mobility analysis has very limited use for sentiment analysis. As a consequence, we propose and test a new data gathering strategy allowing the collection of a larger number of tweets at the occurrence of specific events, yielding statistical ensembles of the size of most surveys.

4.1 Geolocated tweets

The database of geolocated tweets developed by IFISC (see D2.1, Annex II.6) continuously gathers and records a sub-sample of the tweets shared in the micro-blogging platform. This sub-sample is by design limited to tweets with associated geographical information. While these data are of great use for tracking users' movements, its design is not optimal for opinion analysis. This is because a too large fraction of geolocated tweets do not consist of the typical micro-blogging content produced by Twitter users, but they are instead automatically produced by third-party apps (e.g., Foursquare) and lack personal expression. We verified this on an analysis targeting several of the main European airport hubs (AMS, BCN, CDG, FCO, FRA, IST, LHR, MUC, ZRH). In the period between November 2014 and August 2017, we gathered over 500.000 tweets in these airports. However, half of the tweets recorded inside the airports' areas are a product of Foursquare's Swarm App (https://www.swarmapp.com). Travelers arriving at the airport check-in to the airport's location in their phone app, and this is automatically shared via Twitter with a text only stating the venue where the user checked in. 24% of tweets are produced by Instagram, with a textual content characterised by too many hashtags and incomplete sentences that do not allow for semantic analysis. Lastly, 4% of tweets are links to other tweets. We are therefore left with only about 22% of those having real personal content to analyse. This fraction represents insufficient statistics for a sentiment analysis targeted to days of high airport's delays. For instance, focusing on the day within the interval of analysis where, using official DDR2 data, we identify the highest aggregated delay in the European airports (the 23rd June 2016, as a consequence of ATC industrial actions in France), we would have a total ensemble of 946 tweets recorded in the 9 airports listed above. After removing tweets from Instagram and Swarm, we are left with 159 tweets written by 41 users, which does not allow us to provide any conclusion on a statistical basis on the user's opinions.

4.2 Word-based queries

To gather a sufficient data on the opinion of users over a particular topic, it is necessary to implement a different data gathering strategy collecting tweets not on the base of their location, but querying for specific words in different European languages. With this infrastructure, we can capture up to 1800 recent tweets per hour (https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets), and collect historical tweets written in the last seven days (https://developer.twitter.com/en/docs/tweets/search/overview).

A drawback of this new strategy is that it cannot be automated. We cannot just wait and capture tweets from the stream containing too generic keywords – such as delay, cancelled, missed, waiting, rebook, seat, or boarding – because they are used in several different contexts (e.g., wait for a train, a plane, a dinner; cancelled flight, meeting; etc.). Instead, we have to target specific keywords or hashtags emerging in association to specific events. This imposes that the script must run exactly in the days the events are happening, as the historical reach is limited to only one week. Since the query is not aimed at geolocated tweets, only a very small fraction of the tweets collected will be accompanied by geographical information.

4.3 Sentiment analysis

In order to identify and study the attitude of the traveller when posting a tweet, it becomes necessary the use of text analysis, natural language processing and computational linguistics techniques. Existing approaches to these techniques, commonly known as sentiment analysis, can be grouped in three different categories: statistical methods, knowledge-based techniques, and hybrid approaches. Statistical methods use elements from machine learning, grammatical relationships of words, and grammatical dependency relations. Knowledge-based techniques assign affinities between words and emotions. Hybrid approaches leverage both machine learning and elements from knowledge-based methods such as semantic networks, and are especially adapted to detect subtleties and non-evident meanings.

In the case of sentiment analysis on tweets, there are some drawbacks. All these approaches are language specific. While they have been widely tested on English corpuses, the reliability of these methods on other languages is still in early stages. This can be avoided by translating non-English tweets to English by using external services such as Google Cloud Translation API (https://cloud.google.com/translate/). Besides, sentiment analysis precision and accuracy is high when working with middle-range and long correctly written texts (without many misspellings). The Twitter dataset has very particular characteristics: short messages with a maximum of 140 characters (280 from November 2017), usually with abbreviations and contractions (w/, btw, asap, ive, uve, im, …), hashtags, urls, smilies and other emoticons. Even though some approaches use spelling correction algorithms before applying sentiment analysis methods, data cleaning is advisable. As for that, we remove urls, hashtags, emojis and simple contractions before applying any opinion mining method.

The universe of sentiment analysis libraries ranges from pay per use Google Cloud Natural Language API (https://cloud.google.com/natural-language/) to open sources libraries such as NLTK (http://www.nltk.org/) and TextBlob (https://textblob.readthedocs.io/) that need to be locally trained. While NLTK is the leading platform to work with human language data, in our case, we relied on the last one, TextBlob, a simple API for diving into common natural language processing tasks such that wraps and simplifies NLTK.

We have tested two different implementations of sentiment analysis methods:

- Naive Bayes Classifier: being studied since the 1950s, it is a machine learning probabilistic classifier based on Bayes' theorem widely used for text classification. It relies on word frequencies as features. In our case, it has been trained on a dataset of movie reviews related

to opinion handling. The result of this algorithm for a given sentence is a (positive, negative)-tuple with the probability of the sentence being positive or negative.

- Pattern analysis (www.clips.uantwerpen.be/pages/pattern-en#sentiment): developed by CLiPS (Computational Linguistics & Psycholinguistics), a research centre associated with the Linguistics department of the faculty of Arts of the University of Antwerp, this implementation bundles a lexicon of adjectives (e.g., good, bad, amazing, irritating...) that occur frequently in product reviews, annotated with scores for sentiment polarity (positive-negative) and subjectivity (objective-subjective). The result of this algorithm for a given sentence is a (polarity, subjectivity)-tuple, based on the adjectives it contains, where polarity is a value between -1.0 and +1.0 and subjectivity between 0.0 and 1.0.

4.4 Proof of concept on Monarch Airlines' collapse

As a first experiment, we studied the potential of Twitter to evaluate passenger's satisfaction through an analysis involving the Monarch Airlines collapse on 2nd October 2017, when the airline entered ceased operations with immediate effect affecting more than four hundred thousand passengers (https://en.wikipedia.org/wiki/Monarch_Airlines#Administration_and_suspension).

We first gathered tweets between September 30 and October 10 containing the hashtags *#monarchairlines* or *#monarch.* As the date range is more than one week, we ran the script twice in ten days. After discarding duplicated, almost 18,000 tweets were retrieved only to notice that monarch was being used in very different contexts (Monarch Airlines, monarch Charles…). As for that, we narrowed the search by identifying the main hashtag used when mentioning Monarch Airlines, #monarchairlines, and we collected tweets in the same dates containing that hashtag. In this case, we gathered 4,600 tweets from 3,500 users, though the context was narrow enough.

At this point we cleaned the data to prepare it to be analysed. The process involved two different aspects. First of all, we remove non-human users, bots, by using the Botometer software (https://botometer.iuni.iu.edu) developed in the context of the OSoME project (https://osome.iuni.iu.edu/). This software computes the probability of a Twitter user being a bot; by setting the threshold to 0.5, we got rid of 315 users and our dataset was reduced to 4,099 tweets. On the other hand, as said above Twitter messages have to be cleaned by removing urls, emoticons, emojis, correcting contractions and removing the # symbol from hashtags. We did not correct misspellings. Finally, with these plain messages and the help of the widely used language detection software *Compact Language Detector 2* (https://github.com/CLD2Owners/cld2), we selected the subset of tweets written in English. The final dataset contains almost 3,800 tweets. In Figure 4.4.1, we can see the distribution of tweets over the time, noticing that the main peak, with more than a half of the data, corresponds to the day the company collapsed.

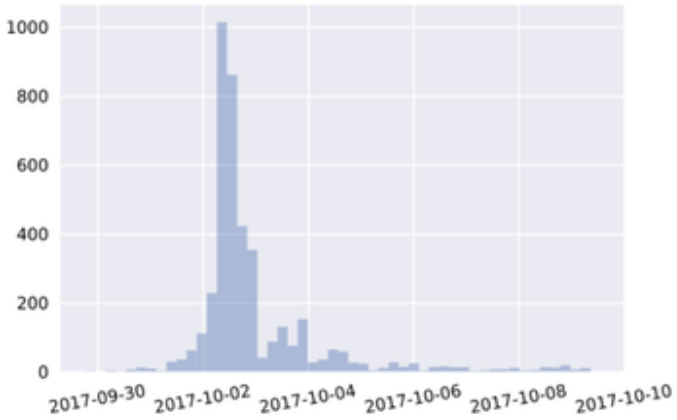

Figure 4.4.1 Histogram with the number of tweets detected containing the hashtag #monarchairlines.

Finally, we applied sentiment analysis methods to evaluate the state of satisfaction of Twitter users that used *#monarchairlines* hashtag between September 29 and October 9. The list of keywords used for stream filtering includes: delay, delays, delayed, cancel, cancels, cancelled, miss, missed, wait, waiting, rebook, rebooks, rebooked, seat, boarding.

Two different methods were tested: the naive Bayes Classifier trained on a dataset of movie reviews and the pattern.en analyser from CLiPS research group. The naive Bayes classifier result is the probability of the sentence of being positive or negative. The distribution of positivity for the naive Bayes Classifier for our dataset is shown in Figure 4.4.2.

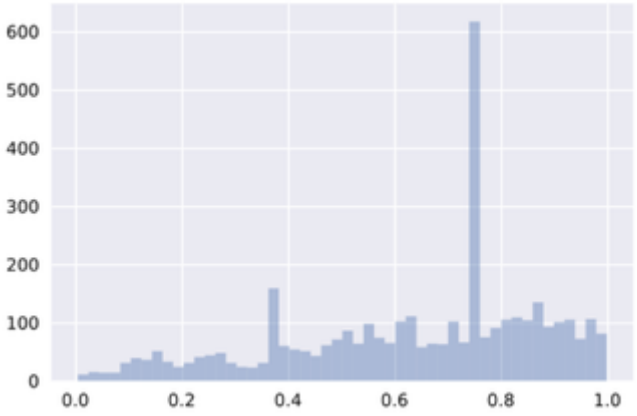

Figure 4.4.2 Histogram of the distribution of positivity of the tweets with the hashtag about the Monarch Airlines cease of operations.

As it can be seen, there are a few negative tweets and there is a large number of tweets with a probability of positivity of approximately 0.75. This seems to crash with the idea of people talking about a company ceasing activity. The pattern classifier result is a tuple containing the probability of the text being subjective and sentiment polarity (negative-positive) ranging from -1 to 1. The distribution of subjectivity and polarity with marginal distributions is displayed in Figure 4.4.3.

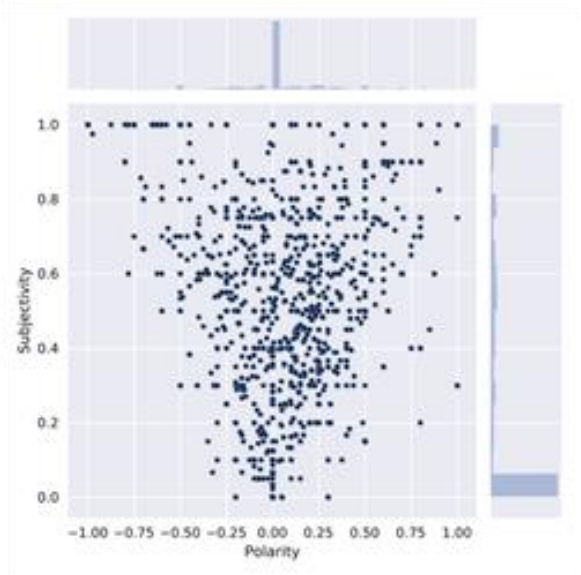

Figure 4.4.3 Polarity of the tweets on the Monarch Airlines collapse. The bars on the top and right are a histogram with the number of tweets in each bin.

51% of the tweets, more than 1900, receive both a neutral subjectivity and a neutral polarity. Objective tweets tend to be neutral on polarity. The higher the subjectivity, the higher the polarity. In particular, 24% of the tweets show a positive polarity and 19% a negative one, being most of them totally neutral.

By random inspection, we observe that while in some cases both algorithms agree on the positivity (It's been 8 years, two airports and some priceless memories and experiences. thank you @monarch #monarchairlines), other messages, mainly containing ironic statements, are not correctly classified. For example, “Great start to the week. Didn't want to go on holiday tomorrow anyway! cheers monarchairlines_uk” receives the maximum polarity and subjectivity and a probability of being positivy of 0.65. and “rt @thegreatswerve: @ryanair cancelled my flight so @opodo_fr booked me one with @monarch this is the worst tweet ever. #monarchairlines #are…” with maximum negative polarity, maximum subjectivity and a probability of being positive of 0.93. Apart from that, when considering tweets with non-zero sentiment in pattern method, the average sentiment is 0.048.

In conclusion, the new data gathering method is effective if specific keywords or hashtags are targeted. However, sentiment analysis tested methods are not reliable enough. Irony must be taken into account and a more systematic content check is needed. As for that, further study should focus on assessing whether the characteristics of tweets (short messages with misspellings, emojis and grammatical errors) should be implemented in the sentiment analysis. We propose creating a specific corpus on air traffic and opinion to train the naive Bayes classifier, or using deep learning techniques.

5 Conclusions and applicability of results

The main conclusion of the work presented here is that the new sources of data indeed bear enough information to allow us to satisfactorily face the next stages of the project.

In terms of mobility, we have introduced methods to obtain mobility flows out of mobile phone records and Twitter. Both data sources are very different. Regarding geographical resolution, mobile phone traces are obtained at the level of tower service areas while the tweets information can appear in a variety of scales from GPS coordinates, to POIs, neighbourhoods, cities or even countries. Given the fragmentation of mobile phone operators in the European context, the coverage of the mobile phone records is normally limited to a single country while the tweets appear with different penetration rates all across the continent. On the temporal side, mobile phone records sample the mobility patterns of the users every few seconds, while the tweets depend on when the user uploads new posts. The methods developed take into account all these characteristics to filter out problematic data and upscale the reliable information extracted out of these new data sources. This includes checking spatial consistency in towers, eliminating accounts with multiple users in Twitter, bots, etc. The resulting mobility flows have been confronted with survey information coming from Eurtostat or the Spanish Statics Office (INE) to validate them and also, in some cases, to adjust the thresholds that maximise the reliability of the final flows. Mobile phone records provide us with information to face the question of door-to-door mobility and mode choice, which will be treated in the door-to-door case study. Twitter data yield interesting results on transport demand at the continental level and opens the possibility of by blending the information with other external sources such as Google API to tackle the question of airport catchment areas, which will be the subject of a subsequent case study, and also in combination with mobile phone records the analysis of inside the airport mobility.

Even though with a more restricted time span, the credit card records allow us also to analyse how passenger expenditure patterns are modified by generalised delays in and close to the airports.

Finally, semantic-sentiment analysis on the tweets' contents, once expanded to all the streaming and not only to geolocated tweets, can offer extra insights on the passenger reactions to such problematic events.