Inventory and Quality Assessment of Data Sources for ATM Socioeconomic and Behavioural Studies

D 2.1
BigData4ATM
Grant: 699260
Call: H2020-SESAR-2015-1
Topic: Data Science in ATM
Consortium coordinator: Nommon
Edition date: 15 November 2016
Edition: 00.01.00
## Authoring & Approval

### Authors of the document

<table>
<thead>
<tr>
<th>Name/Beneficiary</th>
<th>Position/Title</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedro García / Nommon</td>
<td>Project Contributor</td>
<td>15/11/2016</td>
</tr>
<tr>
<td>Javier Torres / Nommon</td>
<td>Project Contributor</td>
<td>15/11/2016</td>
</tr>
<tr>
<td>Riccardo Gallotti / IFISC</td>
<td>Project Contributor</td>
<td>15/11/2016</td>
</tr>
<tr>
<td>Fabio Lamanna / IFISC</td>
<td>Project Contributor</td>
<td>15/11/2016</td>
</tr>
<tr>
<td>Bruno Campanelli / IFISC</td>
<td>Project Contributor</td>
<td>15/11/2016</td>
</tr>
<tr>
<td>Thomas Louail / IFISC</td>
<td>Project Contributor</td>
<td>15/11/2016</td>
</tr>
<tr>
<td>Gennady Andrienko / Fraunhofer</td>
<td>Project Contributor</td>
<td>15/11/2016</td>
</tr>
<tr>
<td>Carla Ciruelos / ISDEFE</td>
<td>Project Contributor</td>
<td>15/11/2016</td>
</tr>
<tr>
<td>Nicole Adler / HUJI</td>
<td>Project Contributor</td>
<td>15/11/2016</td>
</tr>
</tbody>
</table>

### Reviewers internal to the project

<table>
<thead>
<tr>
<th>Name/Beneficiary</th>
<th>Position/Title</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedro García / Nommon</td>
<td>Project Coordinator</td>
<td>17/11/2016</td>
</tr>
<tr>
<td>Ricardo Herranz / Nommon</td>
<td>Project Coordinator</td>
<td>17/11/2016</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Name/Beneficiary</th>
<th>Position/Title</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ricardo Herranz / Nommon</td>
<td>Project Coordinator</td>
<td>18/11/2016</td>
</tr>
</tbody>
</table>

### Rejected By - Representatives of beneficiaries involved in the project

<table>
<thead>
<tr>
<th>Name/Beneficiary</th>
<th>Position/Title</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Document History

<table>
<thead>
<tr>
<th>Edition</th>
<th>Date</th>
<th>Status</th>
<th>Author</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>00.00.01</td>
<td>11/11/2016</td>
<td>Draft</td>
<td>Pedro García</td>
<td>Initial draft</td>
</tr>
<tr>
<td>00.01.00</td>
<td>18/11/2016</td>
<td>Approved for submission to the SJU</td>
<td>Pedro García</td>
<td>Review and approval by BigData4ATM Consortium</td>
</tr>
</tbody>
</table>
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 an analysis of the data sources currently available for socioeconomic and behavioural research in ATM, including both traditional air transport and demographic data sources and new, unconventional passenger data generated by mobile devices. An analysis of these data sources has been carried out, identifying gaps on traditional and potential synergies with these new ‘Big Data’ sources. Finally, the research questions that will be tackled during next stages of the project are outlined.
# Table of Contents

**EXECUTIVE SUMMARY** ........................................................................................................... 5  

1 **INTRODUCTION** ................................................................................................................. 6  
   1.1 **SCOPE AND OBJECTIVES** ......................................................................................... 6  
   1.2 **REFERENCE AND APPLICABLE DOCUMENTS** ....................................................... 6  
   1.3 **LIST OF ACRONYMS** ............................................................................................. 7  
   1.4 **STRUCTURE OF THE DOCUMENT** ......................................................................... 8  

2 **OVERVIEW OF PASSENGER ORIENTED DATA** ...................................................................... 9  
   2.1 **TRADITIONAL DATA** ............................................................................................... 9  
   2.2 **BIG DATA: A NEW OPPORTUNITY** ......................................................................... 10  
      2.2.1 **Big Data sources identified** ............................................................................. 10  
      2.2.2 **State of the art in the analysis of user-centric Big Data sources** .................... 12  

3 **QUALITY ASSESSMENT METHODOLOGY** ........................................................................... 20  

4 **QUALITY ASSESSMENT RESULTS FOR TRADITIONAL DATA** .............................................. 23  
   4.1 **QUALITY ASSESSMENT RESULTS** .......................................................................... 23  
      4.1.1 **Passenger surveys** ......................................................................................... 23  
      4.1.2 **Aviation passenger intelligence solutions** ..................................................... 23  
      4.1.3 **Airfare data (MIDT data)** ............................................................................. 24  
      4.1.4 **CODA data** .................................................................................................. 24  
      4.1.5 **STATFOR data** ............................................................................................. 25  
      4.1.6 **EUROSTAT data** ........................................................................................... 25  
      4.1.7 **INE data** ...................................................................................................... 26  
   4.2 **CONCLUSIONS: GAPS AND ROOM FOR IMPROVEMENT** ...................................... 26  

5 **QUALITY ASSESSMENT RESULTS FOR BIG DATA SOURCES** ........................................... 28  
   5.1 **QUALITY ASSESSMENT RESULTS** .......................................................................... 28  
      5.1.1 **Mobile phone records** .................................................................................... 28  
      5.1.2 **Twitter data** ................................................................................................... 31  
      5.1.3 **Public transport smart card data** .................................................................... 33  
      5.1.4 **FlightRadar24 data** ...................................................................................... 34  
      5.1.5 **Anonymised credit card transactions** ............................................................. 36  
      5.1.6 **Google Maps APIs** ...................................................................................... 37  
      5.1.7 **TomTom data** ............................................................................................... 41  

6 **PASSENGER-ORIENTED BIG DATA SOURCES MAP** .......................................................... 43  

7 **RESEARCH QUESTIONS** ....................................................................................................... 44  
   7.1 **DOOR-TO-DOOR MOBILITY ANALYSIS** .................................................................. 44  
   7.2 **INTRA-AIRPORT MOBILITY ANALYSIS** ............................................................... 45  
   7.3 **EXPENDITURE ANALYSIS** .................................................................................... 45  
   7.4 **OPINION AND SENTIMENT ANALYSIS** ............................................................... 45  

8 **REFERENCES** .................................................................................................................... 46  

APPENDIX A **DATA CHARACTERISTICS FACTSHEETS** ....................................................... 49  

APPENDIX B **DETAILED CHARACTERISATION OF BIG DATA SOURCES** ................................. 66
Executive summary

The goal of BigData4ATM is to investigate how different passenger-centric geolocated data generated from mobile devices can be analysed to extract relevant information about passenger behaviour, and how this information can be used to inform air transport and ATM decision making processes. The purpose of this document is twofold:

- First, this document presents a review of different datasets, including both traditional and Big Data sources, in order to characterise their strengths, limitations and potential synergies.
- Second, the document outlines a set of research questions that will be addressed in the next stages of the project.

The following data sources have been reviewed:

- Conventional data sources: Sabre, IATA PaxIS, OAG, CODA and STATFOR public reports, EUROSTAT, national statistical offices (in particular the Spanish statistics agency, INE) and airport surveys.

For each one of these data sources, a factsheet with the following information has been filled:

- General information: name, link to the source, etc.
- Abstract
- Availability: owner, access conditions, cost, etc.
- Data characteristics: size of the sample, temporal and geographical scope, etc.
- Quality issues
- Other comments

Also, for the Big Data sources, an additional analysis has been conducted, with the aim to characterise aspects such as sample size and temporal and spatial characteristics.

Bases on this analysis, a number of research questions are proposed, which have been grouped into four topics:

- Door to door mobility analysis
- Intra-airport mobility analysis
- Expenditure analysis
- Opinion and sentiment analysis
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 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 ATM system, including the development of passenger-centric door-to-door delay metrics, the improvement of air traffic forecasting models, the analysis of intra-airport passenger behaviour and its impact on ATM, and the assessment of the socioeconomic impact of ATM disruptions.

To achieve these objectives, an analysis of the different existing datasets that are potentially relevant for the project has been conducted. The present document contains a review of both traditional data sources and ‘Big Data’ sources, assessing their quality, identifying their flaws, exploring synergies between them and ultimately identifying their usefulness for BigData4ATM. This document also proposes a list of potential research questions to be addressed during the next stages of the project.

1.2 Reference and applicable documents

- Grant Agreement No 699260 BigData4ATM – Annex 1 Description of the Action
- BigData4ATM D1.1 Project Management Plan, v00.02.00, August 2016

1 “The opinions expressed herein reflect the author’s view only. Under no circumstances shall the SESAR Joint Undertaking be responsible for any use that may be made of the information contained herein.”
### 1.3 List of acronyms

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

*Table 1: List of acronyms*
1.4 Structure of the document

The document is structured as follows:

- Section 2 presents an overview of the passenger data available from traditional air transport and ATM databases and the opportunities that Big Data offers as a source of passenger-centric information. The section also includes a brief literature review on the analysis of these Big Data sources.
- Section 3 presents the methodology followed for data quality assessment.
- Section 4 presents the results of the quality assessment of traditional databases.
- Section 5 presents the results of the quality assessment of the new ‘Big Data’ sources.
- Section 6 provides a visual guide of the information available from the different data sources.
- Section 7 contains a list of research questions identified as potential topics of interest to be tackled during next stages of the project.
2 Overview of passenger oriented data

2.1 Traditional data

Generally, the methods used to collect passenger information are based on surveys. Although surveys provide very rich information, these methods present intrinsic limitations (e.g., incorrect and imprecise answers, dependence on the willingness and availability to answer of the interviewed persons, etc.) and they are also expensive and time-consuming. For these reasons, surveys are conducted with a frequency and a sample size lower than desired, and even for some airports, they are not conducted at all due to budget limitations.

Another source of passenger information are market intelligence data services (such as IATA Passenger Intelligence Services, PaxIS). These services provide aggregated passenger information detailing origin and destination airports, fare data, connecting points, etc. However, these data have a low temporal granularity (they are typically monthly aggregated) and fail to capture important information, such as door-to-door origin-destination pairs and travel times.

EUROCONTROL, through CODA (Central Office for Delay Analysis), provides information on the air traffic delay situation in Europe. This information is supplied to CODA by aircraft operators and ATFM (Air Traffic Flow Management data from EUROCONTROL Network Manager).

EUROCONTROL STATFOR (Statistics and Forecast Service) provides statistics and forecasts on air traffic in Europe. These forecasts are used by an extensive number of planning departments of airlines, air navigation service providers (ANSPs), airports, government authorities, etc. for general planning.

EUROSTAT, the European Statistical Office, also plays an important role as a source of information for air transport. EUROSTAT not only provides official statistics, but also promotes the harmonisation of the statistical methods followed by European countries, allowing the comparison of data across different regions. EUROSTAT air transport database contains low temporal granularity (monthly aggregated) passenger and traffic information.

<table>
<thead>
<tr>
<th>Information</th>
<th>Data source</th>
</tr>
</thead>
</table>
| Passenger information | • Travel surveys  
                      | • Aviation passenger intelligence solutions (Sabre, OAG, IATA PaxIS)  
                      | • EUROSTAT         |
| Traffic statistics & forecasts | • STATFOR                              |
| Delay            | • CODA                                            |

Table 2: Potential uses for traditional data sources
The role of these traditional data sources within the BigData4ATM will be twofold:

- As a first step, they will be used for validation purposes, in order to assess the goodness of the results provided by methodologies based on unconventional data.
- Once novel methodologies are validated, traditional data sources will be merged with unconventional data to provide enhanced passenger-centric information.

2.2 Big Data: a new opportunity

During the past few years, the term ‘Big Data’ has become very popular due to a variety of reasons:

- **Pervasive use of data-generator devices**: nowadays, the amount of data generated in our daily lives is huge. Smartphones, credit cards, public transport smart cards and many other devices produce digital traces that enable new ways of inferring individual behaviour, preferences, mobility patterns, etc.
- **Technological evolution**: recent advances in data processing and storage technologies, such as Cloud Computing, Hadoop or Spark have allowed the analysis of great volumes of data at speeds compatible with the current data generation rates.
- **Technical progress**: advances in the artificial intelligence and machine learning fields are resulting in faster, more reliable algorithms that provide richer information than before.

ATM, due to its inherent complexity, has all the ingredients to become a Big Data success case. The amount of data involved in a typical flight (route data, weather data, passenger data, etc.) is immense and virtually impossible to be analysed through traditional approaches. This makes Big Data methods and tools good candidates to provide a deeper knowledge of the ATM system and ultimately to act as a useful tool for decision making.

2.2.1 Big Data sources identified

In BigData4ATM, it will be explored how to exploit this new opportunity to enrich traditional passenger information. The following data sources have been identified as potentially useful for the project:

- **Anonymised mobile phone records**: Call detail records (CDRs) provide location information about millions of users with high temporal and spatial resolution. CDRs are the registers generated when a mobile phone connected to the network makes or receives a phone call or an SMS, or connects to the Internet. For invoicing purposes, the information regarding the mobile phone antenna to which the user was connected when the call or service was initiated and ended is logged, providing an indication of the geographical position of the user and its timestamp. Mobile phone data can be used to obtain door-to-door origin-destination matrices and information about the access/egress modes of the airport.

- **Internet social networks**: Twitter, through its streaming Application Programming Interfaces (APIs), provides access to global tweet data. Semantic analysis can be applied to tweets in order to get information about social perception of ATM, analyse information propagation during disruptions, etc. In addition, there are some tweets that contain information about the location of the mobile device at GPS level at the moment when the tweet was sent. This may also allow mobility analysis to be performed, studying flows between airports and reconstructing passenger complete door-to-door trips.
• **Public transport smart card records**: smart cards used in public transport generate records that contain information about timestamp, mode and stop where the smart card was used. This information can be used to characterise urban mobility, identify peak hours and analyse/model behaviour during disruptions, maintenance works, etc. In the context of BigData4ATM, public transport smart card records can provide information on the access/egress modes of the airport, in order to develop indicators of airport accessibility and carry out intermodality studies.

• **Anonymised credit card transactions**: this dataset contains information about the purchases made by bank clients in any point of sale (POS) or all the purchases made in any POS belonging to the bank. The information contained consists of the amount spent, POS type (type of business), location and timestamp of the purchase. Some sociodemographic information (age, gender, zip code) is also available for bank clients. The main application of this data source for the project is to evaluate the impact of disruptions on tourist expenditure (both inside and outside airports). Also, it can help evaluate the positive economic impact (in terms of tourist expenditure) of new flight connections.

• **Google Maps APIs**: Google Maps offers a variety of APIs that provide map/location based information, such as routing information and travel times for different transport modes (including public transport) for an origin-destination pair. The information that can be accessed through these APIs will be very useful both on its own, to calculate accessibility indicators for airports, and in combination with other data sources, to enhance mode and route determination algorithms.

• **TomTom data**: TomTom offers a variety of services that provide traffic information, such as real-time and historical traffic data (speed, free-flow speed, number of vehicles) and routing services. The uses of these data are similar to the uses of Google Maps. The main differences are that TomTom provides historical information, but it does not include public transport.

• **Wi-Fi/Bluetooth beacons**: beacons use Wi-Fi or Bluetooth signals to detect other devices and determine their location, in order to carry out location-based actions such as marketing-oriented actions, send notifications, or guide users through a place. Some airports have started deploying this technology and embedding it in their mobile phone Apps in order to provide airport-related information and send personalised alerts. This information can be used to identify kerb-to-gate mobility patterns, which can provide insights on how to improve passenger experience (e.g., sending real-time information about airport services, optimising the allocation of resources by identifying bottlenecks, etc.) and optimise non-aeronautical revenue (e.g., identifying customer journeys and sending personalised offers, optimising commercial layout, etc.). At the moment of writing the present document, access to this kind of data is being negotiated, but it has not yet been granted for the BigData4ATM project.

• **FlightRadar24**: FlightRadar24 is an online database that uses ADS-B receivers to track flights all over the world in real-time. In their database, they provide flight route information, with position at several timestamps, airport arrivals/departures data, with landing and departure times, and delay information. Access to historical data can also be achieved by using a premium account or by collaborating with their network of ADS-B receivers. Although, unlike the abovementioned sources, these are not passenger-centric data, we have included it here because it can be considered a non-conventional data source, based on crowd sourcing, which can be used if access to restricted, official data sources is not achieved.
**Information** | **Data source**
--- | ---
Passenger mobility patterns | • Mobile phone records  
• Internet social networks  
• Public transport smart card  
• Wi-Fi/Bluetooth beacons
Passenger opinions and sentiments | • Internet social networks
Passenger activity and expenditure | • Credit card
Air traffic data | • FlightRadar24
Routing/location based information | • Google Maps APIs  
• TomTom data

*Table 3: Potential uses for ‘Big Data sources*

### 2.2.2 State of the art in the analysis of user-centric Big Data sources

In this section, a review of the state of the art for the general methods used to obtain knowledge from different sources of geolocation information is presented. These sources can be [1]:

- **Time-based:** positions of users are recorded at regularly spaced time moments.
- **Change-based:** a record is generated when a user’s position, speed or movement direction differs from the previous one.
- **Location-based:** a record is made when a user enters or comes close to a specific place, e.g. where a sensor is installed.
- **Event-based:** positions and times are recorded when certain events occur, in particular, when users perform certain activities such as mobile phone calls or sending a tweet.
- **Various combinations of these basic approaches:** for example, GPS tracking devices may combine time-based and change-based recording; the positions are measured at regular time intervals but recorded only when a significant change of position, speed, or direction occurs.

Some of the data produced by the data sources identified in 2.2.1, such as credit card records, fit into the category of ‘episodic movement data’ [2]. This term refers to data about spatial positions of moving objects where the time intervals between the measurements may be quite large and therefore the positions between the measurements cannot always be reliably reconstructed by means of interpolation, map matching, or other methods. Such data can also be called ‘temporally sparse’; however, this term is not very accurate since the temporal resolution of the data may greatly vary and occasionally be very high.

Irrespective of the collection method, the common type of uncertainty in any episodic movement data is the lack of information about the spatial positions of the objects between the recorded positions, which is caused by (sometimes) large time intervals between the recordings. Another frequently occurring type of uncertainty is imprecision of the recorded positions. For example, a sensor may detect an object within its range but may not be able to determine the exact coordinates of the objects. Another example is a mobile phone record that is located within the coverage area of a certain antenna, but cannot be associated to an exact point in space.
Due to these uncertainties, episodic movement data cannot be represented as continuous trajectories, i.e., lines in the spatio-temporal continuum where known (measured) positions are connected by straight or curved segments. Many of the existing visual and computational methods designed for dealing with movement data are explicitly or implicitly based on interpolation of objects’ positions between the measured positions and are therefore not suitable for episodic data. Methods for finding patterns of relative or collective movement of two or more objects (e.g., meeting or flocking) also require fine-resolution data.

Since many of the existing methods are not applicable to episodic movement data, there is a need to find suitable approaches for the analysis of this kind of data. When episodic data are available for a sufficiently long period, it is possible to use large-scale datasets for identifying and semantically interpreting individual and public places ([3],[4]). The use of spatio-temporal aggregation can be seen in references [5] and [6], where the data are aggregated in two complementary ways:

1) **By places**: for each place P (e.g., territory compartment or sensor location) and time interval T, the data about the objects that were registered in P within T are aggregated. The aggregate attributes may be, for example, counts of objects or average times spent in the place.

2) **By pairs of places**: for two places P1 and P2 and time interval T, the data about the objects that moved from P1 to P2 within T are aggregated. The aggregate attributes may be, for instance, counts of objects or average durations of the transitions.

Reference [7] provides an overview of applications of episodic movement data analysis methods. The models for pedestrian monitoring distinguish between microscopic and macroscopic aspects of mobility [8]. Whereas microscopic models describe individual behaviour and provide their trajectories, macroscopic models aim at modelling the moving population and use values such as density, quantity, or speed to characterise pedestrian flows. Both of these views on movement are closely related as macroscopic values can be derived by aggregation from microscopic ones [8].

### 2.2.2.1 Mobile phone records

Nowadays, when almost everyone has a mobile phone, the amount of data that is generated through these devices every second is massive. During the last 10 years, a considerable amount of research has been carried out to analyse these data, and several new techniques and applications have appeared. In this section, a small summary of the state of the art of the extraction of mobility information from mobile phone data is presented.

There are two different kinds of mobile phone data that can be used as a source of information: Event-Driven Cell Phone Data (Call Detail Records) and Network-Driven Events. Call Detail Records (CDRs) are generated every time a mobile phone interacts with the network through a voice call (receiving or sending), SMS or Internet connection. These records are stored by the Mobile Network Operator (MNO) for billing purposes. Network-Driven Events do not require any interaction from the mobile phone user, as they are generated by the MNO for network managing purposes. The availability and characteristics of these Network-Driven Events depend on the MNO, as some operators may store information at different rates or spatial resolution (cell/groups of cells), or not store any information at all. In this project, Network-Driven Events are not being considered, at least for the moment, as the data provision agreements signed so far only cover CDRs.
The format of CDRs is not fully standardised, and it may vary across different MNOs. However, it will always contain at least the following information:

<table>
<thead>
<tr>
<th>User 1 ID</th>
<th>Origin cell ID</th>
<th>Date</th>
<th>Time</th>
<th>Duration / Mb</th>
</tr>
</thead>
</table>

- **User 1 ID**: it is a code that identifies the mobile phone user that connects to the network because of an activity (voice call, SMS or data connection). Before providing CDRs to a third party, it has to be anonymised because of privacy concerns, usually through a hash function.
- **Origin cell ID**: this number identifies the mobile network cell to which user 1 connects.
- **Date**: the date when the event is registered.
- **Time**: the time when the event starts.
- **Duration / Mb**: the duration (for voice calls) or the amount of data (in case of Internet connections) of the event.

With the fields highlighted above, it is already possible to extract information about users’ activity and mobility. Sometimes, CDRs may also include the user ID and/or the cell ID of the receiver of the call/SMS and/or the cell ID where the event finishes and/or an identifier of the type of event. This can be useful to extract more information. For example, if receiver ID and location are available, knowledge about the social network of the population can be inferred [9]. Also, the information about whether an event is generated by the user or not may be useful to complement activity detection models.

**Location information from CDRs**

When no additional information is provided and only cell ID and its coordinates are available, cell coverage areas are usually approximated through a Voronoi tessellation created from the network cell locations. This is done under the assumption that the user will connect to the closest tower. The user location can be approximated through assumptions such as the following:

- The user is at the tower coordinates.
- The user is at the centroid of the Voronoi polygon (or cell sector, if information is available).
- The user is in a specific area inside a Voronoi, based on land use characteristics.
- The user is at a “central” or “relevant” node of the road network that falls inside the Voronoi polygon.

The accuracy of this assumption depends on the cell size, which in turn depends on the network infrastructure. In urban areas, where cells are smaller, these methods will have smaller error than in rural areas. Typically, the order of magnitude of the error in urban areas is expected to be between 100-1000m, while in rural areas the error is between 450-2000m [10].

If the signal strength that the mobile phone receives from the surrounding towers or the time delay between “sent-received” signals is taken into account, it is possible to obtain the triangulated position of the mobile phone. This needs the cooperation of the network operator, as this triangulated position is not currently being recorded (and most possibly, even not calculated) by all MNOs. Several studies have used triangulated information ([11], [12], [13], [14]). For the data used in these studies, the average uncertainty in the position has an average value of 320m, with a median of 220m.
Inferring common locations/activities from CDRs

Once the location information has been extracted from the CDRs, it is possible to start extracting information about a user. Although CDRs may not capture every single detail of a user’s daily activities (due to their relatively low temporal sampling), they are a very powerful source of information to obtain common stay points, such as workplace, home, or other frequent activities/places, and the time intervals when a user stays in each of them. This information allows the estimation of an approximate distribution of the residential and working areas within a city, and the flows of people travelling between them.

Typically, the expansion of the data to the full population is done through the market share of the network operator that provides the data, by comparing with an external source (such as the census, or toll counts) or by a combination of both. To compare with the census, the number of people from the study population that resides in an area has to be obtained (as it is possible to determine where a person lives by studying his activity patterns) and then this number is compared to the census data.

General mobility models

Once common locations, activities and trips from users have been inferred, it is possible to obtain more accurate models for human mobility and to evaluate population flows and travel patterns. Several studies have been published about this topic, and interesting conclusions have been extracted from them. In [15], for example, a model that integrates data from mobile phones with universal mechanisms (preferential return, burst of activities and circadian rhythms) from human mobility is presented. In [12], mobile phone data were used together with vehicle odometer readings to analyse individual mobility and its relationship with environmental variables, such as population density, public transport, land use mix and accessibility to workplaces.

Origin-destination matrices

Usually most of the information that transport planning authorities require to make efficient decisions is stored in the so-called Origin-Destination Matrices. An OD Matrix contains the number of trips between two different regions, for several temporal windows. OD Matrices may vary in terms of spatial resolution (for example, they may store information about trips between tracks or districts inside a city, or even between different cities) or in temporal resolution (hourly, daily, seasonal...).

The usual process to obtain OD matrices is the following:

- Starting from a series of time-location measurements for each user, consecutive locations (from the same user) that are close together are clustered, and considered “stays”.
- Once “stays” are determined, “trips” are all those time-location measurements that are detected between user’s consecutive stays.
- The geographical area under study is divided in regions (the size of the regions is a compromise between resolution and accuracy, as explained before).
- Trips are parameterised by user, origin region, destination region and time, and grouped at different temporal windows, obtaining the sample OD Matrix.
- The sample OD Matrix is expanded to the full population to provide meaningful info.

In [11], the OD Matrix from the metropolitan Area of Boston is calculated from triangulated CDRs, and validated against census survey data. Another example of a study that calculates the OD Matrix
for a big city is [13], where the OD Matrix for Dhaka, Bangladesh is obtained from CDRs that provided location information at cell-based level. As before, user locations (tower locations) were grouped into regions (traffic network nodes, in this case) to calculate the OD Matrix. The study was validated with 13 traffic count locations, with a Root Mean Squared Percent Error of 13.59%.

**Transport mode and route determination**

Transportation mode inference is another of the potential fields where mobile phone data information could help. This kind of information is very useful when public transport authorities try to determine the effectiveness of their policies, or when trying to analyse competition between two different transport modes (plane vs high-speed train, for example). Route assignment is a very similar problem, as different transport modes often imply different routes, and therefore determining one can help determine the other.

In [16], a brief review about methods for inferring transportation mode by estimating travel speed is presented. Different transportation modes usually have different speeds and different characteristics (for example, public transport usually has more stops than private transport). Travel speed can be measured by variance of signal strength, or through switch rate of cells. However, this method is not able to distinguish between modes that have very similar speeds, such as car and buses, for example. Also, it is very difficult to calculate actual travel speed or stops through CDRs. Another approach they introduce, which is very similar, is to infer mode by travel time. Given a trip, different travel times for each transportation mode are estimated through Google Maps, and then compared with the CDR data. This method has a potential flaw, as accurate travel time estimation through CDRs is only possible if enough phone activity is recorded.

An alternative method to determine the transport mode is to combine the data from mobile phones with other sources of data, such as public transport smartcards. This was done in [17], where the market share of public transport vs private transport in Singapore was calculated. They had access to public transport smartcards, which recorded events both when getting on and off from public transport. This way, they had access to the overall OD matrix (through CDRs) and the public transport OD matrix (through smart card data). Overall results showed a good agreement with surveys, but at intra-district level no validation could be carried out. Another important outcome from this paper is that it was possible to evaluate the weak points in public transport connections, by checking the busiest places and where the private transport had a higher share. It is also remarkable that this study used cell-based location information, not triangulated position.

Another interesting study about mode determination from CDR data is presented in [18] and [19]. A method to evaluate route/mode working with Voronoi location information is shown. The method is based on assigning probabilities of connecting to different towers depending on the route followed. It has to be noted that, due to jumps in tower connection, this method of assigning route may not be suitable for every case, as routes may share several Voronoi areas and CDRs do not always provide detailed temporal sampling.

**Tourism**

Tourism is another sector where this kind of technology can be used. Knowing the principal access points of a city, or the most visited places and the routes tourists follow, can help to make decisions that improve the attractiveness of a city, increasing its revenues. An example of the kind of study that can be conducted is [20]. There, tourism information such as tourist nationality, duration of the stay, places visited and attendance to events was extracted for the city of Tartu, Estonia. A very
detailed report about the extraction of tourism statistics from CDRs is [21]. One of the problems about extrapolating the data from roaming into total number of visitors is the lack of reference information. Tourists are not in a census, or measured by the mobile network operator market share. Therefore, alternative sources of reference information have to be used.

2.2.2.2 Twitter data

The use of Twitter geolocated data to study human mobility started in 2011, two years after the “explosion” of the Twitter user base in 2009. An important issue is the non-homogeneous penetration rate, which changes from one region to another for economic and social reasons. The first analysis performed with this kind of data aimed at identifying singular events, i.e., mass gathering that does not repeat in time. In zones of high penetration, it has also become possible to aggregate Twitter georeferenced data in order to acquire information on regular activity patterns in a specific area, on urban mobility and interurban travels. Once corrected for the non-homogeneous penetration rate, international flows can be derived to study migratory patterns or to characterise international tourism.

Penetration rate

An important thing to take into account when using Twitter data to study human mobility is that the micro-blogging platform has a limited user base, and among the users only a fraction of them posts georeferenced tweets. The georeferenced twitter penetration rate can reach values of the order of 0.1%, however it shows a very wide heterogeneity, varying by several orders of magnitude among countries worldwide [22], as well as within the same country [23]. At a large scale, the penetration rate is correlated with per capita GDP, scaling super-linearly with a country’s economic development [22]. Within Europe, the penetration rate of georeferenced tweets shows very little dependence on GDP, and the broad differences observed are most probably associated to sociocultural differences [23].

Any measure of flows at national or international level should necessarily be rescaled to compensate for those penetration rates. Among Twitter users, millennials are notoriously overrepresented. By using face recognition software on the photo profile, it has been estimated that the population pyramid of Twitter users shows also a gender asymmetry: 63% of the users whose photo could be evaluated are classified as males. Also, males have an average age of 26 years, whereas for females the average age is 20 [24].

Singular events

One can automatically identify the location of points-of-interest [25] or detect real-time local events [26]. This method can be useful to quantify the dimension of assemblies of people such as the number of attendees in a football match [27] or in political protests [28].

Regular activity patterns

By aggregating data on different days one can identify typical activity profiles for different areas at different hours of the day ([25], [29], [30], [31]). A preliminary validation analysis associated to air transport has been made by evaluating the correlation between the Twitter activity in the Linate airport of Milano and the number of flights associated to that hour of the day [27].
Urban mobility

Twitter data can be used to determine the dominant land use of an urban area from the tweeting activity performed at different hours in workdays or weekends [25]. Taking advantage of the regularity of human mobility patterns, one can identify the location of a user’s home and that of a predominant regular daily activity (usually identified as the workplace) and reconstruct a home-work commuting origin-destination matrix that is approximatively proportional to those obtained from surveys [29]. For very active users, a set of locations (on average 3 per person) where tweets are regularly made can be identified in order to study the characteristics of the individual mobility networks [30]. Trying instead to identify the home and work locations on the basis of the semantical content of tweets yields worse results, in particular because of the incorrect identification as home of transit stops [31].

National travel

If the penetration rates are high enough (of the order of 0.1%), one can try to estimate the average passenger flows on different segments of interurban roads and on the train network as the number of tweets observed in close proximity of the road or the railway. These averages appear however to be poorly correlated with real traffic flows because shorter road segments tend to be systematically under-sampled [23]. On a more aggregated level, one can evaluate internal migration trends [24], national mobility maps [32], or the evacuation due to natural emergencies [33], however this kind of analysis have not been yet validated.

International travel

Twitter has been pointed numerous times as a possible source of information for global mobility and migration trends, useful in particular in developing areas where no better information is available ([24], [32], [34]).

The volume of international travellers arriving in a country estimated from tweets, by aggregating at national level and rescaling accordingly to the penetration rates at the origin, correlate with data provided by the World Economic Forum [22]. The analysis has been limited to the countries with a penetration rate above 0.05% and the number of arrivals are thus underestimated by a factor of 3. Probably because of the existing relation between penetration rate and GDP, a better correlation can be observed between the number of arrivals estimated and aggregated tourist’s expenses in that country [22].

A similar kind of analysis can be made by selecting a subset of important locations. It is possible, for instance, to identify the flows of international tourists between, and evaluate the international attractiveness of, touristic cities [36] and touristic sites [37]. In these studies, the attractiveness ranking has been defined in three alternative ways: i) the distance travelled to reach the destination; ii) this same distance, divided by the average distance of Twitter users from the location, to account for the inhomogeneous distribution of population on the world surface; iii) the area covered by the visitors’ place of residence. Comparing different Spanish areas, a super linear scaling (exponent ~1.5) between the number of visitors estimated via Twitter (among other data sources) and the city population has been observed [37].
2.2.2.3 Credit card transactions data

All studies on credit card data are very recent (2014 onwards) and many of them are centred on Spain, since this kind of data has been provided for research purposes by a local bank. Most work with this kind of data addresses the analysis of individual spending behaviour and how these data can be used to extract economic indicators at urban level.

Individual behaviour

From the sequence of locations where a transaction occurred, one can extract mobility information at interurban [38] and urban [39] scale. Factors like age are important, with younger people spending less and traveling further for every transaction [39]. This last trend is different between men and women [40]. Average income per business changes widely between neighbourhoods in the same city [41]. The spending patterns can also be strongly influenced by specific events: for the case of a hurricane in Mexico, it has been observed that the lower the income level, the shorter the time needed for economic activity to return to normal levels, with males recovering faster than women [42].

Economic indicators and city scale

The spending behaviour aggregated at urban level allows the relation between the number of transactions and city size to be estimated. The activity of foreigner visitors has a manifestly superlinear (with exponent ~1.5) scaling with the size of the city they visited [37], the log-residuals of this scaling being a reasonable indicator of tourist attractiveness. Locals present an almost linear scaling (1.05) with city population, and its meaning as economic indicator has been confirmed by verifying its correlation with household expenditures [40] and other economic indexes [43].
3 Quality assessment methodology

After identifying the data sources that are potentially useful for BigData4ATM, it is necessary to analyse, classify and assess the quality of the information provided by each one of them. This analysis has been performed by filling, for each data source, the factsheet presented in Figure 3.1. The factsheets corresponding to each data source can be found in Appendix A.

The factsheet contains the following information:

1. **General information**: identification of the data source and contact/support information.
   - 1. Data source name
   - 2. Last update of the factsheet
   - 3. Contact information
   - 4. Support information

2. **Abstract**: brief description of the data source and its potential usefulness for the project.

3. **Availability**: relevant information about owner and readiness to use.
   - 1. Owner
   - 2. Access conditions
   - 3. Mode of access to the data
   - 4. Data cost
   - 5. Access limitations
   - 6. Availability within the project
   - 7. Privacy/Confidentiality issues
   - 8. Security issues
   - 9. State
   - 10. Link to the data

4. **Data characteristics**:
   - 1. Estimated size of the sample
   - 2. Temporal scope
   - 3. Geographical scope
   - 4. Temporal granularity
   - 5. Geographical granularity
   - 6. Delivery frequency
   - 7. Delivery delay
   - 8. Data format

5. **Quality issues**: comments about known flaws of the data source

6. **Comments**: other relevant information
### Data Identification, Collection & Assessment

#### Potential Data Sources Characteristics Spreadsheet

1. **General Information - Identification of the Data Source**
   - **Data source name**
   - **Last update of this file**
   - **Contact information**
   - **Support**

2. **Abstract - Brief description of the data source and its potential usefulness for the project**

3. **Availability - Relevant information about owner and readiness to use**
   - **Owner**
   - **Access conditions**
   - **Data cost**
   - **Access limitations**
   - **Availability within the project**
   - **Privacy / Confidentiality issues**
   - **Security issues**
   - **State**
   - **Link to the data**

4. **Data Characteristics - Temporal, geographical and size characteristics of the information provided**
   - **Estimated size of the sample**
   - **Temporal scope**
   - **Temporal granularity**
   - **Geographical scope**
   - **Geographical granularity**
   - **Delivery frequency**
   - **Delivery delay**
   - **Data format**

5. **Quality Issues - Comments about known flaws of the data source**

6. **Comments - Other relevant information**

---

**Figure 1: Data Factsheet template**

---

**Founding Members**


All rights reserved. Licensed to the SESAR Joint Undertaking under conditions.
Additionally, for the ICT-based data sources, an in-depth analysis was carried out with the objective of better evaluating the usefulness of each data source for the project. The reason for performing this analysis is that, as these are new data sources that have not been used until recently, they are not as well characterised as the traditional data sources.

Although there is a great variety in the origin of the analysed ICT-based data sources, most of them have in common that they are generated by an individual person, and therefore the following detailed information could be obtained:

- Sample characteristics: a detailed characterisation of the sample, including its gender, age and other relevant information when available.
- Temporal characteristics: a characterisation of the temporal granularity of the individual information provided by the data source.
- Spatial characteristics: a characterisation of the level of the spatial resolution of the data source.

The complete analysis for these data sources is presented in Appendix B.
4 Quality assessment results for traditional data

4.1 Quality assessment results

In this section, a summary of the main results of the assessment for each one of the different traditional passenger-centric data sources is presented. This summary includes the main applications and issues for each data source. The factsheets with the complete results of the data quality assessment can be found in Appendix A.

4.1.1 Passenger surveys

Some of the information that BigData4ATM project wants to gather and analyse has been traditionally collected by surveys made by the airports to the passengers. Passenger surveys usually gather information related with the passenger profile, the airports’ catchment area, the access mode to the airport, the connecting points of the passenger and their shopping habits at the airport. An example is that of the EMMA (Estudios de Movilidad en Modo Aéreo, Aerial Mode Mobility Studies) surveys performed by AENA. These surveys are performed in the boarding queues, which provides a sampling covering all the airport destinations and makes it possible for the airport to introduce specific questions in the case of recent developments: new terminal, special events, new high speed train line, etc.

When looking at flaws and limitations of these data, it needs to be remarked that due to the cost of these surveys, they cover only one week, and are performed every year in the main airports (Adolfo Suárez Madrid-Barajas and Barcelona-El Prat) and every three to five years in the other airports of the Spanish network. This reduces the validity and representativeness of the surveys and make it quite difficult to have information for unusual days (bad weather conditions, strikes, big delays) except in the case where these events overlap with the days when surveys are being carried out.

ASQ (ACI’s Airport Service Quality) surveys are another source of information. In this case, the information gathered is similar to the previous one, but they also cover the passengers’ quality perception. These surveys are performed quarterly in the member airports, hence most of the big and medium airports in the world are included in the sample.

4.1.2 Aviation passenger intelligence solutions

Three aviation passenger intelligence data providers have been analysed: OAG, Sabre and IATA PaxIS. The three of them provide the same kind of information: number of passengers for each route, aggregated by month, with connecting airports, fare data and class. This information is obtained
from flight ticket sales from the travel agencies and, through some internal processes, these ticket sales data are cleaned, corrected, extrapolated (as information of the tickets purchased directly to airlines and charter flights is not directly available through Global Distribution System, GDS data) and aggregated.

IFISC and ISDEFE, two of the members of the project Consortium, have already worked with Sabre data, and that is why finally this provider was chosen. The dataset already collected by IFISC and ISDEFE has a temporal scope of 2009-2015, with data for all over the world for more than 140 million routes.

This source of data will be useful as a validation source for the new data sources. Also, as it contains connecting airports, it can be used to compute total flight delay metrics. However, due to the low temporal granularity of the data (monthly aggregated), the usability for studying disruptions or specific events is very limited, forcing to use interpolation methods.

4.1.3 Airfare data (MIDT data)

Air ticket prices information may be of interest for the project: they can be used for air traffic forecasting purposes or to study if there is a relationship between airfares and delay/congestion. Marketing Intelligence Data Tapes (MIDT) provide information on one-way airfares for tickets bought through a Global Distribution System (GDS). There are three large GDSs in the world (Amadeus, Sabre and Travelport), with their market share depending on the area of study. Data purchased is typically monthly aggregated, although depending on the agreement reached with the data provider, a higher temporal resolution may be obtained (involving a higher cost). Spatial granularity is at airport level, including connecting points.

The raw data can be very problematic (missing data, flights with misleading/no price, employee discounts/frequent flyer programs...). This is why, when purchasing this information, it may be interesting to purchase corrected information (either by the GDS or through a re-seller) instead of raw data. Also, data has a very low granularity (airport level, monthly aggregated), which is helpful for direct analysis, but limits the amount of post-processing or customised analysis. The decision to purchase this kind of data will be made depending on the research questions decided and their temporal and geographical scope. As this kind of information is already included (and corrected) in the datasets from Sabre, it will not be necessary to purchase it except if an analysis for a period of time not covered by the available dataset is to be conducted.

4.1.4 CODA data

One of the case studies proposed in the BigData4ATM project is the development of passenger-centric delay metrics which effectively capture how delays in air traffic affect the complete door-to-door travel. Air traffic delays in Europe are analysed by EUROCONTROL, through the Central Office for Delay Analysis (CODA). CODA publishes monthly reports with the delay situation in Europe, with information about air traffic, delay length, causes, punctuality and scheduling indicators aggregated for the ECAC area. This information is reported directly from the airlines. There are also quarterly/annual reports, which contain some information at airport level, such as top 20 delay affected airports and top 20 delay affected city pairs. These reports are publicly available through CODA website.
The information provided, although very useful to get a general view of the delay situation across Europe, might be too coarse for the analysis proposed. Therefore, in order to get delay information that can be combined with the other data sources that will be used within the project (ICT-based high granularity data sources, which usually contain disaggregated information about anonymised individuals), access to more disaggregated data will be requested to EUROCONTROL. The main problem with disaggregated data is that it contains highly sensitive information for the airlines, and access to these data is restricted.

4.1.5 STATFOR data

Statistics and forecasts on air traffic in Europe are provided by EUROCONTROL STATFOR service. It provides two main sources of information:

- STATFOR Interactive Dashboard (SID): a tool that provides monthly updated statistics
- Forecasts reports: public reports for short, medium and long term forecasts on flights and service units

Access to SID is restricted to authorised users and will be requested for BigData4ATM project in case it is needed for the case studies. The information included in this document refers to STATFOR public reports. Forecast reports contain a summary of the methodology used for the forecasts, the trends and aspects that may have affected them, and the inputs and assumptions. Short term flights forecasts contain information about total number of IFR flights in Europe for both short (2 years) and medium (7 years) term at country level. Short term service units forecasts provide information for the total service units in Europe and per-country service units. These short-term flight and service units reports are produced annually, with an update (or up to two updates in the case of service units report) typically at mid-year. The temporal granularity of the information provided in these reports is yearly.

4.1.6 EUROSTAT data

EUROSTAT is the statistical office of the European Union. Its objective is to provide statistics at European level and promote the harmonisation of the statistical methods across the Members States of the EU. EUROSTAT collects and compiles data from country statistical agencies. These data go through an internal EUROSTAT validation process, which ensures the quality of the data delivered. Two main databases from EUROSTAT have been identified as potentially useful for the project: Air transport database (avia) and demographic (population and census) database.

With respect to the air transport database, the following information has been determined as useful:

- Passenger Information: number of monthly passengers per Country/Airport, disaggregated by National / IntraEU / ExtraEU passengers. There is also information about flows of passengers between the main airports.
- Flights Information: number of annual flights per Country/Airport, aircraft type, etc.

The temporal granularity of the data may range from monthly data to annual data. Although the temporal and geographical granularity of this data source might be too low for the analysis proposed in the project, it can be used to validate the results.
Regarding demographic databases, the following information has been determined as useful:

- Census information: a compilation of the information from the census of the different EU countries. It includes socioeconomic characteristics (employment, household types, etc.).
- Population on 1 January: information about the population on 1 January, with basic characterisation (age, sex, country of birth, etc.)

Census data is provided every 10 years, while population data is provided annually. With respect to the geographical granularity, data may range from NUTS (Nomenclature of Units for Territorial Statistics) 3 level to country level. Although the granularity of the data is not very high, it may be useful as a sampling frame for ICT-based data sources. For studies focused on a single country, it may be preferable to use the information from the national statistical office.

### 4.1.7 INE data

Since some of the main non-conventional datasets available for the BigData4ATM project correspond to Spain, as an example of data available from national statistics agencies we have looked at the Spanish National Statistical Office, INE. The main advantage of INE data when compared to EUROSTAT data is the higher level of detail of the demographic information provided. While EUROSTAT data has a geographical granularity of NUTS 3, the Spanish Census provided by INE has a census tract resolution, which is much higher. Data from INE will be used as a sampling frame for ICT-based data that requires an expansion process (such as the mobile phone data provided by Orange Spain).

When looking at flaws and limitations of these data, it needs to be remarked that census are produced by INE every 10 years, with the last one being produced in 2011. Also, due to the methodology applied to comply with data protection legislation, not all census tracts have the same level of detail, with some of them with full demographic information (age and gender combined for several age intervals) and some of them providing only total population, with no information about gender and age.

There is also a more updated data available from INE: municipal registers. However, they are more prone to errors, as while theoretically people are obliged to register in the place they live in, in practice they don’t always do it. Hence, information from municipal registers needs to be handled with care.

### 4.2 Conclusions: gaps and room for improvement

In this section, the main conclusions derived from the quality assessment of traditional data sources are presented. The main flaws and areas with room for improvement identified are the following:

- **Lack of door-to-door information**: one of the challenges set up in the ‘Report of the High Level Group on Aviation Research - Flight Path 2050’ is that 90% of the passengers are able to complete their door-to-door travel in less than 4 hours. However, there is not any source of information for the current door-to-door travel times. There are several sources of information for the duration of the flight segment (CODA data, DDR2), but there is no information about the door-to-kerb/kerb-to-door and kerb-to-gate/gate-to-kerb segments duration. Moreover, these are the trip segments that are more difficult to model, as they depend on several factors such as the airport layout, airport accessibility and passenger behaviour (origin, destination,
luggage check-in, risk aversion...), which causes a great variability in these segments duration. Therefore, a gap identified in traditional data is its inability to effectively measure door-to-door travel times, which is one of the first steps needed for achieving the 4h door-to-door goal.

- **Lack of information on airport access/egress mode**: this point is closely related with the lack of door-to-door information. As trips involving air transport are typically intermodal, good airport accessibility considering different transport modes is important in order to ensure a seamless travel experience for the passenger. Therefore, information on how the airport is connected to the transport network and the use that passengers make of these connections is very valuable to support the development of new public transport services/infrastructures. However, the only sources of this kind of information are passenger surveys, which, as stated before, present several limitations (high cost, they depend on the willingness to answer of the passenger and the veracity on his answers, they require planning, which does not allow unexpected events to be studies, etc.). Obtaining this information from smart passenger devices would overcome these limitations, helping to better integrate air transport in the European transport network.

- **Poor passenger intra-airport behavioural information**: similarly to information on airport access/egress mode, intra-airport behavioural information has traditionally been obtained from surveys. Through these surveys, information related to the buffer time (the time passenger arrives in advance to their flight departure), the shops visited and the expenses made before departing is obtained. As said before, surveys present a set of intrinsic limitations, and therefore this intra-airport information is rarely collected with the desired detail and frequency. If higher quality information could be obtained, more effective marketing decisions could be made, targeting specific types of passengers with personalised offers. Also, this kind of information could be used to detect bottlenecks in airport processes and optimise the allocation of resources within the airport.

- **Data sources of restricted access**: STATFOR Interactive Dashboard (SID) and CODA disaggregated data could be combined with passenger ICT based data to measure the effects of air traffic congestion and delay in passenger door-to-door travel times. However, SID and CODA data contain highly sensitive information for the airlines, and therefore access to these sources is restricted. It would be interesting to explore the potential use of these restricted sources, if access is granted for the project. Another interesting analysis would be to explore how Big Data sources can be used to replicate some of the information obtained from these restricted data sources.

- **Poor granularity**: a common characteristic of many of the traditional datasets analysed (particularly public reports) is that they are highly aggregated, either temporally (typically monthly), geographically (route/airport/country level) or both. Although this level of granularity may be enough for many studies aimed to get an overall picture of the status of the ATM system, it does not allow more detailed studies. As an example, the delay data aggregated at route level provides average delays quarterly, and with that it is not possible to assess how delay affects the complete door-to-door travel times. Another example is the study of disruptions: with monthly aggregated air transport intelligence data, it is not feasible to study how events (for example, a football match) affect the number of passengers on a particular day.
5 Quality assessment results for Big Data sources

5.1 Quality assessment results

In this section, the main results of the assessment for each one of the Big Data sources that we plan to use in the project are presented. Each one of the data sources is described, and the strengths and limitations identified for each one of these data sources are also identified. The complete results of the quality assessment can be found in Appendix A (data factsheets) and Appendix B (Big Data sources detailed analysis).

5.1.1 Mobile phone records

Nommon has access to disaggregated data from Orange Spain through a private agreement between the parties. These data consist of call detail records (CDRs), socio-demographic information about Orange clients and the geographic location of the telecommunication towers. As explained in section 2.2.2.1, analysing the data produced by the Mobile Network Operator during several days makes it possible to extract information about mobility patterns of the population.

Orange Spain also provides roamers CDRs and Mobile Country Code (MCC), which allows tourists’ mobility patterns to be analysed. In the case of roamers, no socio-demographic information is available.

The characterisation of mobile phone records has been performed with the data collected on Wednesday 2nd March 2016. That day, 8.9 million users (almost 20% of Spanish population) connected to the network, generating 629 million records. If more days were analysed, the sample size would exceed the 20% of the population, as Orange's market share is close to 30%. The graphs below show the percentage of users with some phone activity and percentage of records (Figure 2) and the distribution of records per user in each hourly interval (2). There it can be seen that, during the central hours of the day, most of the users active during the day (>60%) have at least a register. During night, the number of active users drops to 30%, which is still a high number. On average, during daytime users have around 7 registers per hour, with this number dropping to 2-3 registers per hour during night.
One of the main advantages of mobile phone data is that it is collected passively, without user interaction. When access to Internet connections/data sessions is available, the information flows and the frequency of temporal sampling increases substantially because data sessions are periodically renewed and some apps may send or request information automatically (for example, looking for updates, or checking e-mail). As nowadays is very common to have Internet connections available, mobile phone users generate geolocated data even if they do not use their phone. In Figure 4, it is represented the probability density function of time intervals between consecutive data sessions. There, these automatic updates can be seen as the spikes that appear at 30 minutes, 60 minutes, etc.
The spatial accuracy of the data depends on the spacing between towers providing coverage in the area. In densely populated areas, distances are of hundreds of metres, while in rural areas distances are of kilometres. An analysis was conducted in the area of Madrid, where the radius of coverage of more than 50% of the towers is less than 300 meters. This spatial accuracy is acceptable for characterising medium and long distance trips; however, short range trips may be undetected. Also, in urban environments, this spatial accuracy does not always allow route and mode of transport to be determined and therefore mobile phone records need to be combined with other data sources.

As this analysis proves, mobile phone data is a powerful source to obtain information about mobility patterns of the population, since the sample size is much bigger than compared to other data sources (around 20% of the population), and both spatial and temporal resolution are good enough to characterise the mobility of users. It also allows the study of the mobility of tourists in Spain, but this information is only available for those tourists that connect to Orange Spain network and, as it is shown in Appendix B, roamers have a slightly lower amount of registers.
Mobile phone data, however, has some inherent limitations. First, it has already been said that the spatial accuracy of mobile phone data may not be enough for some applications, and it needs to be combined with other fine-grained data (like in the case of urban route determination). Also, sociodemographic information provided by the MNO sometimes contain errors (for example, missing values or teenagers that have their mobile phone on their parents’ name). Besides, when looking at the age and gender distribution for the sample, it can be seen that young people are almost inexistent (either because they do not have a mobile phone or because they have it on their parents’ name) and the segment of population with ages between 25 and 60 is overrepresented. This needs to be taken into account when expanding this sample to the total population.

Another limitation is that mobile phone data is geographically limited to the country(ies) where the MNO has presence. At the moment of writing this deliverable, only Orange Spain data was available, which limits the analysis possible that we can conduct to the Spanish territory. However, the CDRs produced by Orange clients when they are abroad are stored in the dataset, so it will be explored what valuable information can be extracted from those registers.

### 5.1.2 Twitter data

IFISC has been collecting Twitter data from the public API Stream since 2014. The process is still ongoing and covers all European countries, and eventually can have a worldwide coverage for further needs. Here we focus on geolocated tweets, i.e., only those tweets associated to geographical coordinates (latitude and longitude). This geolocation needs to be enabled by the user. Each document in the database corresponds to an anonymised tweet, and the typical (minimum) information of a record consists of the following fields:

- User identification code
- Tweet identification code
- Date and time of the creation of the tweet
- Coordinates associated to the tweet
- Text contained in the tweet
Regarding sample characteristics, the Twitter user pool is biased towards young people, typically around 20s-early30s. Moreover, there are differences amongst Twitter users in terms of geolocated tweets: some users tend to share their location more than others. This seems to be related with cultural aspects, as users from certain countries (Spain, UK) are more prone to publish this kind of information. In terms of sample size, it depends on the duration of the period considered to study the data. For a single day, the sample size is about 100,000 users across Europe, with around 225,000 tweets. However, if the time frame considered is extended to a whole year, the number of users grows to almost 3,750,000 users, with more than 76 million tweets. Therefore, for a single day, an active user will produce an average of 2.2 tweets, while, for the whole year, this average grows up to 20 tweets per user.

Regarding the temporal characteristics of the dataset, in Figure 10 and Figure 11, a similar pattern to the ones found with mobile phone data is present: more tweets and active users during central hours of the day. However, the reduction in tweets and users during the night is more marked. This is because Twitter actually needs user interaction with the mobile (it requires the user to send a tweet), while this is not the case for mobile phone registers when data connections are enabled.
Twitter data are typically biased by the characteristics of the users: as explained in section 2.2.2.2, several studies that characterise this have been presented. Despite those issues, it has been shown that they are reliable in terms of for certain mobility analyses. Twitter data are highly correlated with official statistics when used to get information about Origin/Destination flows in urban areas Unlike mobile phone data, which are restricted to the country of the MNO, Twitter data cover all European countries, and therefore it can be effectively used to study aggregated flows of passengers between different European cities.

Moreover, Twitter not only provides location information, but it contains text information, which can be semantically analysed to study passengers’ opinion about ATM performance during disruptions, or how effectively announcements from ATM stakeholders propagate through the Twitter network.

## 5.1.3 Public transport smart card data

Access to the Madrid public transport smart card data was achieved during this first stage of the project. The dataset provided comprehends three working days of March 2016, and contains the validations made by smart card owners when they accessed any kind of public transport in the region of Madrid: commuter trains, metro, intercity buses and municipal (EMT) buses. Due to the functioning of Madrid public transport, only records of entering the system are available, as only for a few subway and train stations users have to use their tickets/smart cards to get out. However, airport metro and train stations do need this ‘tap out’ validation. Airport accessibility indicators and access modes used by passengers could be derived by combining these data with the information about the location of the different public transport stops.

At the moment of writing this deliverable, only information about the location of the intercity buses stops was available, with no information about the location of the rest of public transport modes stops. Therefore, the complete analysis of the quality of the dataset could not be performed, and only a preliminary characterisation of the dataset was done, leaving the detailed characterisation for when the rest of the dataset arrives.

Regarding sample characteristics, for each one of the three days of the dataset, around 1.1 million different user IDs appear. This is the number of active smart card owners detected in those days. The average number of validations per user was around 3.3.

In this preliminary analysis, it was detected that the most used public transport mode were subway and city buses, with subway being the most used during peak hours and city buses the most used during valley hours (see Figure 12). It can also be seen that intercity buses correspond to a small fraction of the total number of validations.
From the results of this preliminary assessment it can be derived that these data can be used to assess the modal split of the access/egress from the airport. However, two main issues need to be solved:

- The location of the stops available up to now is limited to intercity buses, but it is expected to have the complete stops dataset soon.
- Madrid public transport users do not have to use their smart card when egressing public transport: this will need to be solved through different models. However, when studying the trips to the airport, this is mitigated by the fact that subway and train stations at the airport require a validation of the ticket at the exit.

Also, during the week before closing this deliverable, access to similar information for the public transport system of Malaga was achieved. As soon as these data arrives, an analysis will be made and this deliverable will be updated with those data.

### 5.1.4 FlightRadar24 data

IFISC has been collecting data from the Flightradar24 website since early 2015. Data can be collected freely from the website, but explicit permission (which has been granted to IFISC) is required in order to use it. Each document in the database corresponds to a flight, where the typical document contains the following fields:
<table>
<thead>
<tr>
<th>Field Name</th>
<th>Type and Format</th>
<th>Note(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>regnumber</td>
<td>string</td>
<td>Tail number of the aircraft operating the flight</td>
</tr>
<tr>
<td>flight</td>
<td>string or list of strings</td>
<td>Flight number (may be more than one)</td>
</tr>
<tr>
<td>airline</td>
<td>string (2 characters)</td>
<td>Operating airline (IATA code)</td>
</tr>
<tr>
<td>iata_to</td>
<td>string (3 characters)</td>
<td>Origin airport (IATA code)</td>
</tr>
<tr>
<td>iata_from</td>
<td>string (3 characters)</td>
<td>Destination airport (IATA code)</td>
</tr>
<tr>
<td>scheduled_departure</td>
<td>UNIX timestamp</td>
<td>Scheduled Off-Block Time (SOBT)</td>
</tr>
<tr>
<td>scheduled_arrival</td>
<td>UNIX timestamp</td>
<td>Scheduled In-Block Time (SIBT)</td>
</tr>
<tr>
<td>real_departure</td>
<td>UNIX timestamp</td>
<td>Actual Take-Off Time (ATOT)</td>
</tr>
<tr>
<td>real_arrival</td>
<td>UNIX timestamp</td>
<td>Actual Landing Time (ALDT)</td>
</tr>
<tr>
<td>collected_at</td>
<td>string</td>
<td>Date and time, formatted as yyyy-mm-dd HH:MM:SS</td>
</tr>
</tbody>
</table>

Subtracting the scheduled departure from the real departure yields the flight's departure delay plus the taxi-out time. Likewise, subtracting scheduled arrival from real arrival gives arrival delay plus taxi-in time. Delays and taxi times cannot therefore be separated, unless one of the two is known from a different dataset. This should not be a problem when considering the delay from the point of view of the passengers, as the only thing that matters is the time they get off the aircraft.

Documents which do not possess a valid value for each of the above-mentioned fields are discarded and not considered in the analysis. The number of "malformed" documents found in each day is usually zero or negligible in the US and Europe, although outside these regions it is a serious issue.

Consistency issues between related flights may also appear:

1. Two aircraft at the same airport (both arriving/departing, or possibly one arriving and one departing) are mistakenly identified with the same tail number. This can result in a variety of inconsistencies, such as flights terminating after the previous leg in the rotation has started.

2. An aircraft moves from a zone with good coverage to a zone without coverage and re-emerges later. In this case, the rotation will have one or more "holes" in it.

3. The aircraft's schedule may itself be malformed, with one leg scheduled to depart before the previous one is completed. According to what we have discussed in private communications with industry experts, this is something that actually takes place.

These problems can be solved with different strategies depending on what is needed. If mapping each flight to a specific aircraft is not needed (for example when calculating the total traffic at one airport during a period of time), no action needs to be taken. The simplest measure is to just "split" rotations at every error, i.e. all the flights after a "hole" can be considered as operated by a different aircraft. Another option is to discard the entire rotation as soon as an error is found, but it is likely excessive, as it would result in around 10% of the flights being discarded.
Regarding data coverage, the Figure 13 shows the amount of valid flight entries in the ECAC area for each day since data collection has started until the 25th of September 2016. Gaps and periods with abnormally low numbers of entries are due to technical problems either on the IFISC or the Flightradar24 sides. The ticks are placed at the beginning of the corresponding month. For comparison, the average number of flights per day in Europe (excluding non-EU countries) during 2015 is 20776 according to EUROSTAT, and according to EUROCONTROL, the number of flights/day in the ECAC area ranges from around 22000 during January 2015 to above 30000 during the summer (detailed day-by-day data is not publicly available). According to the figure, it seems that when there are not technical problems, the number of flights recorded through FlightRadar24 is very similar to the official data.

![Figure 13: Number of daily flights recorded by IFISC](image)

**5.1.5 Anonymised credit card transactions**

Access to this kind of data for the BigData4ATM project has not yet been granted. However, as it is data of interest for the project and negotiations to get access to it are currently ongoing, it has been included in this document. IFISC and Nomnom have previous experience with this kind of data in the FP7 project EUNOIA (http://eunoia-project.eu/doc/about/). The analysis presented here is a summary of the analysis made for the EUNOIA project. There, an anonymised credit card transactions dataset was provided by BBVA, one of the largest Spanish banks. The geographical scope of this dataset was restricted to the provinces of Madrid and Barcelona, and covered two years, 2011 and 2012. The dataset contained the following information:

- Transactions made by BBVA clients at any Point of Sale (POS)
- Transactions made by non-BBVA clients at any BBVA POS

A typical transaction record contains information about the amount spent, timestamp of the transaction, user anonymised identifier, POS location and business type. The information regarding BBVA clients was a little bit richer, containing basic profile information about the client (gender, age, home zip code), while for non-BBVA clients the only information available was the country of their bank. This dataset is not available for BigData4ATM, as the data provision agreement expired with the end of EUNOIA. At the moment of writing this deliverable, an extension of the agreement, which will also grant access to more recent data, is being negotiated. An in-depth analysis of this dataset can be found in [41].
Regarding data quality and potential usefulness for the BigData4ATM project, the number of users with a credit card activity frequent enough to derive mobility patterns is very low (more than 95% of the individuals have less than 1 transaction per week on average). However, from the BigData4ATM perspective, the interest of this data source is not to derive mobility patterns, but to characterise passenger expenditure and how it is impacted by air traffic disruptions.

5.1.6 Google Maps APIs

Google Maps offers a variety of map/location-based data APIs. A preliminary analysis on these APIs has been conducted, and those that may be the most useful for BigData4ATM have been identified. These are the following APIs:

**Google Maps Directions API.** This API calculates route options and their associated travel times given an origin and a destination. These routes can be obtained for different transport modes, such as walking, private transport or public transport. Also, this API accepts waypoints or departure/arrival time information to better characterise the trip. This API is very useful to determine the route followed during trips detected by other source (such as mobile phone data or Twitter data). Also, this API can be used to calculate indicators of airport accessibility and connections with other transport modes.

**Google Maps Distance API.** This API calculates distance and travel time between locations, with mode and departure/arrival time options. The main difference between this API and Directions API is that this API does not return the route information. However, it can calculate several origin-destination pairs during the same call to the web service.

**Google Places API.** This API provides information about places present in an area. These places can be stores, services, localities, points of interests, etc. This information could be useful for adding a better characterisation of the passenger behaviour by inferring the activities performed by people during their visits to a place.

**Google Maps Geocoding API.** This API translates human-readable addresses into geographic coordinates and vice versa. This service would be useful for the analysis of credit card data, where businesses coordinates are not always correct.

**Google Maps Roads API.** This API generates the route followed by a user from GPS traces. It may be interesting if the project gets access to GPS data (like data from a mobile App).

**Google Maps Geolocation API.** This API provides a location based on cell phone towers and Wi-Fi nodes that the mobile phone can detect. This information can be useful to locate mobile phone users when they are abroad, as mobile network operators only know the location of their own towers.

From these APIs, Google Maps Directions API was identified as the most interesting for the project, and therefore, a detailed analysis of this API has been carried out. The main benefits of this API are the following:

- **Transport network:** Google Maps Directions API uses Google road network, which has global coverage and is constantly being updated. Also, Google Maps includes public transport data from a large set of cities, (list available in [https://maps.google.com/landing/transit/cities/](https://maps.google.com/landing/transit/cities/)). Having this kind of information already pre-calculated with a high level of accuracy is crucial for a project like BigData4ATM, where there would be no time to create a transport network, gather public transport timetables and merge all these into an operational network model.
• **Traffic model**: Google Maps Directions API has an embedded traffic model that takes into account both historical and real-time traffic data when calculating routes. Therefore, the travel time estimations are expected to be quite accurate. During the detailed characterisation, it was determined that the real-time traffic data had influence only on estimates for the next 2 hours from the time when the estimation was requested. As before, building a traffic model is out of the scope of the project, and therefore access to this information is a valuable addition.

These two benefits are important because they provide information about the door-to-kerb and kerb-to-door segments, which are by far the most difficult segments of a trip to analyse and characterise, due to the multiple possibilities of origin/destination, mode to get to/from the airport and route followed.

Some limitations were also identified for this data source, the most important being the following:

• **Number of daily queries**: free access to Google Maps Directions API grants a maximum of 2,500 queries per day, which is a number way below the needs of the project. This limitation can be overcome by using a Premium Plan, which allows up to 100,000 queries per day.

• **Sponsored flights**: when asking this API for route information between a pair of points that are connected by flights, Google may suggest some air routes. However, these routes are not necessarily the best alternatives, as they are sponsored routes and therefore, for determining flights, the information from Google is not reliable. But, as there are many ATM-related sources (Sabre, FlightRadar24) that provide flight data, this is not a big issue.

• **No access to historical data**: Google Maps Directions API provides travel time estimations, not historical travel times. This can be a limitation when analysing disruptions, as the real times to/from airport for a specific date and time are not available.

An analysis of the driving travel times from Nuevos Ministerios (a central zone in Madrid) to the airport was carried out. According to the next figures, the traffic model embedded in Google Maps Directions API provides different travel times for different times of the day and times of the week. However, it seems that it does not take into account additional information (e.g., the travel times for a Monday in August are the same as the travel times for a Monday in November). This is somehow overcome by Google by using real-time traffic data to adjust the estimations, but this real-time data only influences estimations within 2 hours from the query.
Although not perfect for our purposes (in particular, due to the lack of historical data), the Google Maps Directions API has been identified as a very valuable data source for BigData4ATM, which provides unique capabilities (such as global public transport routes and timetables information) and an embedded transport network and traffic model.
Airport accessibility

Airport accessibility is one of the key factors that may influence passenger decisions about choosing their flight when they need to travel. Airport accessibility can therefore be a tool for assessing competition between airports as well as with other transport modes. There exists a broad variety of accessibility indicators, ranging from very simple ones (travel time, number of trips to a zone) to complex, compound indicators. As an example of application for Google Maps Directions API, a map of the driving travel times to airport has been obtained for the city of Madrid for three times of the day: morning peak, valley hour and afternoon peak. Using the public transport data from Google Maps APIs, similar maps could be obtained for other transport modes, such as bus, subway or train.

![Figure 16: Driving travel times to airport at morning peak](image1.png)  ![Figure 17: Driving travel times to airport at valley hour](image2.png)

Route/mode determination

Google Maps Directions API data can also be used to enrich the analysis made with other data sources. In the case of mobile phone data, the mobility of a great sample of the population is characterised, but this characterisation lacks some important details, such as the route followed or the transport mode used. By combining mobile phone data with Google data, it is sometimes possible to overcome these gaps.

As an example of application, a preliminary characterisation of the modal share in the trips between Madrid and Catalonia has been made with the combination of mobile phone and Google Maps data. First, with Google Maps Directions API, possible routes between Madrid and Catalonia for high-speed train (blue) and road (red) are calculated. Plane routes (green) for that day were also obtained from ATM databases. Then, mobile phone registers are compared against these routes, and mode can be determined (it is to be noted that, even for people travelling by plane, sometimes mobile phone registers appear during landing and take-off). Similarly, this methodology will be also explored to determine the transport mode used to get to the airport.
5.1.7 TomTom data

TomTom, similarly to Google Maps APIs, has a service that provides routing and traffic information based on TomTom’s database, which is a collection of anonymous data from vehicles with GPS devices. The services provided by TomTom are the following:

**Online Traffic Flow.** This service returns information about the current speed and free flow speed for a road segment or the roads present in an area. This is useful for routing options, in order to suggest routes that avoid congestion, but it is not of particular interest for BigData4ATM.

**Online Routing.** This service calculates different possible routes and travel times between an origin and a destination location, only for private vehicle. Similarly to Google, TomTom uses historical data to obtain traffic patterns and predict journey times.

**Traffic Stats.** This service provides historical travel times and speeds on any road (similar to online traffic flows, but showing historical data) or route (similar to online routing, but for specific days in the past).

As done with Google Maps Directions API, an analysis of the driving travel times between Nuevos Ministerios and Madrid airport was carried out, looking for the variations of the travel times with hour of the day and day of the week. The results of this analysis are present in the next figures:
TomTom data provide similar results to Google data, but with some differences:

- TomTom travel times during the night are constant, while those of Google vary. It seems that, when there is little congestion, TomTom uses some free flow estimations.
- The mid-day peak that appears during working days is much bigger in Google data than in TomTom data. By looking at the public transport smart card data hourly validations, this peak is quite considerable, which would suggest that Google estimation could be more accurate.

The influence of real time traffic could not be studied in such a detailed way as in the Google case due to the limitations of the TomTom trial account, which allowed just a few queries. However, the following information about how TomTom embeds real time traffic information into travel time estimations was found in support webpages: when calculating a route, TomTom estimates the travelling time based on its database and the current traffic, which may add some seconds to the duration of the travel (traffic delay). For future calculations, the traffic delay equals to zero while in routes calculated for the present it may be positive depending on traffic conditions. According to these considerations, if the estimated congestion in the present is not greater than the average congestion for the same weekday, the planned travel time for the present will be the same as the planned time in a future day for the same weekday.

As it can be seen, TomTom and Google are very similar data sources, which somehow complement themselves: Google has public transport data, although it does not provide access to historical data, while TomTom provides access to historical traffic data while not taking into account public transport.

![Figure 21: Variation with day of the week of the driving travel times from Nuevos Ministerios to Madrid airport](image)
6 Passenger-oriented ‘Big Data’ sources map

- **Door-to-Door**
  - Mobile phone data
  - Twitter data

- **Gate-to-Gate**
  - FlightRadar24 data

- **Kerb-to-Door**
  - Google data
  - Public transport Smart Card data
  - Mobile phone data

- **Intra-airport**
  - Beacon data
  - Boarding pass data
  - Credit card data
7 Research questions

In this section a set of potential research questions to be tackled during WP3 are outlined. These research questions were initially defined by the Consortium partners, and were later on refined with the inputs gathered from the External Experts Advisory Board during the 1st BigData4ATM Stakeholder Workshop that took place in Amsterdam the 14th of November 2016. These research questions can be grouped in four categories:

- Door-to-door mobility analysis
- Intra-airport mobility analysis
- Expenditure analysis
- Opinion and sentiment analysis

7.1 Door-to-door mobility analysis

The data sources that will be used for the research questions associated to this topic are:

- Mobile phone records
- Twitter
- Public transport smart card
- Google Maps APIs

The research questions associated to door-to-door mobility analysis are related to:

Airport catchment area:

- Determine from where people go to the airport and the transport mode used by analysing mobile phone data and merging it with public transport data.
- Study new, more comprehensive airport accessibility indicators by merging supply information (public transport and driving travel times) and demand information (passengers that used each transport mode to access the airport).
- Analyse the competition/complementarity with other transport modes: what would be the influence of a new public transport connection to the airport? Is high-speed train always competing with air transport or can it act like a feeder for long-range flights?
- Study how this airport catchment area and accessibility indicators could be used for air traffic forecasting.
Door-to-door passenger behaviour:

- Determine the door-to-door travel time, either by directly measuring it (with sources like mobile phone data), by modelling it (from accessibility indicators), or through a combination thereof.
- Infer different passenger behavioural patterns (business/leisure, accommodation, length of the stay, visited places etc.) by analysing several days of data.
- Analyse how congestion impacts passenger behaviour: do they shift access mode? Do they arrive earlier to the airport?
- Propose new metrics that capture how delays in ATM affect total door-to-door travel time, and study how this information could be used to prioritise flights so as to minimise impact on passengers.

### 7.2 Intra-airport mobility analysis

The data sources that could be used for the research questions associated to this topic are:

- Wi-Fi/Bluetooth beacons
- Boarding pass

The main focus of this analysis would be to obtain the kerb-to-gate mobility patterns, and try to extract meaningful information from them related to airport processes, with the aim to detect and predict flows of people in different areas of the airport, identify bottlenecks and why they are produced, and ultimately optimise the use of airport resources for a better integration of landside and airside processes. As access to the abovementioned data sources are still under negotiation, these research questions are still preliminary and will be jointly refined with the data provider if access to such data is finally obtained.

### 7.3 Expenditure analysis

The main data source that will be used for this topic is credit card transactions data. The research questions associated to expenditure analysis will be related both to expenditure inside the airport and outside the airport (e.g., tourist expenditure). In particular, we will explore whether the analysis of passenger expenditure patterns can provide new insights on the impact of ATM disruptions: how do expenditure patterns change when disruptions occur? How much does it cost for the tourist destinations? Is expenditure in airports significantly increased due to a longer stay of the passengers?

### 7.4 Opinion and sentiment analysis

The main data source that will be used for this topic is Twitter. Here again, we will initially focus on the impact of ATM disruptions: how does the general public perceive the responsibility of different agents (ANSPs, airports, airlines)? What is the temporal dynamics of the tweets related to the disruption (i.e. since when and for how long is the disruption an active topic)? How does information from official channels spread? How does this influence attitude towards ATM.
8 References


Appendix A  Data characteristics factsheets

The factsheets of each data source analysed by the BigData4ATM project can be found in the next pages in alphabetical order. The factsheets included in this document correspond to the following data sources:

A.1 AIRFARE DATA ........................................................................................................ 50
A.2 CREDIT CARD DATA ................................................................................................ 51
A.3 CODA PUBLIC REPORTS .......................................................................................... 52
A.4 EUROSTAT AIR TRANSPORT DATABASES ............................................................... 53
A.5 EUROSTAT DEMOGRAPHICAL DATABASES ............................................................. 54
A.6 FLIGHTRADAR24 ...................................................................................................... 55
A.7 GOOGLE MAPS DIRECTIONS API DATA ................................................................. 56
A.8 IATA PAXIS DATA ................................................................................................. 57
A.9 INE POPULATION DATA .......................................................................................... 58
A.10 OAG DATA ............................................................................................................. 59
A.11 ORANGE SPAIN MOBILE PHONE RECORDS ......................................................... 60
A.12 PUBLIC TRANSPORT SMART CARD DATA ......................................................... 61
A.13 SABRE DATA ......................................................................................................... 62
A.14 STATFOR PUBLIC REPORTS .................................................................................. 63
A.15 TOMTOM DATA ...................................................................................................... 64
A.16 TWITTER DATA .................................................................................................... 65
A.1 Airfare data

1. General Information - Identification of the data source

   - Data source name: Airfare data (MIDT)
   - Last update of this file: 15/07/2015
   - Contact information: Private
   - Support: Not applicable

2. Abstract - Brief description of the data source and its potential usefulness for the project

   It may be of interest to the project to detect whether airfares are connected to delay/cancellation i.e. airfares are lower/higher as a function of congestion. Market intelligence data (MIDT) provides information on one-way airfares for tickets bought through a Global Distribution System (GDS). Worldwide there are 3 large GDSs and Amadeus probably has the largest market share in Europe. It may be easier to purchase data from an MIDT-reseller who is able to clean the data and correct the sample. Such resellers include Interline, Amadeus (belonging to AITA) and OAG. Purchasing historical data normally reduces the price (approximately $20,000 for a year of data).

3. Availability - Relevant information about owner and readers to use

   - Owner: Private (several sellers of such data)
   - Data access: Private agreement
   - Data cost: Depending on the seller, temporal scope and geographical scope. Price within the $25,000 range for a year of data.
   - Access limitations: Will probably be a one-off purchase
   - Availability within the project: All partners
   - Privacy / Confidentiality issues: Data is not at a personal level, so it will not be affected by privacy/confidentiality issues
   - Security issues: Not applicable
   - State: To explain
   - Link to the data: Not applicable

4. Data Characteristics - Temporal, geographical and size characteristics of the information provided

   - Estimated size of the sample: Will be a function of the market size of the data producer, generally monthly aggregated. Daily information may be available, but at a higher cost.
   - Temporal scope: Expendable on agreement
   - Geographical scope: Single destination date with information on any connecting hubs
   - Temporal granularity: Expendable on agreement
   - Geographical granularity: Depending on agreement
   - Delivery frequency: One-off purchase
   - Delivery delay: Several months
   - Data format: Not applicable

5. Quality Issues - Comments about known flaws of the data source

   The data can be very problematic (much missing data and flights at 0 price due to frequent flyer programs and employee discounts) which is why it might be preferable to purchase from a reseller who has information to clean the data.

   We would need to ensure that the airfare includes all additional charges and not simply the base price which can be very misleading.

6. Comments - Other relevant information

   Project partners have experience with airfare data provided by Interline and OAG.
A.2 Credit card records

### 1. General Information - Identification of the data source

- **Data source name:** Credit Card Data
- **Last update of this file:** 2016
- **Contact information:** 
  - Person: Alessio Mancuso (MAC
  - Email: a.mancuso@post@liverpool.ac.uk)

### 2. Abstract - Brief description of the data source and its usefulness for the project

Credit card transactions record information about credit card users every time they make a purchase.

The key features of this data source are:
- **Anonymity:** The data is anonymized, ensuring user privacy.
- **Volume:** The dataset contains a significant number of transactions, covering a wide geographic area.

### 3. Availability - Relevant information about owner and readiness to use

- **Access conditions:** Through a commercial agreement.
- **Data access:** Through an API or batch.
- **Data cost:** Free.
- **Access limitations:** To be determined.
- **Availability within the project:** To be determined.
- **Privacy / Confidentiality issues:** The data is encrypted and stored in a secure environment.
- **Security issues:** No known security issues.
- **State of the data:** Data is currently being processed.

### 4. Data Characteristics - Temporal, geographical and size characteristics of the information provided

- **Temporal scope:** The data is updated quarterly.
- **Geographical scope:** The dataset covers the entire European Union.
- **Geographical granularity:** Data is available at the city level.
- **Delivery frequency:** The data is delivered on a monthly basis.
- **Data format:** The data is available in CSV format.

### 5. Quality issues - Comments about known flaws of the data source

- **Data accuracy:** The data is generally accurate, but it may contain occasional errors.
- **Data consistency:** The data is consistent across different datasets.

### 6. Comments - Other relevant information

- **Data mining:** The data is suitable for various data mining tasks.
- **Data visualization:** The data can be effectively visualized to present insights.

This dataset is expected to be updated quarterly, providing researchers with the latest information on consumer behavior.
A.3 CODA public reports

<table>
<thead>
<tr>
<th>BigData4ATM - WP2 Data identification, collection &amp; assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. General information - identification of the data source</td>
</tr>
<tr>
<td>Data source name: CODA - control office for delay analysis belonging to Eurocontrol</td>
</tr>
<tr>
<td>Last update of this file: 29/01/2020</td>
</tr>
<tr>
<td>Contact information: Not applicable</td>
</tr>
<tr>
<td>Support: Not applicable</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Abstract - Brief description of the data source and its potential usefulness for the project</th>
</tr>
</thead>
<tbody>
<tr>
<td>In order to study how delays propagate in the complete door-to-door travel, flight delay information is needed. CODA publishes monthly reports about traffic delays in Europe. The delay information contained in these reports is gathered from the airlines. Delays are classified according to their cause, including reactionary delays, ATM caused, airline caused amongst others.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Availability - Relevant information about owner and readiness to use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Owner: Eurocontrol</td>
</tr>
<tr>
<td>Access conditions: Public</td>
</tr>
<tr>
<td>Data access: Online</td>
</tr>
<tr>
<td>Data cost: Free</td>
</tr>
<tr>
<td>Access limitations: None</td>
</tr>
<tr>
<td>Availability within the project: Available for all members</td>
</tr>
<tr>
<td>Privacy / Confidentiality issues: Public monthly reports do not comprehend any of these issues.</td>
</tr>
<tr>
<td>Security issues: None</td>
</tr>
<tr>
<td>Status: Already available</td>
</tr>
<tr>
<td>Link to the data: <a href="https://www.eurocontrol.int/articles/data-publications">https://www.eurocontrol.int/articles/data-publications</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Characteristics - Temporal, geographical and size characteristics of the information provided</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated size of the sample: Approximately 60% of IFR flights in the ECAC area</td>
</tr>
<tr>
<td>Temporal scope: From 3AM onwards</td>
</tr>
<tr>
<td>Geographical scope: ECAC area</td>
</tr>
<tr>
<td>Temporal granularity: Monthly aggregated</td>
</tr>
<tr>
<td>Geographical granularity: ECAC area aggregated</td>
</tr>
<tr>
<td>Delivery frequency: Monthly</td>
</tr>
<tr>
<td>Delivery delay: Any delay</td>
</tr>
<tr>
<td>Data format: .csv</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality issues - Comments about known flaws of the data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>CODA data, as it is obtained mainly from airlines reports, is subject to their criteria for delay classification and calculation. Also, CODA data does not include all the flights in the ECAC area. The granularity of the information provided in the reports may be too low, and therefore limit the utility of this data source within the project.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comments - Other relevant information</th>
</tr>
</thead>
<tbody>
<tr>
<td>There are also quarterly annual reports, which contain some information disaggregated at airport level (top arrivals/departure delays, top affected ports).</td>
</tr>
</tbody>
</table>
A.4 EUROSTAT air transport databases

<table>
<thead>
<tr>
<th>Table 2.1 - Data Identification, Collection &amp; Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data source name</strong></td>
</tr>
<tr>
<td><strong>Last update of this file</strong></td>
</tr>
<tr>
<td><strong>Contact information</strong></td>
</tr>
<tr>
<td><strong>Support</strong></td>
</tr>
</tbody>
</table>

2. **Abstract - brief description of the data source and its potential usefulness for the project**

EUROSTAT is the statistical office of the European Union. Amongst others, it provides air transport specific information at EU level, which will be useful to validate the results of the project.

- Passenger information: number of monthly passengers per Country/Airport. Disaggregated by National / interEU / extraEU passengers. Flows of passengers between the main airports.
- Flight information: number of arrivals/ departures per Country/Airport, aircraft type, etc.

3. **Availability - relevant information about owner and readiness to use**

| Owner | Public (EU)  |
| Access conditions | Open Sources  |
| Data access | Free  |
| Data cost | Free  |
| Access limitations | No limitations identified  |
| Availability within the project | All BigData4ATM project partners  |
| Privacy / Confidentiality issues | No confidentiality issues identified  |
| Security issues | No security issues identified  |
| State | Already Available  |
| Link to the data | [https://ec.europa.eu/eurostat/web/transport/data/database](https://ec.europa.eu/eurostat/web/transport/data/database)  |

4. **Data Characteristics - Temporal, geographical and use characteristics of the information provided**

- Estimated size of the sample: Data represents data values collected from national authorities. Information about sample size is not available.
- Temporal scope: Depending on the data queried, but typically from 2000 onwards.
- Geographical scope: EU level.
- Temporal granularity: Depends on the date queried. It may range from monthly data to annual data.
- Geographical granularity: Depends on the date queried. It may range from airport level, airport level + country level, country level.
- Delivery Frequency: Depends on the delivery of the data, which may be every 6 months to deliver the data. Typically, annual data is released around 9 months after the reference period.
- Delivery Delay: 15Y
- Data format: 13Y

5. **Issues - Comments about known flaws of the data source**

Data temporal granularity might be too coarse to perform the analysis proposed within the project.

Although there is information of flows between airports, it is only available for main airports.

Timeliness might be an issue when studying recent events, as the 0-6 months delay might be incompatible with project deadlines.

6. **Comments - Other relevant information**

Data is collected and compiled by EUROSTAT from country agencies. Data goes through an internal validation process, which should ensure the quality of the data.
### A.5 EUROSTAT demographical databases

#### 2. General information - Identification of the data source

<table>
<thead>
<tr>
<th>Data source name</th>
<th>EUROSTAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last update of this title</td>
<td>29/07/2012</td>
</tr>
<tr>
<td>Contact information</td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

#### 3. Availability - Relevant information about owner and responsibilities to use

<table>
<thead>
<tr>
<th>Owner</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access condition</td>
<td>Open source</td>
</tr>
<tr>
<td>Data access</td>
<td>Batch</td>
</tr>
<tr>
<td>Data cost</td>
<td>Free</td>
</tr>
<tr>
<td>Access limitations</td>
<td>No limitations identified</td>
</tr>
<tr>
<td>Availability within the project</td>
<td>All BigData4ATM project partners</td>
</tr>
<tr>
<td>Privacy / Confidentiality issues</td>
<td>No privacy/confidentiality issues identified</td>
</tr>
<tr>
<td>Security issues</td>
<td>No security issues identified</td>
</tr>
<tr>
<td>Status</td>
<td>Already available</td>
</tr>
<tr>
<td>Estimated size of the sample</td>
<td>Data represents total values collected from national authorities. Information about sample size is not available.</td>
</tr>
<tr>
<td>Temporal scope</td>
<td>Data is collected and updated by EUROSTAT from national statistical offices. Data goes through an internal validation process, which should ensure the quality of the data.</td>
</tr>
<tr>
<td>Temporal granularity</td>
<td>EU</td>
</tr>
<tr>
<td>Geographical granularity</td>
<td>Country data, every 10 years. Population by 31st December, annually.</td>
</tr>
<tr>
<td>Delivery frequency</td>
<td>Country data, every 10 years. Population by 31st December, annually.</td>
</tr>
<tr>
<td>Data format</td>
<td>Data is collected and updated by EUROSTAT from national statistical offices. Data goes through an internal validation process, which should ensure the quality of the data.</td>
</tr>
</tbody>
</table>
### A.6 FlightRadar24

#### 1. General information - Identification of the data source

<table>
<thead>
<tr>
<th>Data source name</th>
<th>FlightRadar24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last update of this file</td>
<td>25/01/2018</td>
</tr>
<tr>
<td>Contact information</td>
<td><a href="https://www.flightradar24.com/data">For the data</a></td>
</tr>
</tbody>
</table>

#### 2. Abstract - Brief description of the data source and its usefulness for the project

Data collected from the API of FlightRadar24
- The key features of this data source are:
  - The coverage of the data is world-wide with special resolution in Europe (ECAC area) and North America
  - Data is available (2015-prepresent)
  - Passive collection, which allows the study of unprogressed events
- The congestion events are detected in real time

#### 3. Availability - Relevant information about owner and readiness to use

<table>
<thead>
<tr>
<th>Owner</th>
<th>FlightRadar24, public API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access conditions</td>
<td>Through public API, access agreement &amp; explicit permission by the owner to use the data for academic research</td>
</tr>
<tr>
<td>Data access</td>
<td>API</td>
</tr>
<tr>
<td>Data cost</td>
<td>Free, although premium plans can be purchased</td>
</tr>
<tr>
<td>Access limitations</td>
<td>The raw data cannot be shared with third parties</td>
</tr>
<tr>
<td>Availability within the project</td>
<td>The raw data cannot be shared, sharing processed data is possible</td>
</tr>
<tr>
<td>Privacy / Confidentiality issues</td>
<td>The data refers to flight performance, there is no privacy issues</td>
</tr>
<tr>
<td>Security issues</td>
<td>No security issues</td>
</tr>
<tr>
<td>Status</td>
<td>Data from mid 2015 preseant, data is continuously base updated</td>
</tr>
<tr>
<td>Link to the data</td>
<td><a href="https://www.flightradar24.com/data">https://www.flightradar24.com/data</a></td>
</tr>
</tbody>
</table>

#### 4. DUBI Characteristics - Temporal, geographical and use characteristics of the information provided

<table>
<thead>
<tr>
<th>Informed size of the sample</th>
<th>The database contains information of around 8 million flights within the ECAC area and around 24 million flights worldwide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographical scope</td>
<td>World wide, the quality of the data fluctuates but it is good enough in the ECAC area and in North America</td>
</tr>
<tr>
<td>Temporal granularity</td>
<td>The information refers to airports and flights. In case the flights in some cases it is possible to extract trajectories with coordinates</td>
</tr>
<tr>
<td>Geographical granularity</td>
<td>To be determined</td>
</tr>
<tr>
<td>Delivery frequency</td>
<td>To be determined</td>
</tr>
<tr>
<td>Delivery delay</td>
<td>Boeing08</td>
</tr>
</tbody>
</table>

#### 5. Quality Issues - Comments about known flaws of the data source

The data is collected by a cooperative effort. The quality in the flight performance data is lower than the one collected by the BTS and ECDA but the coverage is wider.
### A.7 Google Maps Directions API

| **824.903.0205**: Data Identification, Collection & Assessment |
| **7.1 Data Identification** |
| **Potential data sources characterization spreadsheet** |

#### 1. General information - Identification of the data source

<table>
<thead>
<tr>
<th>Data source name</th>
<th>Google data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last update of this file</td>
<td>19/01/2016</td>
</tr>
<tr>
<td>Contact information</td>
<td>Pablo Garcia (Rearman)</td>
</tr>
<tr>
<td>Support</td>
<td><a href="https://developers.google.com/maps/documentation/directions">https://developers.google.com/maps/documentation/directions</a></td>
</tr>
</tbody>
</table>

#### 2. Abstract - Brief description of the data source and its potential usefulness for the project

Google Maps API can be used to obtain route indications, travel time and travel distances. This will be useful to complement Mobile Phone data or Twitter data, which have a lower level of spatial resolution and are not able to determine route/trajects by themselves.

#### 3. Availability - Relevant information about ownership and readiness to use

<table>
<thead>
<tr>
<th>Source</th>
<th>Private (google)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access conditions</td>
<td>For sale (premium account) / Free (limited usage)</td>
</tr>
<tr>
<td>Data access</td>
<td>Through an API</td>
</tr>
<tr>
<td>Data cost</td>
<td>Premium price costs around 6,000 €</td>
</tr>
<tr>
<td>Access limitations</td>
<td>Premium plan: 100,000 requests per day, Free account: 2,500 requests per day. More info at: <a href="https://developers.google.com/maps/premium">https://developers.google.com/maps/premium</a></td>
</tr>
</tbody>
</table>

#### 4. Data Characteristics - Temporal, geographical and use characteristics of the information provided

- Estimated size of the sample: Not applicable
- Temporal scope: Not applicable
- Geographical scope: Worldwide
- Temporal granularity: Not applicable
- Geographical granularity: OSM coordinate level
- Delivery frequency: Not applicable
- Delivery delay: Not applicable
- Data format: JSON

#### 5. Quality issues - Comments about known flaws of the data source

This data source is not useful for characterizing the complete departure/sectoral routes, an flight leg is obtained from sponsored data and does not provide variables/relevant route indications.

#### 6. Comments - Other relevant information

Google Maps Directions API can take into account traffic information. It seems that it uses a model made out from historical data, which also takes into account real-time data for short-term predictions.
## A.8 IATA PaxIS

<table>
<thead>
<tr>
<th>Data Source Name</th>
<th>IATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last Update of this file</td>
<td>11/07/2016</td>
</tr>
<tr>
<td>Contact Information</td>
<td>Private</td>
</tr>
</tbody>
</table>

**Abstract - Brief Description of the Data Source and its Potential Usefulness for the Project**

The data is similar to the airlines' monthly passengers traveling single, multiple, round trip, connecting points, and flights. This information will be useful to verify the new data sources and in the case study relating with intra-airport mobility patterns to estimate how many passengers will be affected by an upgrade in the airline schedule reducing connecting times.

The SIS Data includes worldwide historical schedule data back to January 2000 plus 12 months of future schedules. Full Schedule report, Sample Work reports, 4 report reports, Cost Savings reports, and Map Utility.

**Availability - Relevant Information about owner and readiness to use**

<table>
<thead>
<tr>
<th>Owner</th>
<th>Private (IATA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access conditions</td>
<td>For sale</td>
</tr>
<tr>
<td>Data access</td>
<td>Through an API</td>
</tr>
<tr>
<td>Data size</td>
<td>24.000-12 month access</td>
</tr>
<tr>
<td>Access Restrictions</td>
<td>3 Users</td>
</tr>
</tbody>
</table>

**Data Characteristics - Temporal, geographical and size characteristics of the information provided**

<table>
<thead>
<tr>
<th>Time scale of the sample</th>
<th>Original sample unknown. Data provided theoretically contains an estimation for the 100% of regular commercial flights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal scope</td>
<td>Data available from 2000 onwards, but temporal scope depends on access conditions</td>
</tr>
<tr>
<td>Geographical scope</td>
<td>Europe. It is possible to buy worldwide but it is more expensive</td>
</tr>
<tr>
<td>Delivery frequency</td>
<td>Monthly</td>
</tr>
<tr>
<td>Geographic granularity</td>
<td>Airport level</td>
</tr>
<tr>
<td>Delivery readiness</td>
<td>Monthly</td>
</tr>
<tr>
<td>Data format</td>
<td>CSV</td>
</tr>
</tbody>
</table>

**Quality Issues - Comments about known flaws of the data source**

This information comes mainly from the travel operators, so tickets bought directly to the airlines are annotated and the estimation is not always perfect. It does not cover charter or private flights.

**Comments - Other relevant information**


### A.9 INE population data

#### 1. General Information - Identification of the data source

- **Data source name**: INE (Spanish National Statistics Office)
- **Last update of the file**: 24.07.2012
- **Contact information**: Net applicable
- **Support**: [http://www.ine.es/docs/detalle/0,53846,4412,00.detalle.pdf](http://www.ine.es/docs/detalle/0,53846,4412,00.detalle.pdf)

#### 2. Abstract - Brief description of the data source and its potential usefulness for the project

INE is the Spanish National Statistics Office. It provides demographic information for Spain at a higher level of detail than EUROSTAT. In particular, the data considered to be useful is the 2011 Spanish Census, where population data, characterized by age groups and gender, is available at census tract level.

#### 3. Availability - Relevant information about owner and readiness to use

- **Owner**: Public
- **Access conditions**: Open source
- **Data access**: Data
- **Data cost**: Free
- **Access limitations**: No limitations identified
- **Availability within the project**: 45/300
- **Privacy / Confidentiality issues**: No confidentiality issues identified
- **Security issues**: No security issues identified
- **State**: Already available

#### 4. Data Characteristics - Temporal, geographical and size characteristics of the information provided

- **Estimated size of the sample**: 4.2 million people (around 10% of total population)
- **Temporal scope**: 2011
- **Geographical scope**: Spain
- **Geographical granularity**: Census tract level
- **Delivery frequency**: Census data, every 10 years
- **Delivery delay**: Not relevant (next census is planned for 2021, after BigData4ATM has finished)
- **Data format**: CSV

#### 5. Quality Issues - Comments about known flaws of the data source

Census data was taken every 10 years, with the last one being produced in 2011. This makes these data a little bit obsolete when used for some studies.

Due to statistical methodology applied to survey with data protection legislation, not all census tracts have the same level of detail. Some of them contain information of population both by age and gender, some of them only by gender, and some of them may not contain any categorical information at all. Census data is rounded up to the closest 5.

#### 6. Comments - Other relevant information

Although INE was only discussed 2011 Census Data, INE provides more aggregated and updated data (municipal level) that may be useful for the project. However, these registers are many more in volume.
### A.10 OAG

<table>
<thead>
<tr>
<th><strong>Catalogue of Data</strong></th>
<th><strong>WP 2 Data identification, collection &amp; assessment</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. General information - identification of the data source</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Data source name</strong></td>
<td>OAG</td>
</tr>
<tr>
<td><strong>Last update of the file</strong></td>
<td>28/07/2019</td>
</tr>
<tr>
<td><strong>Contact information</strong></td>
<td>Private</td>
</tr>
<tr>
<td><strong>Support</strong></td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

2. **Abstract - Brief description of the data source and its potential usefulness for the project**

   The data is similar to the SITA One monthly passengers departing origin, destination, 2 transaction points, airport, class and point of sale. In other databases they had information related with schedules, future traffic, area. This information will be useful to verify the new data sources and to the case study related with intra-airport mobility patterns to estimate how many passengers will be affected by an upgrade in the airlines schedules reducing connecting times.

3. **Availability - Relevant information about owner and readiness to use**

   | **Owner** | Private (OAG) |
   | **Access conditions** | For sale |
   | **Data access** | Through API |
   | **Data set** | 26.500 E 12 month access |
   | **Access limitations** | Not applicable |
   | **Availability within the project** | Not purchased |
   | **Privacy / Confidentiality issues** | Data is considered as confidential information. Security requirements are specified in the data access agreement |
   | **Security issues** | To explore |
   | **Name** | **http://www.oag.com/** |
   | **Link to the data** | |

4. **Data characteristics - Temporal, geographical and size characteristics of the information provided**

   | **Estimated size of the sample** | 100% of regular commercial/Flights |
   | **Temporal scope** | The information is available from 2000 to May 2018 |
   | **Geographical scope** | worldwide, it is possible to buy just Europe |
   | **Temporal granularity** | Monthly |
   | **Geographical granularity** | Airport level (it can be aggregated by city, state, country or replica) |
   | **Delivery frequency** | Monthly |
   | **Delivery delay** | One month |
   | **Data format** | CSV |

5. **Quality issues - comments about known flaws of the data source**

   The information comes mainly from the travel agencies, so tickets bought directly to the airlines are automated and this automation is not always perfect. It does not cover charter or private flights.

6. **Comments - other relevant information**
## A.11 Mobile phone records

<table>
<thead>
<tr>
<th><strong>BigData4ATM - WP2 data identification, collection &amp; assessment</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E2C data identification</strong></td>
</tr>
<tr>
<td><strong>Permissible data sources characteristics spreadsheet</strong></td>
</tr>
</tbody>
</table>

### 1. General Information - Identification of the data source

<table>
<thead>
<tr>
<th>Data source name</th>
<th>Orange Spain Mobile Phone Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last update of the file</td>
<td>1/01/2016</td>
</tr>
<tr>
<td>Contact information</td>
<td>Pedro Orriols (ATM4EU)</td>
</tr>
<tr>
<td>Support</td>
<td>not applicable</td>
</tr>
</tbody>
</table>

### 2. Abstract - Brief description of the data source and its usefulness for the project

Mobile phone call (SMS) records contain the geolocated information of mobile phone users every time they make/receive a call, send/receive an SMS or use a data connection to access the internet. The key features of this data source are:

-)throws the capability to identify slow-to-fail trips
- big samples (more than 20% of the total population)
- patron collections, which allows the study of unpredicted events
- contains data of both non-in-motion and non-motion (Orange Spain clients traveling abroad)

### 3. Availability - relevant information about owner and readiness to use

<table>
<thead>
<tr>
<th>CNRN</th>
<th>Private (crosseq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access conditions</td>
<td>Through private agreement</td>
</tr>
<tr>
<td>Data access</td>
<td>See CNRN</td>
</tr>
<tr>
<td>Data cost</td>
<td>Fees specified in the data access agreement on a project by project basis</td>
</tr>
<tr>
<td>Access limitations</td>
<td>No limitations</td>
</tr>
<tr>
<td>Availability within the project</td>
<td>Only Norwegian has access to the disaggregated data</td>
</tr>
<tr>
<td>Privacy / Confidentiality issues</td>
<td>No issues, as data has been previously anonymised by crosseq</td>
</tr>
<tr>
<td>Security issues</td>
<td>Data is considered as confidential information. Security requirements are specified in the data access agreement</td>
</tr>
<tr>
<td>Share</td>
<td>Access Available</td>
</tr>
<tr>
<td>Link to the data</td>
<td>not applicable</td>
</tr>
</tbody>
</table>

### 4. Data Characteristics - Temporal, geographical and size characteristics of the information provided

- **Estimated size of the sample**: 20-30% of Spanish total population (depend on the day)
- **Temporal scope**: From November 2014 to January 2015, and from February 2016 onwards
- **Geographical scope**: Spain
- **Temporal granularity**: Depends on the user usage of the mobile phone. Typically one register every half an hour.
- **Geographical granularity**: National network area (from 200-200m in urban areas to 2.2km in rural areas)
- **Delivery frequency**: Monthly
- **Delivery delay**: One hour
- **Data format**: CSV

### 5. Quality Issues - Comments about known flaws of the data source

Geographic granularity of mobile phone data depends on the mobile network area density, giving spatial granularity of around 200m in big cities, but up to 3km in rural areas.

User profile information may not always be accurate, as the phone (who is the person that appears in the mobile network operator file) may not be the user of the mobile phone (taxpayers who have their phones paid by their parents, for example).

When no network coverage maps are available, coverage areas are used as a proxy of antenna coverage, which is not the optimal approach (they do not take into account obstacles, or antenna technology).

### 6. Comments - Other relevant information

As this data source also captures the mobility of the non-motion (or non-motion) that connect to the Orange network in Spain, it will be explored if this can be exploited to determine mobility patterns and door-to-door trips across Europe.
A.12 Public transport smart card data

<table>
<thead>
<tr>
<th>Data source name</th>
<th>Madrid public transport smart card data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last update of this file</td>
<td>17/01/2016</td>
</tr>
<tr>
<td>Contact Information</td>
<td>pedro carrillo (research)</td>
</tr>
<tr>
<td>Support</td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

**Abstract** - Brief description of the data source and its potential usefulness for the project

This dataset contains all the registers of the Madrid public transport smart card users. A register is produced whenever a user enters in the public transport, and, for some cases, when he gets out. This kind of information will be useful to calculate the mode of transport used to access/leave the airport.

**Availability** - Relevant information about owner and readiness to use

<table>
<thead>
<tr>
<th>Owner</th>
<th>Private (CITVA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access conditions</td>
<td>Through private agreement</td>
</tr>
<tr>
<td>Data access</td>
<td>Batch</td>
</tr>
<tr>
<td>Data cost</td>
<td>Depends on the agreement</td>
</tr>
<tr>
<td>Access limitations</td>
<td>No limitations</td>
</tr>
<tr>
<td>Accessibility within the project</td>
<td>Only measures access to disaggregated data</td>
</tr>
<tr>
<td>Privacy / Confidentiality issues</td>
<td>Tax issue, as data has previously been anonymized by data provider</td>
</tr>
<tr>
<td>Security issues</td>
<td>Data is considered confidential. Security requirements are specified in the data access agreement</td>
</tr>
<tr>
<td>Status</td>
<td>Already available</td>
</tr>
<tr>
<td>Link to the data</td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

**Data Characteristics** - Temporal, geographical and size characteristics of the information provided

- **Estimated size of the sample**
  - Around 60% of Madrid public transport users
  - 9.1 and 10th of March 2016
- **Temporal scope**
  - Madrid public transport
- **Temporal granularity**
  - Precise timestamp of the validation
- **Geographical granularity**
  - GPS coordinates of the public transport stop
- **Delivery frequency**
  - Not applicable
- **Delivery Delay**
  - Not applicable
- **Data format**
  - CSV

**Quality issues** - Comments about known flaws of the data source

For most of the public transport system, it is not needed for the user to use their smart card to get out of the system. Therefore, for many cases, it is only available the register that corresponds to the start of the trip. Also, in many cases, the location of the intensity buses is available, and no location information about metro or train stops, which causes the dataset to not being completely useful.

**Comments** - Other relevant information

- Not applicable
- Not applicable
- Not applicable
- Not applicable
- Not applicable
- Not applicable
# A.13 Sabre

## 1. General Information - Identification of the data source

<table>
<thead>
<tr>
<th>Data source name</th>
<th>SABRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last update of the file</td>
<td>26/01/2016</td>
</tr>
<tr>
<td>Contact information</td>
<td>Jordi L. Namarnirk (RSC)</td>
</tr>
<tr>
<td>Support</td>
<td><a href="https://www.sabre.com/">For the data</a></td>
</tr>
</tbody>
</table>

## 2. Abstract - brief description of the data source and its usefulness for the project

Dataset collected by SABRE and purchased within the context of the SESAR project TRZK.

- Market sector data with origin, destination, intermediate stops, fares and class of the passengers
- Date availability: 2012-2016
- Data in terms of passengers per month in each route
- Price of the ticket

## 3. Availability - Relevant information about owner and readiness to use

<table>
<thead>
<tr>
<th>Owner</th>
<th>SABRE, purchased by RSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access conditions</td>
<td>Providing of an annual access fee to access information during one year</td>
</tr>
<tr>
<td>Data access</td>
<td>SABRE platform</td>
</tr>
<tr>
<td>Data cost</td>
<td>70,000 euros for one year access</td>
</tr>
<tr>
<td>Access limitations</td>
<td>The raw data cannot be shared with third parties</td>
</tr>
<tr>
<td>Availability within the project</td>
<td>The raw data cannot be shared, sharing processed data is possible</td>
</tr>
<tr>
<td>Privacy / Confidentiality issues</td>
<td>The data refers to aggregated passenger flows, there are no privacy issues</td>
</tr>
<tr>
<td>Security issues</td>
<td>The data cannot leave RSC, access to the data is only possible in the intranet</td>
</tr>
<tr>
<td>Data format</td>
<td>Data from 2009-2015</td>
</tr>
<tr>
<td>Link to the data</td>
<td><a href="https://www.sabre.com/">https://www.sabre.com/</a></td>
</tr>
</tbody>
</table>

## 4. Data Characteristics - Temporal, geographical and size characteristics of the information provided

- **Estimated size of the sample**: The database contains market sector data worldwide. The flow of passengers is detailed in over 140 million "routes", which includes origin, final destinations and intermediate stops (connecting airports).
- **Temporal scope**: 2009-2015
- **Geographical scope**: Worldwide, the quality of the data fluctuates but it is good enough in the ECAC area and in North America.
- **Geographical granularity**: Monthly
- **Data format**: To be determined
- **Delivery frequency**: To be determined
- **Delivery delay**: MongoDB

## 5. Quality Issues - Comments about known flaws of the data source

The monthly resolution forced us to interpretate smaller time units due to be used. SABRE also did some estimations in case the information on the flight ticket sales was not fully available to them. This can introduce some biases.

## 6. Comments - Other relevant information
A.14 STATFOR public reports

<table>
<thead>
<tr>
<th>BigData4ATM – WP9 Data identification, collection &amp; assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1. Data identification</td>
</tr>
<tr>
<td>Pooled data source characteristics spreadsheet</td>
</tr>
</tbody>
</table>

1. General Information - Identification of the data source

<table>
<thead>
<tr>
<th>Data source name</th>
<th>STATFOR reports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last update of the file</td>
<td>06/10/2016</td>
</tr>
<tr>
<td>Contact information</td>
<td>not applicable</td>
</tr>
</tbody>
</table>

2. Abstract - Brief description of the data source and its potential usefulness for the project

STATFOR is the Statistics and Forecast Service of EUROCONTROL, and has the objective to provide statistics and forecasts on air traffic in Europe. STATFOR offers information through the SID (STATFOR Interactive Dashboard) and through periodic reports. As access to SID has not been granted to the project yet, here we describe the information provided through the periodic reports. This information consists of:
- Traffic forecasts for short, medium and long term
- Service units forecasts for short and medium term

Short term: ≤ 2 years, medium term: (2-5 years) and long term: >5 years forecasts

3. Availability - Relevant information about owner and restrictions to use

<table>
<thead>
<tr>
<th>Owner</th>
<th>access control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>public</td>
</tr>
<tr>
<td></td>
<td>online</td>
</tr>
<tr>
<td></td>
<td>free</td>
</tr>
<tr>
<td></td>
<td>none</td>
</tr>
</tbody>
</table>

Access limitations: none

Availability within the project: Available for all members

Privacy / confidentiality issues: Reports do not contain any of these issues

Security issues: None

Stat: Easily available

http://www.eurocontrol.int/article/forecasts

4. Data Characteristics - Temporal, geographical and size characteristics of the information provided

<table>
<thead>
<tr>
<th>Estimated size of the sample</th>
<th>not applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal scope</td>
<td>not applicable</td>
</tr>
<tr>
<td>Geographical scope</td>
<td>reflects</td>
</tr>
<tr>
<td>Temporal granularity</td>
<td>nearly</td>
</tr>
<tr>
<td>Geographical granularity</td>
<td>country level</td>
</tr>
<tr>
<td>Delivery frequency</td>
<td>two flight reports per year, up to three service units reports per year</td>
</tr>
<tr>
<td>Delivery delay</td>
<td>two months</td>
</tr>
<tr>
<td>Data format</td>
<td>pdf</td>
</tr>
</tbody>
</table>

5. Quality issues - Comments about known flaws of the data source

Granularity is too coarse (country level, yearly) for the studies conducted for the project, therefore access to SID will be requested

6. Comments - Other relevant information


## A.15 TomTom

### 1. General information - Identification of the data source

<table>
<thead>
<tr>
<th>Data source name</th>
<th>TomTom data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last update of this file</td>
<td>30/09/2016</td>
</tr>
<tr>
<td>Contact information</td>
<td>Javier TomTom (Netherlands)</td>
</tr>
<tr>
<td>Support</td>
<td><a href="http://developer.tomtom.com/support">http://developer.tomtom.com/support</a></td>
</tr>
</tbody>
</table>

### 2. Abstract - Brief description of the data source and its potential usefulness for the project

TomTom On-the-road sensing API can be used to obtain route indications, travel time and travel distances by road in private transport. This will be useful to complement Mobile Phone data or Twitter data, which have a lower level of spatial resolution and are not able to determine the travelling route by themselves.

### 3. Availability - Relevant information about owner and readiness to use

<table>
<thead>
<tr>
<th>Owner</th>
<th>Private (TomTom)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access conditions</td>
<td>Free (Premium account) / Free (limited usage)</td>
</tr>
<tr>
<td>Data access</td>
<td>Through an API</td>
</tr>
<tr>
<td>Data cost</td>
<td>Contact with sales department. No public information about pricing</td>
</tr>
<tr>
<td>Access limitations</td>
<td>Free access; 1,000 calls per day and 5 calls per second</td>
</tr>
<tr>
<td>Availability within the project</td>
<td>BigData4ATM project partners that have access to the data</td>
</tr>
<tr>
<td>Privacy / Confidentiality issues</td>
<td>No confidentiality issues identified</td>
</tr>
<tr>
<td>Security issues</td>
<td>No security issues identified</td>
</tr>
<tr>
<td>Issues</td>
<td>Already available</td>
</tr>
<tr>
<td>Link to the data</td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

### 4. Data characteristics - Temporal, geographical and size characteristics of the information provided

<table>
<thead>
<tr>
<th>Estimated size of the sample</th>
<th>Not applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal scope</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Geographical extent</td>
<td>Worldwide</td>
</tr>
<tr>
<td>Temporal granularity</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Geographical granularity</td>
<td>GPS coordinates level</td>
</tr>
<tr>
<td>Delivery frequency</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Delivery-Delay</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Data format</td>
<td>xml, json or csv</td>
</tr>
</tbody>
</table>

### 5. Quality issues - Comments about known flaws of the data source

This data source is not valid for characterising the complete door-to-door mode, as there is no information about public transport (bus, subway, etc.) and flights. Only private road vehicle data is provided.

### 6. Comments - Other relevant information

TomTom Online Routing API uses real-time (when the departure time of the travel is close to “now”) and the information is available) and historic traffic information.

TomTom Online Routing API allows to see waypoints in the route and returns an optimised route order for the waypoints.
A.16 Twitter

1. General Information - Identification of the data source
   - Data source name: Twitter geolocated data
   - Last update of this file: 22/06/2015
   - Contact information: Jane J. Fernandez (PITT)
   - Support: For EU data: https://dax.twitter.com/overview/documentation

2. Abstract - Brief description of the data source and its usefulness for the project
   Data collected through the Twitter API with a fraction of the global geolocated tweets posted across the world.
   The key features of this data source are:
   - High coverage of the area is consistent with
   - Large sample (2012-present)
   - Passive collection, which allows the study of unreported events
   - Users may employ this media as a channel to express their problems with the air transport network

3. Availability - Relevant information about owner and readiness to use
   - Owner: Twitter, public API
   - Ready: Through public API access agreement
   - Data access: API
   - Data cost: Free
   - Access limitations: The raw data cannot be shared with third parties.
   - Availability within the project: The raw data cannot be shared, sharing processed data is possible.
   - Privacy / Confidentiality issues: The data contains strictly personal information. An automatic method has been developed to store only anonymized information and delete private data.
   - Security issues: Data is considered as confidential information. Security requirements are specified in the data access agreement.
   - Data from 2012-present, data is continuously being acquired.
   - API documentation: https://dev.twitter.com/overview/documentation

4. Data Characteristics - Temporal, geographical and size characteristics of the information provided
   - Estimated size of the sample: 7.2 billion tweets, at a million new tweets per day.
   - Temporal scope: From 2012 to present.
   - Geographical scope: Worldwide, the focus is on the full Europe even though the penetration rate of this technology is country-dependent.
   - Geographical granularity: Depends on the use of Twitter on mobile phones.
   - Geographical granularity: 64 coordinates of the points where the message was sent.
   - Geographical granularity: To be determined.
   - Geographical granularity: MongoDB

5. Quality Issues - Comments about known flaws of the data source
   - The use of geo-coordinated Twitter is not homogeneous across the continent, correction factors are needed to take this into account.
   - The sample of the population is biased toward younger individuals.

6. Comments - Other relevant information
Appendix B  Detailed characterisation of Big Data sources

B.1 Mobile phone data

B.1.1 Data inputs

Nommon Solutions and Technologies has access to disaggregated data from Orange Spain through a private agreement between the parties. These data consist of call detail records (CDRs), sociodemographic information about Orange clients and the geographic location of the telecommunication towers.

This analysis has been performed for the 2nd of March of 2016, Wednesday, a typical working day.

B.1.2 CDRs

Every time a mobile phone interacts with the network through a voice call (receiving or sending), SMS or internet data connection, this event is recorded by the network operator for billing purposes. This is called a Call Detail Record (CDR). The CDR format is not standardised, and it may vary across the different mobile network operators. Typically, the main information of CDRs, from which it is possible to obtain users’ activity and mobility information, is the following:

| User 1 ID | Origin tower cell ID | Date | Time |

• User 1 ID: it is a code that identifies the mobile phone user that connects to the network because of an activity (voice call, SMS or data connection). Before providing CDRs to a third party, it has to be anonymised because of privacy concerns, usually through a hash function.

• Origin cell ID: this number identifies the mobile network cell to which user 1 connects.

• Date: the date when the event is registered.

• Time: the time when the event starts.

Sometimes, the CDRs contains additional information, such as the receiver of the call/SMS and the cell ID to which he is connected, Mobile Country Code (MCC) of the user(s) or the time when the call/data session finishes.

Orange Spain also provides roamers CDRs, which, together with the MCC, allows the analysis of tourists’ mobility and activity patterns.

B.1.3 Socio-demographic information

For each client, Nommon has access to the following information:

| User 1 ID | Nationality | Gender | Year of birth |

With this data it is possible to segment the population of the sample into different groups for future uses, such as studying the behaviour of each group and expanding the sample to the whole population.
B.1.4 Sample size

In the following table, the number of Orange clients is presented, together with some extra information about roamers. We can distinguish three types of roamers:

- Orange clients with registers both in Spain and abroad (roamers-out)
- Orange clients with registers only abroad (roamers-out)
- Non Orange clients with registers in the Orange Spain network (roamers-in)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain’s population</td>
<td>46,624,382</td>
</tr>
<tr>
<td>Sample users</td>
<td>8,877,537</td>
</tr>
<tr>
<td>Orange clients with records only in Spain (not roamers-in)</td>
<td>8,498,512</td>
</tr>
<tr>
<td>Orange clients with records both in Spain and abroad</td>
<td>16,344</td>
</tr>
<tr>
<td>Orange clients with records only abroad</td>
<td>30,476</td>
</tr>
<tr>
<td>Roamers-in</td>
<td>332,205</td>
</tr>
<tr>
<td>Rate of the sample to the total population</td>
<td>19%</td>
</tr>
<tr>
<td>Total records</td>
<td>628,926,007</td>
</tr>
</tbody>
</table>

It has to be said that the numbers that appear in the table above refer only to clients with some activity during the day of study (2 March 2016). This is why these numbers are different from Orange’s total number of clients, which according to official statistics is around 13 million. Also, official statistics typically include Mobile Virtual Network Operators (MVNO) owned by Orange, which have been excluded from the data provision agreement. For the analysis of longer periods, and if MVNO data is finally available for analysis, the sample size will presumably be higher than 19%.

B.1.5 Distribution of the records per user

The table below shows the distribution of the number of records per user. It is seen that more than half of the population has 50 or more records per day.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>71</td>
</tr>
<tr>
<td>Maximum</td>
<td>15,968</td>
</tr>
<tr>
<td>Percentile 85</td>
<td>121</td>
</tr>
<tr>
<td>Percentile 50</td>
<td>49</td>
</tr>
<tr>
<td>Percentile 15</td>
<td>8</td>
</tr>
<tr>
<td>Minimum</td>
<td>1</td>
</tr>
</tbody>
</table>
B.1.6 Socio-demographics

Figure 22: Gender distribution of the sample

Figure 23: Age distribution of the sample

Figure 24: Availability of gender and age of the sample

Figure 25: Comparison between Orange sample and Spanish population pyramids
In the graph above it is compared the distribution of the sample to the distribution of the Spanish population according to INE (Spanish Statistical Office). It is noticed that distributions differ markedly. The percentage of people aged 35-59 in the sample is much greater than in Spain, whereas in the sample there is almost no people under 20 years. These differences may be explained by the following reasons:

- Children do not have mobile phones.
- Teenagers have the contract under the name of their parents, what explains the higher percentage of people aged 30-55 years in the sample than in the Spanish population and the lack of population under 25 in the sample.
- Teenagers may opt for Mobile Virtual Network Operators (MVNO). In the future, data of some MVNO which use the Orange’s network is expected to be available, so this problem may be mitigated.

### B.1.7 Temporal characteristics

**Figure 26: Histogram of the number of records per user**

**Figure 27: Maximum time between records per user**
The graph of Figure 28 represents the time at which maximum time between records occurs. This graph has been built as follows: if a user has his maximum time between records during 1:00 and 9:00, this user is represented in all the bars between [1-2) and [8, 9).

**Hourly intervals**

These three graphs show that the activity of users is much higher during the day than during the night, as it could be expected. However, there is still activity at night time, mainly due to automatic connections made by apps. The peaks of activity are at 12-14 h and 18-20 h and there is a minor decrease at lunchtime.
Three-hour intervals

![Figure 32: Users and records per three hour intervals](image1)

![Figure 33: Records per user per three hour intervals](image2)

![Figure 34: Number of three hour intervals with user activity](image3)

Six-hour intervals

![Figure 35: Users and records per six hour intervals](image4)

![Figure 36: Records per user per six hour intervals](image5)
Time intervals between consecutive mobile data sessions

When access to Internet connections/data sessions is available, the information flow and the frequency of temporal sampling increases substantially compared to only voice calls and SMS events. This is because time to time data sessions may be renewed and some apps do not require user interaction, and may send or request information automatically (for example, looking for updates, or checking e-mail). This can be seen in the figure above, where the time between data sessions has a peak every 30 minutes, 60 minutes, etc.

B.1.8 Spatial characteristics

Orange has more than 16,000 telecommunications towers in Spain in order to provide coverage to its clients. The spacing between these towers depends on the population present in the area to which they provide coverage (in urban, densely populated areas, distances are of hundreds of metres, while in rural areas distances are of kilometres).

The coverage area for each Mobile Network Tower is approximated through a Voronoi Map. This works under the assumption that users connect to their closest tower. This assumption works reasonably well for most of the cases, and has been the traditional approach to work with CDRs when no real coverage maps or triangulated signal information is available.

The Voronoi Polygons created for the tower’s network shows the differences in tower density between urban and rural areas. It is also easy to distinguish the main roads.
Converting the Voronoi Polygons to circles with the same area, the radius of coverage of the towers is distributed as follows:

<table>
<thead>
<tr>
<th>Distribution of radius of coverage (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
</tr>
<tr>
<td>Percentile 20</td>
</tr>
<tr>
<td>Percentile 40</td>
</tr>
<tr>
<td>Percentile 50</td>
</tr>
<tr>
<td>Percentile 60</td>
</tr>
<tr>
<td>Percentile 80</td>
</tr>
<tr>
<td>Maximum</td>
</tr>
<tr>
<td>Average</td>
</tr>
</tbody>
</table>

Looking at this graph, it may appear that the spatial density is not good enough to analyse the dynamics of the population properly, but this is because a big percentage of the towers are placed in rural areas, where Voronoi polygons are too large and only few people live. To compare how different is the size of Voronoi polygons in urban areas, the same analysis has been conducted in the metropolitan area of Madrid, where there are 613 towers covering an area of 264 km². In the picture it is easily observable that the further we are from the city centre, the larger the Voronoi polygons.
The radius of coverage of the towers is distributed as follows:

<table>
<thead>
<tr>
<th>Distribution of radius of coverage (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
</tr>
<tr>
<td>Percentile 20</td>
</tr>
<tr>
<td>Percentile 40</td>
</tr>
<tr>
<td>Percentile 50</td>
</tr>
<tr>
<td>Percentile 60</td>
</tr>
<tr>
<td>Percentile 80</td>
</tr>
<tr>
<td>Maximum</td>
</tr>
<tr>
<td>Average</td>
</tr>
</tbody>
</table>

More than 50% of the towers have a radius of coverage shorter than 300 m, and those towers whose radius of coverage is larger than 500 m are located in the outskirts of the city, where the population density is lower than in the city centre.
B.1.9 Roamers

Roamers-in

Roamers-in are those agents who are not Orange Spain clients that connect to Orange Mobile Phone Network. These agents may be either people from other countries or clients of other Spanish MVNOs that do not have coverage in certain areas and use Orange network.

As they are not Orange Spain clients, there is no data about their socio-demographics so it will not be feasible to segment this group in sub-groups depending on gender and age.

From this data, the origin country of people visiting Spain can be obtained, and also, analysing their mobility, it may be possible to determine the transport mode used (for example, if their first/last records are close to an airport it can be assumed that they have travelled by plane).

According to the data from Orange, the origin country of roamers-in is the indicated in the graph below.

![Graph showing origin country of roamers-in](image)

Comparing this distribution of the origin country of roamers-in with the data about foreign visitors available from INE for March 2016, there are significant differences in the percentage of some countries such as Germany, France and Italy.

![Graph showing origin country of tourists according to INE](image)

This difference may be mainly explained by the market share of Orange and its agreements with local companies in those countries. In the case of France, it may also be explained by people crossing the border between Spain and France for short journeys that are not measured by INE.
When looking at the temporal characteristics of the registers produced by roamers-in, it can be seen that the activity of foreign people in Spain is lower than average as it could be expected due to expensive roaming rates.

![Figure 44: Number of records per roamer-in](image)

![Figure 45: Roamers-in and records per three hour interval](image)

![Figure 46: Distribution of records per roamers-in per three hour interval](image)

![Figure 47: Number of three hour intervals with roamers-in activity](image)
Roamers-out

Roamers-out are those clients of Orange Spain, who connect to another mobile operator when staying abroad.

Although the location of the cells to which roamers-out connect is not available (as these cells belong to other MNO), some information can be extracted. First, cell id contains the MCC, from which the country can be inferred. Also, with external sources (such as OpenCellID), a proxy of the cell location can be obtained. With this information, similarly to what we have said above with roamers-in, we can obtain the countries that Spanish population visit and the transport mode used.

The total number of out-roamers in the sample is 46,820 clients. 16,344 of these clients have some record in Spain, so they have travelled in the analysed day.

![Figure 48: Comparison between roamers-out sample and Spanish population pyramids](image1)

Compared to the pyramid of the whole sample, it is observed that in the subgroup of out-roamers the percentage of men is higher and also that people travelling abroad is younger than average.

Roamers-out may have records in different countries in the same day, so the sum of the percentage of each country is higher than 100%.

![Figure 49: Destination countries for roamers-out](image2)

When looking at the temporal characteristics of the registers produced by roamers-out, it can be seen that their activity seems high enough to determine their mobility patterns.
Figure 50: Number of records per roamer-out

Figure 51: Roamers-out and records per three hour interval

Figure 52: Distribution of records/user in each interval

Figure 53: Number of three hour intervals with roamer-out activity
B.2 Google Maps

B.2.1 Google Maps APIs overview

Google Maps offers a variety of map/location-based data APIs. A preliminary analysis on these APIs has been conducted, and those that may be the most useful for BigData4ATM have been identified. These are the following APIs:

**Google Maps Directions API**

This API calculates route options and their associated travel times given an origin and a destination. These routes can be obtained for different transport modes, such as walking, private transport or public transport. Also, this API accepts waypoints or departure/arrival time information to better characterise the trip. This API is very useful to determine the route followed during trips detected by other source (such as mobile phone data or Twitter data). Also, this API can be used to calculate indicators of airport accessibility and connections with other transport modes.

**Google Maps Distance API**

This API calculates distance and travel time between locations, with mode and departure/arrival time options. The main difference between this API and Directions API is that this API does not return the route information, and also that it can calculate several origin-destination pairs during the same call to the web service.

**Google Places API**

This API provides information about places present in an area. These places can be stores, services, localities, points of interests, etc. This information would be useful for adding a better characterisation of the passenger behaviour by inferring the activities performed by people during their visits to a place.

**Google Maps Geocoding API**

This API translates human-readable addresses into geographic coordinates and vice versa. This service is useful for the analysis of credit card data, where businesses coordinates are not always correct.

**Google Maps Roads API**

This API generates the route followed by a user from GPS traces. It may be interesting if the project gets access to GPS data, such as data from a mobile App.

**Google Maps Geolocation API**

This API provides a location based on cell phone towers and Wi-Fi nodes that the mobile phone can detect. This information can be useful to locate mobile phone users when they are abroad, as mobile network operators only know the location of their own towers.
B.2.2 Google Maps Directions API in-depth analysis

During the preliminary analysis conducted in B.2.1, Google Maps Directions API was identified as the API from which more information useful for BigData4ATM could be obtained. Therefore, a more detailed analysis of this API was carried out to evaluate whether it was worth purchasing a premium plan or not.

Characteristics

As said before, the Google Maps Directions API takes a pair of locations and calculates the route and travel time between them. The locations can be expressed in latitude/longitude or a string containing an address (in this case, the Directions API will geocode this string and convert it into coordinates). The main advantages of this API for BigData4ATM are the following:

Google transport network. Google Maps Directions API uses Google road network, which has global coverage and is constantly being updated. Also, Google Maps includes public transport data from a large set of cities (list available in https://maps.google.com/landing/transit/cities/). Having this kind of information already precalculated with a high level of accuracy is crucial for a project like BigData4ATM, where there would be no time to create a transport network, gather public transport timetables and merge all these into an operational network model.

Traffic model. Google Maps Directions API has an embedded traffic model that takes into account both historical and real-time traffic data when calculating routes. Therefore, travel time estimations are expected to be quite accurate. As already mentioned, building a traffic model is out of the scope of the project, and therefore access to this information would be of great value for the project.

Moreover, the information provided by the API can be enhanced with additional parameters. Here are detailed the parameters that are more directly related with the project, but more parameters can be included for route calculation (such as language, optimistic/pessimistic traffic predictions, avoiding tolls/highways, etc.).

Mode. This parameter sets the transport mode used for the trip. The following modes are supported: driving, cycling, walking and public transport. When specifying the mode as public transport, the public transport mode(s) preferred can also be specified, between the following: bus, subway, train or tram.

Departure/Arrival time. The API accepts a departure time or an arrival time (but not both) for the trip. This parameter has an impact on the route calculation:

- For private transport, as the API uses its own traffic model, the routes proposed during peak hours, when roads are typically congested, may not be the same as the routes during valley hours.
- For public transport, as the API has timetables information embedded, travel times (or even routes proposed) may differ with the departure/arrival time specified.

As one of the objectives of BigData4ATM is to measure how flight delays impact into total door-to-door travel time, this characteristic seems to be key for the project. However, it has to be remarked that the API departure/arrival time cannot be set to a past time.
Waypoints. This parameter consists in a set of intermediate locations through which the route calculated must pass through. This has interest for the case of Twitter data, where tweets made during the travel may help to better determine the route, or for the case of consecutive trips with intermediate stops detected from mobile phone data.

Alternatives. If specified, the API can calculate more than one route option for the origin/destination given. Similarly to waypoints, this capability of the API can help to better determine the route followed when combined with other data sources.

Limitations
The following limitations of the Google Maps Directions API have been identified:

Number of daily queries. Free access to Google Maps Directions API grants a maximum of 2,500 queries per day. This number of queries is way below the estimations of the queries that will be needed for the project. As an example, when working with mobile phone data, a typical day has around 16,000 Spanish people that have records both in Spain and in another country, which shows a potential trip to the airport that needs to be studied, to which we should add potential trips to/from the airport made by roamers. This limitation can be overcome with the Premium Plan, which allows a maximum of 100,000 queries, which is considered acceptable for the studies planned to be carried out during the project.

Sponsored flights. When asking Google Maps for directions between a pair of points that are connected by flights, Google may suggest some air routes. However, these air routes provided by Google Maps are sponsored, and therefore the information provided does not cover all possible flights and, moreover, it may show completely unreasonable alternatives (due to the cost or the number of flight connections). But, as other data sources such as mobile phone records are able to detect origin and destination airports, Google Maps Directions API will only be used to determine the door-to-kerb and kerb-to-door route, mitigating this limitation.

Historical data are not accessible. As said before, Google Maps Directions API provides travel time estimations that take into account traffic levels. This is done through an embedded traffic model that takes into account actual traffic and historical traffic data to provide travel time estimations. However, travel time estimations are only available for future dates, which means that the traffic data for specific days in the past is not accessible. This can be a limitation when analysing disruptions, as the travel times will not take into account the real traffic that occurred during that day. A further analysis of the traffic model has been carried out in next section.

Traffic model analysis
As said in previous sections, Google Maps Directions API uses its own traffic model to provide travel time estimations. This model takes into account historical and real-time traffic data to produce its travel time predictions. In order to understand how this model merges these historical and real-time data, an analysis of the model has been carried out.

The analysis consisted in obtaining the travel time between two locations for different dates and times of the day, and comparing the different results provided by the model. The locations considered for the analysis were Nuevos Ministerios (a centric, business area of the city of Madrid) and the Madrid-Barajas airport.
Three characteristics of the model were studied through this analysis:

- Variation of travel time with the hour of the day
- Variation of travel time for different days
- Influence of real-time traffic in travel time estimations

Variation of travel time with the hour of the day

The main purpose of this part of the analysis was to check whether the traffic model was able to capture the dynamics of the city traffic or not. Therefore, for a typical working day and a weekend day, driving travel times from Nuevos Ministerios to Madrid Airport were obtained for each hour of the day. To ensure that no real-time traffic data was altering the results of this analysis, the days of study chosen were Tuesday 5th September 2017 and Sunday 10th September 2017 (more than one year after the day when the analysis was made). Results are present in Figure 55.
As it can be seen, for a working day, the model is able to capture the morning and afternoon peaks, together with a midday peak (explained by people using their vehicles during lunch break). Also, the shortest travel times occur during early morning, which is consistent with the results expected. In contrast, the travel times distribution during the weekend is very different, much more constant and with no big peaks. Therefore, it seems that the traffic model adequately reflects the behaviour of the city traffic.

**Variation of travel time for different days**

The main purpose of this part of the analysis was to check whether the traffic model takes into account the day of the week. For this purpose, driving travel times from Nuevos Ministerios to Madrid Airport were obtained for the days of the week of the 4th September 2017 for morning peak (8:00 AM). Again, the effect of real-time traffic into the travel time estimations was not desired, therefore the week of study was chosen one year after the analysis. Results are shown in Figure 56.

![Variation of travel time with day of the week](image)

**Figure 56: Travel time vs day of the week**

By looking at the results, it is clear that the model takes into account the day of the week, as the travel times are different. It is also worth noting the big difference between working days and weekends, where the congestion is much lower and therefore travel times are smaller. Also, different weeks were analysed (weeks in August, October and Christmas week), obtaining the same travel times. Therefore, it seems that the traffic model takes into account only day of the week, but not the specificities of each month or festivities.

**Influence of real-time traffic on travel time estimations**

The main purpose of this part of the analysis was to check the influence of real-time traffic on travel time estimations. This was done by asking Google for estimations of the driving travel times from Nuevos Ministerios to Madrid Airport for the “current time” (which was Tuesday 30th August 2016 at 12:45) and successive times. Then, these times were compared with the equivalent times for an August Tuesday one year in advance. The results are shown in Figure 57.
Figure 57: Influence of real-time traffic

The results show that the traffic of 30\textsuperscript{th} August was lighter than the traffic predicted by the model (which is reasonable, as at the end of August there are still people on holidays and we saw before that the traffic model does not take into account further detail than day of the week, so these seasonal variations would not be captured by the “historical model”). Also, it can be seen that travel time estimations get equal for 2h or more from the “current time”. This means that the model only takes into account the real-time traffic for travel time estimates within the next 2 hours.

Potential applications

In this section, a couple of potential applications that have been identified during the analysis of the Google Maps Directions API are detailed, with some preliminary results.

Airport accessibility. Airport accessibility is one of the key factors that may influence passengers’ decisions about choosing their flight when they need to travel. Airport accessibility can therefore be a tool for assessing competition between airports or with other transport modes. There exists a broad variety of accessibility indicators, ranging from very simple ones (travel time, number of trips to a zone) to complex, compound indicators. As an example of application for Google Maps Directions API, a map of the driving travel times to airport has been obtained for the city of Madrid for three times of the day: morning peak, valley hour and afternoon peak:

Figure 58: Driving travel times at morning peak
These basic travel times can be obtained with Google Maps Directions API for other transport modes, such as railway, bus or subway, giving a picture of the different access modes to the airport. Also, this information can be combined with other data sources (such as trips from Twitter/mobile phone data, or expenditure data from bank cards) to create more sophisticated accessibility indicators.

**Route/Mode determination.** Google Maps Directions API data can also be used to enrich the analysis made with other data sources. In the case of mobile phone data, the mobility of a great sample of the population is characterised, but this characterisation lacks some important details, such as route followed or transport mode used. By combining mobile phone data with Google data, it is sometimes possible to overcome these gaps.

As an example of application, a preliminary characterisation of the modal share in the trips between Madrid and Catalonia has been made with the combination of mobile phone and Google Maps data. First, with Google Maps Directions API possible routes between Madrid and Catalonia for high-speed train (blue) and road (red) are calculated. Plane routes (green) for that day were also obtained from ATM databases. Then, mobile phone registers are compared against these routes, and mode can be determined (even for people travelling by plane, sometimes mobile phone registers appear during landing and take-off). This methodology will also be explored to determine the transport mode used to get to the airport.
Conclusions

After the analysis made, it can be concluded that the information provided by Google Maps APIs can be very useful for BigData4ATM. This data source has a great value on its own, due to the availability of public transport networks and timetables for a great number of cities, which offers the possibility of performing interesting airport accessibility analysis. In addition, it has important synergies with other data sources that will be used within the project (such as mobile phone data, Twitter data or public transport smart card data). However, due to the limited number of queries that can be performed with a free account, a Premium Plan shall be used. Also, limitations regarding the traffic model of the Directions API need to be taken into account when analysing past events.
B.3 TomTom Traffic services

Similarly to Google Maps APIs, TomTom offers a variety of map and location based services. The information provided by TomTom comes from their GPS navigator users. The TomTom services analysed are described below.

B.3.1 Online Traffic Flow

Online Traffic Flow is a suite of web services based on the real-time traffic data from TomTom. It provides:

Flow Segment Data

Given the coordinates (EPSG4326) of a point, it returns information about current speed, free-flow speed and a confidence parameter, which measures the quality of the provided data, of the road fragment closest to it.

Flow Tiles

Given a location, it returns either the current speed or the relative current speed to free-flow speed (in green, current speed equals to free-flow speed) of its road segments.

B.3.2 Online Routing

Similarly to Google Maps Directions, Online Routing calculates as many possible routes as desired between an origin and a destination location, optionally via way-points, selecting the preferred route (fastest, shortest, eco) and taking into account factors such as traffic and departure/arrival time.

It is also possible to select the transport mode (car, bus, truck, bicycle and pedestrian) and the characteristics of the vehicle (height, width, load) which will be used to calculate the available routes for each type of transport and the specific speed of each transport. In this API it is not possible to use public transport.

To calculate the possible routes, TomTom uses historical data to study past traffic patterns and predict journey times. This calculation is based on actual average speed data, rather than permitted speed limits.

We carried out an analysis to study the output data with origin Nuevos Ministerios and destination Madrid Barajas (Madrid Airport). Different characteristics of the model were studied through this analysis.
Possible routes

To get different routes between origin and destination, we selected the three fastest routes. For each route, we obtained a summary of the route, which includes the length, the travel time, the departure and arrival date, and a list of points (latitude, longitude) of the route making it possible to display the routes.

<table>
<thead>
<tr>
<th>Route</th>
<th>Length (m)</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route 0</td>
<td>13,192</td>
<td>942</td>
</tr>
<tr>
<td>Route 1</td>
<td>14,398</td>
<td>983</td>
</tr>
<tr>
<td>Route 2</td>
<td>18,847</td>
<td>1,468</td>
</tr>
</tbody>
</table>

Variation of travel time depending on hour and day of the week

In order to analyse the variation of the travel time depending on the hour of the day and the day of the week, we carried out an analysis of the travel time hour by hour for the days of the week of the 4th September 2017. The main purpose was to check whether the traffic model takes into account the hour and the day of the week.

Variation of travel time with the hour of the day

As it can be seen in the chart, the analysis captures the dynamics of the city traffic along the day and also along the weekdays. In the working days there are three peaks (explained by people using their vehicle to go to work at morning, lunch time and evening), whereas in the weekend, there is not a clear pattern. The shortest travel time is on Sunday at midday, when the fastest route is also the shortest (travel through the city centre) since there is no congestion in the city.
Variation of travel time with the day of the week

These graphs, which indicate the travel time from Nuevos Ministerios to Madrid Barajas for each day from the 4th September 2017 (Monday) to 10th September 2017 (Sunday) at 8.00 am., make it clear that the travel time depends on the day of the week, not only if it is a working or non-working day, obtaining different results for each day of the week. The chart indicates that there is a big difference between working and non-working days, since the car is mainly used to go to work creating more congestion on working days and therefore longer periods of travelling.
The analyses explained in this section were also carried out for different months and seasons of the year obtaining the same results for any date. This means that the only parameter that is taken into account when planning a future route is the day of the week, regardless of the month/season.

**Variation of travel time depending on hour and day of the week**

When calculating a route, TomTom estimates the travel time based on its database and the current traffic, which may add some seconds to the duration of the travel (Traffic delay). For future calculations, the traffic delay equals to zero while in routes calculated for the present it may be positive depending on traffic conditions. According to these considerations, if the estimated congestion in the present is not greater than the average congestion for the same weekday, the planned travel time for the present will be the same as the planned time in a future day for the same weekday.
B.3.3 Traffic Stats

The Traffic Stats portal makes it possible to analyse historical travel times and speeds on any road or route, over any calendar period and time of day based on TomTom’s database (collection of anonymous journeys made by vehicles with GPS devices).

Custom Area Analysis

Custom Area Analysis (CAA) allows users to query the TomTom historical database of vehicle trips to determine precise traffic and travel characteristics in an area.

The analysis was carried out in the city of Madrid for the dates 1st April to 30th April 2016.

The main output of CAA is the actual distribution of speeds in each road segment in the period of time analysed. As output we will get, for each road segment, the number of measurements used for calculation and the distribution of speeds:

- Id of the segment
- Number of measurements used for calculation (hits)
- Average speed.
- Median speed.
- Standard deviation of the speed.
- 5th, 10th, 15th... 95th percentile speed.

Custom Travel Times

The Custom Travel Times (CTT) calculates a statistics for a defined route between an origin and a destination, optionally via waypoints, using TomTom historical database of vehicle trips, as in CAA.

The analysis was carried out in the motorway A-1 exiting Madrid towards Burgos from dates 18th July to 24th July 2016.
As in CAA, the output of CTT is the actual distribution of speeds in each road segment in the period of time analysed and the average and median travel time. As output we will get, for each road segment, the number of measurements used for calculation, the average and median travel time and the distribution of speeds:

- Id of the segment
- Length of the segment
- Number of measurements used for calculation (hits)
- Average speed.
- Median speed.
- Average travel-time.
- Median travel-time.
- Standard deviation of the speed.
- 5th, 10th, 15th... 95th percentile speed.

From these images, two conclusions can be drawn. The first is that the further we are from Madrid the less hits we have due to people travelling to villages close to Madrid who do not complete the whole route by the motorway, and the second is that after road junctions, the speed goes down because of the incorporation of more cars to the road, increasing congestion.

**B.3.4 Conclusions**

After analysing and comparing TomTom and Google Maps data, it can be said that both products are quite similar, providing very useful information for BigData4ATM.

The main benefit of TomTom over Google Maps is the possibility to get historical data in order to analyse exceptional cases as it could be the behaviour of the traffic when roads are closed due to accidents, works, pollution limitations, etc. On the other hand, the inability to calculate routes by public transport it is a great drawback when estimating the accessibility to airports.

Due to the limited number of queries that can be performed with a free account, a Google Maps Premium Plan is needed.
B.4 Madrid Regional Transport Consortium (CRTM) smart card data

B.4.1 Data overview

Smart cards offer great advantages over traditional paying methods for public transport: transactions are faster and easier, and timetables are easier to keep as drivers do not need to sell the tickets. These are the most visible benefits of the smart Card systems, but there is also a secondary advantage: the huge amount of data it generates. These data can provide information about the use of public transport, peak hours, most crowded stations/lines, disruptions... By analysing the data produced by smart cards, useful knowledge can be obtained to inform public transport decision making processes.

The Madrid Regional Transport Consortium Smart Card data contains data from the following public transport services:

- Cercanías train (Commuter train)
- Subway
- EMT buses (Madrid city buses)
- Intercity buses

The smart card data provided by Madrid Regional Transport Consortium (CRTM) contains the following information:

- Validation data: a register every time a CRTM smart card user enters public transport
- Intercity bus stops: locations of each intercity bus stop

At the moment of writing the deliverable, the geographic locations of the stops for the Cercanías, Subway and EMT buses were not available. It is expected that these data will be provided during the following months. Therefore, the analysis presented here is not complete and will be finished when the rest of the data arrives.

Smart card data were provided for three days, Tuesday 8th, Wednesday 9th and Thursday 10th March 2016.

B.4.2 Sample characteristics

By analysing the registers for the 3 days of data that were provided, the number of users is around 1,100,000 users. This number of users keeps constant between the three days of data, with less than 20,000 users of difference. It is worth noting that there are other means of accessing Madrid public transport, so there are public transport users that do not own a smart card. According to data providers, smart card users represent around 60% of the total Madrid public transport users.
When looking at the total number of validations for each day, it is also quite constant across the three days, around 3,700,000.

It is also expected to receive profile information about public transport users, as there exist different fares for different groups (young people, third age, etc.). However, at the moment of writing this deliverable, this information was not available yet.

**B.4.3 Temporal characteristics**

In this section the characteristics of the dataset regarding registers distribution during the day have been studied.

When looking at the number of validations for each hour of the day, for each one of the CRTM operators, we get the following curves:
Figure 69: Validations for 08-03-2016

Figure 70: Validations for 09-03-2016

Figure 71: Validations for 10-03-2016
As it can be seen, distributions are quite similar for all the days. As the three days correspond to working days of the same week, it is something reasonable. It is worth noting that, according to these data, there is a peak of trips at lunchtime (which was also detected with Google Maps Directions API). Also, subway is the most used public transport mode during peak hours, while city buses is the most used during valley hours.

When looking at the number of validations per user distribution (Figure 72), we can see that it is also similar across the three days of data. The most typical use of public transport corresponds to two validations per day. Also, four validations per day is more common than three validations per day (explained by recurrent trips that require transfers between different public transport modes). There are some users with a high number of validations (more than 20), which will be analysed in depth when information about the locations of the stops is available.

Another interesting temporal characteristic is the distribution of the number of users with registers during the day. To study this, the duration of the day was divided in 3h intervals, and the number of users with registers in each of these time intervals was obtained. Also, the distribution of the number of intervals for which there is user activity is presented in Figure 74. As it can be seen, most public transport users have presence only in two different 3h intervals. It is reasonable to assume that this is caused by home-work and work-home trips using public transport.
**Figure 73:** Active users per time interval

**Figure 74:** Number of intervals with activity per user
B.5 Credit card transactions data

The analysis presented here is a summary of the data characterisation done during the EUNOIA project (http://eunoia-project.eu/). However, as negotiations for the access to this kind of data are currently ongoing and it is likely that this data will be used for BigData4ATM, it was decided to include this characterisation in this document.

The geographical scope of this dataset contained was restricted to the provinces of Madrid and Barcelona and covered two years, 2011 and 2012. The dataset contained the following information:

- Transactions made by BBVA clients at any Point of Sale (POS)
- Transactions made by non-BBVA clients at any BBVA POS

A typical transaction record contains information about the amount spent, timestamp of the transaction, user anonymised identifier, POS location and business type.

The information regarding BBVA clients was a little bit richer, containing basic profile information about the client (gender, age, home zip code), while for non-BBVA clients the only information available was the country of their bank. This dataset is not available for BigData4ATM, as the data provision agreement expired with the end of EUNOIA. At the moment of writing this deliverable, an extension of the agreement, which would also grant access to more recent data, is being negotiated.

An in-depth analysis of this dataset can be found in [39].

B.5.1 Sample characteristics

Aggregated stats

In the following for both cities concerned (Madrid and Barcelona), we give the numbers for two distinct spatial delineations:

- the full province (largest spatial extension we have, data provided by BBVA);
- the metropolitan area (we filtered out the transactions in businesses that are not located in municipalities included in the metropolitan area).

<table>
<thead>
<tr>
<th>2nd March 2011</th>
<th>Barcelona (province)</th>
<th>Barcelona (metro area)</th>
<th>Madrid (province)</th>
<th>Madrid (metro area)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of transactions</td>
<td>40390</td>
<td>26015</td>
<td>89565</td>
<td>81139</td>
</tr>
<tr>
<td>% of all transactions in 2011</td>
<td>0.27 %</td>
<td>0.28%</td>
<td>0.28%</td>
<td>0.28%</td>
</tr>
<tr>
<td>Number of businesses</td>
<td>16059</td>
<td>10216</td>
<td>20403</td>
<td>18137</td>
</tr>
<tr>
<td>% of all businesses appearing in 2011</td>
<td>14 %</td>
<td>14%</td>
<td>18%</td>
<td>19%</td>
</tr>
<tr>
<td>Number of customers</td>
<td>32098</td>
<td>20749</td>
<td>71146</td>
<td>64549</td>
</tr>
<tr>
<td>% of all customers appearing in 2011</td>
<td>5.3%</td>
<td>4%</td>
<td>5%</td>
<td>4%</td>
</tr>
</tbody>
</table>
### Table 1

#### March 2, 2011

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Barcelona (province)</th>
<th>Barcelona (metro area)</th>
<th>Madrid (province)</th>
<th>Madrid (metro area)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-19</td>
<td>0.2</td>
<td>0.5</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>20-24</td>
<td>2.5</td>
<td>3.15</td>
<td>5.4</td>
<td>5.5</td>
</tr>
<tr>
<td>25-29</td>
<td>8.3</td>
<td>8.16</td>
<td>11.5</td>
<td>11</td>
</tr>
<tr>
<td>30-34</td>
<td>16</td>
<td>13.9</td>
<td>15.5</td>
<td>15</td>
</tr>
<tr>
<td>35-39</td>
<td>18.5</td>
<td>15.9</td>
<td>15.7</td>
<td>16</td>
</tr>
<tr>
<td>40-44</td>
<td>15.5</td>
<td>13.7</td>
<td>13.1</td>
<td>13</td>
</tr>
<tr>
<td>45-49</td>
<td>12</td>
<td>11.6</td>
<td>10.8</td>
<td>11</td>
</tr>
<tr>
<td>50-54</td>
<td>9.3</td>
<td>9.9</td>
<td>9.2</td>
<td>9.1</td>
</tr>
<tr>
<td>55-59</td>
<td>6.7</td>
<td>8</td>
<td>6.8</td>
<td>7</td>
</tr>
<tr>
<td>60-64</td>
<td>4.9</td>
<td>6.3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>65-69</td>
<td>3</td>
<td>4.1</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td>70-74</td>
<td>1.5</td>
<td>2.1</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>75-79</td>
<td>0.9</td>
<td>1.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>80-84</td>
<td>0.5</td>
<td>0.8</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>85-89</td>
<td>0.1</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>90-94</td>
<td>0.03</td>
<td>0.07</td>
<td>0.04</td>
<td>0.04</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Barcelona (province)</th>
<th>Barcelona (metro area)</th>
<th>Madrid (province)</th>
<th>Madrid (metro area)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>48</td>
<td>47</td>
<td>47</td>
<td>47</td>
</tr>
<tr>
<td>Women</td>
<td>52</td>
<td>53</td>
<td>53</td>
<td>53</td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>Year 2011</th>
<th>Barcelona (province)</th>
<th>Barcelona (metro area)</th>
<th>Madrid (province)</th>
<th>Madrid (metro area)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>53</td>
<td>53</td>
<td>53</td>
<td>53</td>
</tr>
<tr>
<td>Women</td>
<td>47</td>
<td>47</td>
<td>47</td>
<td>47</td>
</tr>
</tbody>
</table>
B.5.2 Temporal characteristics

The x-range in Figure 77 was limited to 0-200 so that the peaks periodicity appears more clearly. These peaks are also observable when one plots the number of transactions vs. day: they have a 1-week periodicity, as there are more card transactions during week-ends. The peak that appears at the beginning of the curve corresponds to users with transactions that are very close between them.
B.6 Twitter data

B.6.1 Data overview

IFISC has been collecting Twitter data from the public API Stream since 2014. The process is still ongoing and covers all European countries, and eventually can have a worldwide coverage for further needs. Here we focus on geolocated tweets, in particular on the data associated to geographical coordinates (latitude and longitude) of emission of the communications, where allowed by the user. Each document in the database corresponds to an anonymised tweet, and the typical (minimum) information of a record consists of the following fields:

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Type and Format</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>uid</td>
<td>integer</td>
<td>user identification number</td>
</tr>
<tr>
<td>id</td>
<td>integer</td>
<td>identification of the tweet</td>
</tr>
<tr>
<td>created_at</td>
<td>datetime object</td>
<td>year, month, day, hour, minute, second of the emission of the tweet</td>
</tr>
<tr>
<td>coordinates</td>
<td>list of floats</td>
<td>coordinates of the emission of the tweet [latitude, longitude]</td>
</tr>
<tr>
<td>text</td>
<td>string</td>
<td>content of the tweet</td>
</tr>
</tbody>
</table>

The field related to the content of the tweet may be the source of further investigation over the language detection and hometown country of the users; thanks to the users’ communication history, location and language, we are able to detect the most probable place of residence and native language for each user, getting more insights on his/her mobility patterns.

Twitter data are typically biased by the characteristics of the users: despite those issues, it has been shown, by cross-checking them with other data sources, that they are reliable for certain types of mobility analyses. In particular, Twitter data are highly correlated with official statistics when used to get information about origin-destination flows in urban areas.

B.6.2 Sample characteristics

We selected a subsample of the data to explore and show the characteristics of the data with different temporal horizons. We have the following number of tweets and users in our dataset:

<table>
<thead>
<tr>
<th>Period</th>
<th>Number of Tweets</th>
<th>Number of Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>76,007,853</td>
<td>3,746,083</td>
</tr>
<tr>
<td>03/2016</td>
<td>6,295,723</td>
<td>734,823</td>
</tr>
<tr>
<td>02/03/2016</td>
<td>224,883</td>
<td>100,516</td>
</tr>
</tbody>
</table>

In order to get a representative sample of geolocated data, we evaluate the penetration rate of Twitter geolocated messages. This is a measure of how much the population of a country is inclined to emit geolocated information through social media. Starting from the original dataset of tweets taken over Europe (2014 to 2016), we first detect the country of emission for each tweet; then for
each user we assign a country of residence according to the country where they tweet more. We finally define the penetration rate as the ratio between the number of Twitter users that share geolocated data over the number of inhabitants of each country.

B.6.3 Temporal characteristics

Here we present some temporal characteristics of the Twitter dataset covering the European region per year (full 2015), per month (March 2016) and per day (2\textsuperscript{nd} March 2016)

Yearly activity
Figure 81: Average Twitter activity on working days
Figure 82: Average Twitter activity on weekends

Figure 83: Average user activity on working days
Figure 84: Average user activity on weekends

Figure 85: Distribution of average inter-event times
Figure 86: Distribution of user activity over days
Figure 87: Distribution of average number of tweets/day
Figure 88: Monthly Twitter coverage

Figure 89: Average Twitter activity on working days

Figure 90: Average Twitter activity on weekends

Figure 91: Average user activity on working days

Figure 92: Average user activity on weekends
Figure 93: Distribution of average inter-event times
Figure 94: Distribution of user activity over days
Figure 95: Distribution of user activity over days

Daily activity

Figure 96: Daily Twitter coverage

Figure 97: Twitter activity per hour of day
Figure 98: Twitter users per hour of day
Figure 99: Inter-event times

Figure 100: Number of tweets per day
B.7 FlightRadar24

B.7.1 Data inputs

IFISC has been collecting data from the Flightradar24 website since early 2015. Data can be collected freely from the website, but explicit permission (which has been granted to IFISC) is required in order to use it. Each document in the database corresponds to a flight, where the typical document contains the following fields:

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Type and Format</th>
<th>Note(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>regnumber</td>
<td>string</td>
<td>Tail number of the aircraft operating the flight</td>
</tr>
<tr>
<td>flight</td>
<td>string or list of strings</td>
<td>Flight number (may be more than one)</td>
</tr>
<tr>
<td>airline</td>
<td>string (2 characters)</td>
<td>Operating airline (IATA code)</td>
</tr>
<tr>
<td>iata_to</td>
<td>string (3 characters)</td>
<td>Origin airport (IATA code)</td>
</tr>
<tr>
<td>iata_from</td>
<td>string (3 characters)</td>
<td>Destination airport (IATA code)</td>
</tr>
<tr>
<td>scheduled_departure</td>
<td>UNIX timestamp</td>
<td>Scheduled Off-Block Time (SOBT)</td>
</tr>
<tr>
<td>scheduled_arrival</td>
<td>UNIX timestamp</td>
<td>Scheduled In-Block Time (SIBT)</td>
</tr>
<tr>
<td>real_departure</td>
<td>UNIX timestamp</td>
<td>Actual Take-Off Time (ATOT)</td>
</tr>
<tr>
<td>real_arrival</td>
<td>UNIX timestamp</td>
<td>Actual Landing Time (ALDT)</td>
</tr>
<tr>
<td>collected_at</td>
<td>string</td>
<td>Date and time, formatted as yyyy-mm-dd HH:MM:SS</td>
</tr>
</tbody>
</table>

Several things should be noted:

- Subtracting the scheduled departure from the real one yields the flight’s departure delay plus the taxi-out time. Likewise, subtracting scheduled arrival from real arrival gives arrival delay plus taxi-in time. Delays and taxi times cannot therefore be separated, unless one of the two is known from a different dataset. Flightradar24 classifies a flight as delayed when taxi time plus delay exceeds fifteen minutes. Taxi times are available for US flights from the website of the Bureau of Transportation Statistics, while for Europe they are not publicly available, but are provided upon request by EUROCONTROL’s Central Office for Delay Analysis.
- Documents which do not possess a valid value for each of the above-mentioned fields are discarded and not considered in the analysis. The number of “malformed” documents found in each day is usually zero or negligible in the US and Europe.
- Hereafter we will use the terms Europe and ECAC (European Civil Aviation Conference) interchangeably, unless otherwise specified. The only states which are part of ECAC but neither of EUROCONTROL nor the European Union are Azerbaijan, Iceland and San Marino.
B.7.2 Data quality and coverage

Daily activity

Data quality issues may arise in two different ways:

1. Database documents with incomplete or manifestly incorrect information, e.g. missing tail number or destination airport, flight length incompatible with aircraft type, etc.
2. Consistency issues between related flights.

Problems of the first type affect in a very serious way the coverage outside ECAC and US, but are negligible inside these two areas, and therefore are outside of the scope of this document. The problems of the second type do not affect any of the analyses presented here, but may become relevant depending on the way the data are used.

Flight entries can be grouped by tail number and then each group can be sorted chronologically by scheduled departure, scheduled arrival, real departure or real arrival. If the information regarding the aircraft rotation is entirely consistent, the origin airport of each leg in the rotation should always coincide with the destination of the previous leg. It should be noted that the flights are collected by going through each airport several times per day and looking at the flights which have already departed or landed, so diversions and cancellations are not going to cause problems in this sense.

In some cases, the above conditions are not satisfied. This can be due to several causes:

1. Two aircraft at the same airport (both arriving/departing, or possibly one arriving and one departing) are mistakenly identified with the same tail number. This can result in a variety of inconsistencies, such as a flight terminating after the previous leg in the rotation has started.
2. An aircraft moves from a zone with good coverage to a zone without coverage and re-emerges later. In this case, the rotation will have one or more "holes" in it.
3. The aircraft schedule may itself be malformed, with one leg scheduled to depart before the previous one is completed. According to what we have discussed in private communications with industry experts, this is something that actually takes place.

These problems can be solved with different strategies depending on what is needed. If mapping each flight to a specific aircraft is not needed (for example, when calculating the total traffic at one airport during a period of time), no action needs to be taken. The simplest measure is to just “split” rotations at every error, i.e. all the flights after a “hole” can be considered as operated by a different aircraft. Another option is to discard the entire rotation as soon as an error is found, but it is likely excessive, as it would result in around 10% of the flights being discarded.

In the following, the analysis is performed over all flights with complete individual information, regardless of their rotation-level consistency, with at least one of the origin and destination airports located in the ECAC area.
Coverage

The figure above shows the amount of valid flight entries in the ECAC area for each day from the start of data collection until 25th September 2016. Gaps and periods with abnormally low numbers of entries are due to technical problems either on the IFISC or the Flightradar24 sides. For comparison, the average number of flights per day in Europe (excluding non-EU countries) during 2015 is 20,776 according to Eurostat, and according to EUROCONTROL, the number of IFR flights per day in the ECAC area ranges from around 22,000 during January 2015 to above 30,000 during the summer.

B.7.3 Temporal characteristics

Here we show several of the temporal features of the data. The following results were computed over the whole dataset.
B.7.4 Spatial characteristics

Here we show several of the spatial features of the data. As in the previous section, the results shown are computed over the whole dataset.

The radius of each point is proportional to the total number of departures in the sample. The top 10 are explicitly named and ranked on the right side.
Figure 105: Histogram of airports ranked by total departures

Figure 106: Distribution of airports by total departures
B.8 Public transportation timetables

In order to analyse and model door-to-door mobility patterns and travel times it is necessary to acquire information on the ground transportation network. Fortunately, in recent years this kind of information has become publicly available thanks to numerous national and municipal open data projects. We list here a few major examples.

B.8.1 Great Britain

Data available for download: http://datadryad.org/resource/doi:10.5061/dryad.pc8m3
Data descriptor: http://www.nature.com/articles/sdata201456

Figure 107: Multilayer network at national scale

Figure 108: Multilayer network at urban scale
B.8.2 USA

The Bureau of Transportation Statistics recently published their new National Transit Map [http://www.rita.dot.gov/bts/ntm/map](http://www.rita.dot.gov/bts/ntm/map). The initial National Transit Map consists of General Transit Feed Specification (GTFS) data feeds [https://developers.google.com/transit/gtfs/](https://developers.google.com/transit/gtfs/). Data from 270 transit agencies provided information on over 398,000 stops and stations and almost 10,000 routes. Development of the National Transit Map is a continuing process and another update is expected to be released by the end of 2016.

![Figure 109: USA participating agencies](image1.png)

![Figure 110: Local example](image2.png)
B.8.3 Further urban feeds (New York City, Barcelona, Paris…) 

Similar data are continuously provided at urban level by several transportation agencies. IFISC has direct experience working with the data of cities like New York, Barcelona and Paris, but more data, in most cases provided in the standard GTFS format, are aggregated in registries such as:

- [https://transit.land/feed-registry/](https://transit.land/feed-registry/)
- [http://transitfeeds.com/feeds](http://transitfeeds.com/feeds)

![Figure 111: Shortest paths in NYC](image1)

![Figure 112: Barcelona public transport network](image2)