BigData4ATM

Passenger-centric Big Data Sources for Socioeconomic and Behavioural Research in ATM

Big Data Analytics for Socioeconomic and Behavioural Research in ATM

State-of-the-art and Future Challenges

May 2016

Contents

Executive summary

A sound understanding of the behavioural and societal factors that influence transport demand and supply — including economic, social, demographic and cultural issues — is essential for shaping transport policies according to the values, needs and expectations of our society. However, research on the relationship between passengers' behaviour and air traffic management (ATM) performance is relatively scarce. There is a lack of understanding of the impact of passengers' behaviour on ATM, as well as of the impact of ATM on individual passengers and society at large. Research in these areas has so far been constrained by the limited availability of behavioural data, typically obtained from cross-sectional (static) demographic and economic datasets, often consisting of very small samples, and usually complemented with assumptions about the permanence of behavioural traits over time. The pervasive penetration of modern information and communication technologies (ICT) and the emergence of big data analytics open new opportunities to overcome this situation: for the first time, thanks to the proliferation of smart personal devices and interconnected services, we have large-scale, detailed longitudinal (dynamic) data allowing us to test hypotheses about travellers' behaviour.

The overall 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 ICT-based spatio-temporal data — including mobile phone records, data from indoor geolocation technologies, credit card records and data from Internet social networks, among others — with the aim of characterise 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.

1. Introduction

The paramount goal of the European transport policy, as defined in the European Commission's 2011 White Paper on Transport (European Commission, 2011a), is to "establish a system that underpins European economic progress, enhances competitiveness and offers high quality mobility services while using resources more efficiently". Particular emphasis is put on the need for a multimodal, passenger-centric transport system, able to provide seamless door-to-door travel and facilitate better modal choices. In line with these objectives, the long-term vision for the European aviation sector outlined in the report 'Flightpath 2050 - Europe's Vision for Aviation' (European Commission, 2011b) identifies five challenges that aviation will have to face at the 2050 horizon: Meeting Societal and Market Needs (Challenge 1); Maintaining and Extending Industrial Leadership (Challenge 2); Protecting the Environment and the Energy Supply (Challenge 3); Ensuring Safety and Security (Challenge 4); and Prioritising Research, Testing Capabilities and Education (Challenge 5). The report envisages a passenger-centric air transport system thoroughly integrated with other transport modes, with the ultimate goal of taking travellers and their baggage from door to door predictably and efficiently while enhancing passenger experience and rendering the transport system more resilient against disruptive events. As part of Challenge 1, the Flightpath 2050 report highlights the role of aviation as an enabler of socioeconomic development and stresses the importance of customer orientation, drawing attention to the key role of the ATM system in realising this vision and defining five high-level goals:

1. European citizens are able to make informed mobility choices and have affordable access to one another, taking into account: economy, speed, and tailored level of service.
2. 90% of travellers within Europe are able to complete their journey, door-to-door within 4 hours.
3. Flights arrive within 1 minute of the planned arrival time regardless of weather conditions. The transport system is resilient against disruptive events and is capable of automatically and dynamically reconfiguring the journey within the network to meet the needs of the traveller if disruption occurs.
4. An air traffic management system is in place that provides a range of services to handle at least 25 million flights a year of all types of vehicles.
5. A coherent ground infrastructure is developed including: airports, vertiports and heliports with the relevant servicing and connecting facilities, also to other modes.

In contrast with this high-level vision, ATM operations have so far lacked a passenger-oriented perspective, with performance objectives and decision criteria (e.g., flight prioritisation rules) not necessarily taking into account the ultimate consequences for the passenger (Cook et al., 2013a). Further research is needed to provide new insights on the interactions between the ATM system and passengers' needs, choices and behaviour. However, current methods used to collect data on passengers' activities are limited in accuracy and validity: traditional methods based on observations and surveys present intrinsic limitations (e.g., incorrect and imprecise answers, dependence on the availability and willingness to answer of the interviewed persons, etc.), and they are also expensive and time-consuming; useful data can also be collected from other sources such as air traffic databases, travel reservation systems or market intelligence data services (e.g., IATA Passenger Intelligence Services, PaxIS), but these data typically fail to capture important information, such as door-to-door origin-destination pairs and travel times. The generalised use of geolocated devices in our daily activities opens new opportunities to collect rich data and overcome many of the limitations of traditional methods. The very same ICT tools that are enabling new forms of bidirectional communication with the passenger are also making it possible to gather permanently updated information on passengers' activity and mobility patterns with an unprecedented level of detail.

2. New big data sources for the analysis of travel behaviour

The generalisation of the use of portable communication devices has brought new opportunities for collecting data on human behaviour, opening the door to a systematic statistical treatment of many social aspects (Lazer et al., 2009). Some examples include the analysis of the structure of social networks (Liben-Nowell et al., 2005; Onnela et al., 2007; Grabowicz et al., 2012), human cognitive limitations (Gonçalves et al., 2011), information diffusion and social contagion (Crane et al., 2008; Lehmann et al., 2012), the role of social groups (Onnela et al., 2009; Grabowicz et al., 2012), the interaction between social relations and mobility (Grabowicz et al., 2014; Picornell et al., 2015), language coexistence (Mocanu et al., 2013) or political movements (Borge-Holthoefer et al., 2011; González-Bailón et al., 2011; Connover et al., 2013).

The analysis of human mobility is one of the questions to which the wealth of new data has notably contributed (Brockmann et al., 2006; Gonzalez et al., 2008; Balcan et al., 2009; Noulas et al., 2012; Bagrow et al., 2012; Lenormand et al., 2014; Tizzoni et al., 2014; Jurdak et al., 2015). Statistical characteristics of mobility patterns have been studied (Brockmann et al., 2006; Gonzalez et al., 2008), finding a heavy-tail decay in the distribution of displacement lengths across users. Most of the trips are short in everyday mobility, but some are extraordinarily long. Besides, trips are not directed symmetrically in space, but each user shows a particular structure (Gonzalez et al., 2008). The duration of stay in each location also shows a skewed distribution with a few preferred places clearly ranking on the top of the list, typically corresponding to home and work (Song et al., 2010). Recently, geolocated data have also been used to analyse the structure of urban areas (Lenormand et al., 2014; Louail et al., 2014; Louail et al., 2015), road and long range train traffic (Lenormand et al., 2014b), land use (Soto et al., 2011; Frías-Martinez et al., 2012; Lenormand et al., 2015) and mobility between cities (Lenormand et al., 2015b) or countries (Hawelka et al., 2014). The new ICT technologies are permitting the analysis of human mobility discerning different user profiles in terms of age, socioeconomic level and place of residence (Lenormand et al. 2015b, Lenormand et al. 2015c). It has been shown that it is possible to track movements across the full continent and to extract origin-destination matrices at an unprecedented scale and at a very low cost (Lenormand et al., 2014; Lenormand et al., 2015b). In addition, the limitations of the different data sources have been explored, finding that at large scales (larger than 2 km), Twitter and cell phone data provide similar results (Lenormand et al., 2014). All these results evidence the potential of ICT data, either from online social media, cell phone or credit card records, to characterise door-to-door mobility, from initial origin to final destination. These data sources provide also access to public content, for instance in Twitter messages. The combination of content/semantic analysis and geolocation is a powerful resource still to be explored in a systematic way.

The potential of these new data sources for the transport sector is huge, but it also comes with a number of challenges. We have more data, but not always with explanatory power about the underlying decisions of individuals. We also have a much higher coverage than with traditional surveys, but this sample size often comes at the expense of low quality, noisy or biased data. The ability to mine, blend and analyse data from multiple sources will be of paramount importance to overcome these issues and offer reliable and comprehensive information.

3. Socioeconomic and behavioural research in ATM: challenges and opportunities

The profound transformation that the current European ATM system is undergoing goes beyond technological considerations. With society being one of the stakeholders of the ATM Performance Partnership (SESAR, 2007; EUROCONTROL, 2010a), users' requirements need to be better accommodated to enable the execution of flights as close as possible to the passengers' intention. Planning for and better adapting the ATM system to the future requires systematic socioeconomic monitoring to understand the factors influencing changes in demand and flights patterns (EUROCONTROL, 2009b).

A variety of studies can be found on factors influencing future air traffic on the economic dimension. However, the societal dimension, essential in the paradigm shift from a flight-centric to a passenger-centric air transport system, has been little explored. Some studies have looked at the perception of air transport and ATM by the European society and their expectations about it, by means of the compilation of existing surveys (Kinchin, 2004), the analysis of print media mentions of ATM performance (Mahaud and Arrighi de Casanova, 2004), and the use of questionnaires and discussion groups (Cook and Tanner, 2005), revealing a general unawareness of ATM performance and a lack of proper coverage of ATM issues in the press. Ignorance of the frontiers and areas of action of the different ATM sub-systems is also reflected. Worries and perceptions vary from country to country, and it is suggested that attitudinal responses are strongly driven by recent events in a given location (Cook and Tanner, 2005). These studies are however restricted on the number of people and countries reached and cover only one of the many social aspects affecting air traffic. A comprehensive assessment of the impact of social and behavioural factors (behavioural attitudes, mobility needs, tolerance to waiting times, transport modes preferences, etc.), as well as of the net impact that SESAR and other ongoing ATM modernisation will have on the passenger, remains to be done. In the following sections we dig deeper into four application areas addressed by BigData4ATM: (i) passenger-centric performance metrics, (ii) air traffic forecasts, (iii) airport landside and airside processes integrated optimisation, and (iv) ATM disruption impact assessment.

3.1 Passenger-centric performance metrics

In addition to monitoring and reporting on the performance of the ATM network in terms of delays, the Network Manager, through the Central Office for Delay Analysis (CODA), provides a monitoring and analysis function for all delay reasons (ATFM, airline, airport, etc.). This enables correlation between airline and network reported delays, and is used in schedule and turnaround planning, enabling better punctuality.

The SESAR WP-E project POEM (Cook et al., 2012, 2013a, 2013b) has proposed the use of new metrics to measure the net effect of flight delays on the passenger full trip. The project has shown the importance of using passenger-centric (rather than flight-centric) metrics for a comprehensive assessment of the ATM system performance and of the effectiveness of policies aimed to improve passenger rights. Passenger-centric metrics are built on the knowledge of passengers' itineraries, which are not easy to get: data sources are sometimes incomplete, not covering data from markets such as low-fare and charter carriers, and usually restrict itineraries to the air segment (i.e., they do not contain door-to-door information).

The use of passenger-centric geolocation data opens several opportunities to extend the work started by POEM by considering door-to-door trips and incorporating the study of the economic consequences of such

delays (e.g., their impact on future travel decisions). These opportunities include: (i) reconstructing passenger itineraries beyond the air segment, by building origin-destination matrices from mobile phone data and geolocation apps; (ii) learning about the economic impact of flights delays on passenger expenditure patterns, measured from credit card data; and (iii) understanding users' motivation to travel again, change transport mode, etc., and acquiring a better knowledge of passengers' perception of ATM performance through opinion mining and sentiment analysis. Additionally, these new data sources provide the capability to capture in almost real time passengers' behaviour in the presence of special events such as strikes, flight cancellations, news of political instability in a destination, security policies, etc.

3.2 Air traffic forecasts

Air traffic forecasts are an essential input for ATM stakeholders in order to plan response to future air traffic needs. The EUROCONTROL's Statistics and Forecast Service (STATFOR) provides forecasts that are used as direct inputs into the Network Strategy Plan, the Network Operations Plan and the Network Performance Plans. These forecasts are also a pre-requisite for the establishment of the unit rates used to calculate the route and terminal charges. Traffic forecasts are also used by an extensive number of planning departments of airlines, ANSPs, airports, government authorities, etc. for general planning.

Air traffic forecasts are derived in a range of ways depending on the time and data available and also on the questions that the forecasts are intended to address. For the purpose of technology renewal, airport design or new routes assignment, a medium- to long-term perspective is needed, while for contingency actions, short-term forecast is more relevant.

The most common forecasting models are those based on time series analysis. They rely on the assumption that patterns observed historically will continue over time and use trend extrapolation and decomposition techniques, without seeking to understand the causes behind the observed patterns. Time series models are not able to explain what might be expected in the future in response to policy changes or alternative future scenarios, but they are able to respond quickly to changes in demand and can be used as benchmarks against which more sophisticated models can be compared (Airports Commission, 2013).

Causal methods look for cause-effect relationships between explanatory variables and demand patterns. Explanatory variables are passenger socioeconomic characteristics, GDP, weather conditions, population, travel costs, consumer spending, etc. In this group of models we find (i) regression models; (ii) gravity models, which estimate travel flows between two points (e.g., city pairs, country pairs, etc.) as being proportional to generation/attraction variables (e.g., the population) and inversely proportional to a generalised distance or cost between origin and destination. The population considered can be a subset of the total population, as these models are not behavioural and do not distinguish between different types of travellers; and (iii) qualitative models (e.g., Delphi method), which rely on the judgement of experts. They can be used to predict a significant change in historical patterns or in cases where there is limited or no data available. They are normally used for the assessment of how new technological or other developments would affect the forecast.

The effect of passengers' profiles and their relationship with the origin and destination locations are not fully considered by the aforementioned models. The same origin-destination pairs can be connected by different routes and by a combination of modes. Other travel modes can benefit certain routes, but at the same time they can act as competitors in the case of short distance trips. Passenger profile not only affects route, mode and company election, but also how they are modified by changes in other variables such as price, waiting

time, etc. Elasticity of air travel demand depends on travel nature: short-haul routes and leisure trips are more sensitive to travel price than long-haul routes or business travels, while business trips are more sensitive to GDP changes than leisure trips (IATA, 2008; Airports Commission, 2013).

Different models have been developed trying to consider these effects in a consistent manner, in an attempt to predict demand under technological and socioeconomic changes (ICAO, 2006; Airports Commission, 2013; OECD ITF, 2014). However, these models require large amounts of specialised data, and often need to be simplified due to the lack of quality data (OECD ITF, 2014). Historically used data sources such as MIDT / BSP and PaxIS are insufficient or difficult to get due to legal issues. Additionally, some aspects of passengers profile, real origin-destination pairs and travel purposes are not reflected by such data, so they have to be obtained from surveys that are expensive, time consuming, and provide small samples. The large amount of data produced by the extended use of geolocated devices offers new opportunities to replace and/or complement these data and get new behavioural insights.

3.3 Integrated optimisation of airport landside and airside processes

Passengers' behaviour at the terminal and landside has a major impact on flights delays (EUROCONTROL, 2009a). Many companies and airports are analysing and testing different options to reduce bottlenecks and waiting times at the check-in, security, passport control and gate queues. The modelling of passenger movement at the airport terminal has traditional made use of aggregated approaches, where the behaviour of single entities (agents) is represented by aggregated flows. However, aggregated models face important limitations. The fundamental weakness is that they are not based on a coherent theory of passenger behaviour, and therefore are not suitable to capture agents' responses to measures influencing their behaviour. In an attempt to address these shortcomings, agent-based microsimulation models have been developed considering agents' intentions, environmental perception and individual interactions. Schultz and Fricke (2011) have proposed an individual-based movement model based on a stochastic approach, including system knowledge about characteristics of handling processes, infrastructure knowledge (navigation, orientation) and perception and processing of provided information (signage). Other analyses have additionally incorporated passenger group dynamics (Cheng, 2014), showing that group behaviour has significant influence on the performance and utilisation of airport services. Different companies are also developing multi-agent airport terminal simulators, such as CAST Terminal, developed by the Airport Research Center. CAST Terminal offers a three-dimensional environment combined with an agent-based model of passenger movement behaviour to identify bottlenecks and solutions to complex operational problems and planning tasks. The relevance of the subject is proven by the companies involved in the simulator development, including BAA, Frankfurt Airport, Zurich Airport, EUROCONTROL and Airbus. However, the practical application of agent-based models is still hindered by the need for detailed data to specify, calibrate, and validate the models. The collection of the data produced by the extended use of geolocated mobile devices opens new opportunities to overcome this problem. Thanks to indoor position systems (iBeacons, Wi-Fi-based and Bluetooth-based positioning systems, etc.), we are now able to track passenger movements in a more detailed manner. The challenge here is to develop appropriate methods for merging and analysing these data sources so as to extract consistent and reliable information on passenger movement patterns, and build new models with a stronger behavioural basis allowing the integrated optimisation of airport landside and airside processes.

3.4 Socioeconomic impact of ATM disruptions

In recent years, there have been numerous examples of severe storms and other natural events that have caused widespread disruption to air transport, leaving passengers stranded in airports and cities all over the world (ICAO, 2013). Flight disruptions are defined as situations where a scheduled flight is cancelled, or delayed for two hours or more, within 48 hours of the original scheduled departure time (European Union, 2004). Disruptions impose capacity reductions and are measured in terms of traffic impact and delay impact (EUROCONTROL, 2015). In the worst cases, the disruption extends well beyond the immediate geographical region, causing a knock-on effect to passengers travelling to other regions.

The recent volcanic ash cloud crises have shown that a network approach is crucial to minimise the impact of disruptions, and that a better understanding of disruptions and their effects is essential for the effective design and implementation of a comprehensive set of measures at network level. The ash cloud crisis in 2010 triggered numerous analyses related to the impact of disruptions. More than 100,000 flights and 10 million passengers were affected and 5,000 additional flights took place to reposition aircraft and crews and to repatriate passengers. EUROCONTROL analysed how the 46% of cancelled flights and 43% of delayed flights impacted on the future air traffic forecast due to the need to adjust historical (April and May 2010) data (EUROCONTROL, 2010c). Other studies addressed the economic impact of the ash cloud consequences. IATA focused on airlines losses, concluding that the closures of the European airspace caused airline losses of around US$400 million per day from scheduled services (IATA, 2010). These numbers are the most conservative estimation based on the verifiable data available on scheduled services. Starting from these data, other analyses were conducted aiming to measure not only the economic impact of the disruption on the airlines but also on alternative modes of transport, like car rentals and railways. As an example, Eurostar reported that it carried 50,000 extra passengers on 15 April, and registered an increase of 33% on 17 April. P&O Ferries of France declared that their services between Britain, Spain, France and the Netherlands were fully booked and that they had to employ extra personnel.

One reasonable hypothesis is that there are differences in passengers' behaviour if the disruption can be known in advance or not. However, the conventional analyses do not reveal such differences. For example, the analyses of the Air France-KLM strike in September 2014 report a €400 million cost for the airlines, 8,500 flight cancellations and 1 million passengers affected, but there is little information about ticket changes or other means of transport (Anna Aero, 2014; The New York Times, 2014).

Finally, disruptions have other consequences well beyond the direct economic impact for airlines (Rose, 2014): passengers might miss important business or personal events, and other economic sectors (e.g., tourism) can be severely damaged. So far, the attempt to measure these indirect costs has had a moderate success. Here again, the large amount of data generated by smart devices offers new opportunities to study passengers' behaviour in situations of disruption and measure how mobility patterns, travel times and expenditure flows are modified.

4. The BigData4ATM project

BigData4ATM (www.bigdata4atm-sesar.eu) is a research project within SESAR Exploratory Research which investigates 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 how this information can be used to inform ATM decision making processes.

4.1 Project objectives

BigData4ATM pursues the following objectives:

1. to develop a set of methodologies and algorithms to acquire, integrate and analyse multiple distributed sources of non-conventional ICT-based spatio-temporal data — including mobile phone records, data from geolocation technologies, credit card records and data from Internet social networks, among others — with the aim of eliciting 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.

4.2 Approach

With the emergence of big data, some voices have raised concerns about the risk of focusing on descriptive work and predictive, non-explanatory models, abandoning theory (Boyd, 2010; Graham, 2012). BigData4ATM will adopt an integrative approach between empirical data analysis and theoretical modelling, with the aim to take advantage of the opportunities offered by big data for the formulation, calibration, and testing of new models of passengers' behaviour and their interaction with the European ATM system.

The proposed research strategy comprises three main stages:

- Data collection and assessment
- Big data analysis and modelling
- Case studies

The first stage will deal with the collection of the different datasets required for the execution of the project and the analysis of their strengths and limitations. This analysis will help refine the questions that will be addressed during the second stage, aimed at extracting relevant behavioural information from the different passenger-centric sources of georeferenced data. Finally, the third stage will evaluate the potential of these new data sources to improve the quality of decision making in ATM in several specific applications.

Data collection and assessment

Data collection

BigData4ATM has as one of its cornerstones the integration of massive repositories of heterogeneous data, with special focus on new spatio-temporal data coming from smart technologies, but also including more traditional data sources for calibration or validation purposes. The datasets that will be analysed include:

- Twitter geolocated data collected using the Twitter streaming APIs. Twitter geolocated data contain information about the location of the mobile device (latitude, longitude) each time a tweet is sent, as long as geolocation is enabled. Potential uses of these data are the reconstruction of door-to-door origin-destination matrices and passenger connectivity, and the semantic analysis of tweets to analyse information propagation (e.g., efficiency of information broadcast about ATM disruptions) and social perception of ATM;
- anonymised mobile phone records, available through private commercial agreements with mobile network operators. Anonymised mobile phone records (call detail records, CDRs) provide location information about millions of users with high temporal and spatial resolution. CDRs are 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 base transceiver station (BTS) tower 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. Mobile phone records provide higher temporal resolution than mobile apps, as well as bigger and presumably less biased samples, allowing the obtention of more accurate origin-destination matrices and the exploration of problems like the connection of air transport with other modes. On the other hand, while data from Twitter can be used almost worldwide, the usefulness of the data provided by a mobile network operator is in most cases confined to the national level;
- geolocation information from iBeacons and Wi-Fi hotspots in airports, which can provide valuable information about passengers' behaviour within the airport. These sensors detect mobile devices in their surroundings, providing an estimation of the device location with a high level of accuracy; and
- anonymised electronic transactions (credit and debit card usage records), which will be used to analyse passenger expenditure patterns. Each transaction contains information about the amount spent, business type (e.g., supermarket) and location (latitude, longitude), as well as some sociodemographic information (e.g., age, gender, etc.) of the clients.

These new data sources will be complemented when necessary with more conventional data sources, which will serve a twofold purpose: (i) they will be useful for calibrating and validating the models built on the new geolocation data sources; (ii) they will be blended with the new geolocation data to enrich the information provided by each dataset separately. The data that will be used include:

- public databases on sociodemographic data, data on economic activities, etc. available upon request or accessible through open data initiatives;
- air transport and ATM databases. Global origin-destination and passenger connectivity data, schedule information, etc. will be obtained from aviation passenger intelligence solutions provided by companies like Sabre and OAG. Data on air traffic delay are expected to be obtained from EUROCONTROL's Central Office for Delay Analysis (CODA).

Data quality assessment

The collected datasets will be assessed on validity, integrity, quality, and spatial and temporal resolution, which will allow us to define in more detail the scope of the analyses that can be performed on each dataset (or each combination of datasets), as well as the associated limitations, leading to the final set of research questions to be tackled in the data analysis and modelling stage.

Big data analysis and modelling

The previous datasets will be mined and analysed to extract relevant and actionable information. The goal is to understand how traditional data sources (e.g., air traffic data) can be merged with new ICT-based geolocation data to develop an enhanced understanding of passenger behavioural drivers and ATM socioeconomic impact. The data analysis process will typically involve the following steps:

1. pre-processing and data cleaning, to clean up errors and prepare the data structures;
2. study of the representativeness of the new digital data sources when used as a passengers' survey sample and correction of the sample to minimise bias;
3. knowledge extraction. To extract knowledge from the data, we will make use of a variety of techniques including exploratory data analysis and visual analytics, spatial statistics and machine learning methods. Classical approaches aimed at the analysis of spatial data from a static viewpoint will be combined with techniques for the analysis of the temporal dynamics and the development of explanatory and predictive models;
4. extrapolation to the total population.

The research questions that will be addressed can be classified in four main groups: (i) door-to-door mobility analysis, (ii) intra-airport passenger movement analysis, (iii) expenditure analysis and (iv) opinion and sentiment analysis.

Door-to-door mobility analysis

The purpose here is to explore the potential of geolocated data from smart devices to capture passengers' door-to-door mobility, from their real origin (e.g., their residence areas) to their final destination. Modal origin-destination matrices and travel times in normal and abnormal conditions will be inferred from mobile phone records within single countries and from Twitter for the whole continent, building on recent work in this area (Andrienko et al., 2013; Lenormand et al., 2014; Picornell et al., 2015). The quality of the origin-destination matrices will be assessed by comparing them with data from other sources, such as travel surveys. Finally, we will develop different types of models of daily people flows and network evolution (including spatial interaction models, reaction-diffusion models and models of complex evolving networks) aimed to reproduce the observed patterns.

Intra-airport passenger movement analysis

Geolocation data from iBeacons and Wi-Fi hotspots in airports will be analysed to better characterise passengers' behaviour inside the airport and measure the time spent in the check-in desks, security control, passport control and boarding queues. We will explore the potential of this information to predict bottlenecks and to inform the formulation, calibration and validation of reliable models of passenger movement and activity such those proposed by Schultz and Fricke (2011).

Expenditure analysis

The credit card dataset will be used to extract expenditure patterns within and outside airports, here again building on recent work in the field (Lenormand, 2015a). We will explore the impact of different events on expenditure flows, and will propose different types of spatial interaction models, such as Boltzmann, Lotka and Volterra models (Wilson, 2010), trying to reproduce the coupling between short-term dynamics (daily flows) and the structural evolution of the expenditure flows.

Opinion and sentiment analysis

Natural language processing, text analysis and computational linguistics will be used to analyse the content of the public messages in Twitter (and possibly other social networks) to identify and extract subjective information about air transport and ATM. Keyword spotting, lexical affinity and machine learning techniques will be used to discriminate positive from negative opinions and tag emotional states. We will first explore the public perception about ATM stakeholders in normal conditions, and will then analyse how it is modified in the presence of disruptions (e.g., a strike), in search for the positive or negative impression of the public towards different agents. The temporal evolution of these aftershocks will be analysed to explore whether public attention vanishes in time as with other regular pieces of information (e.g., memes) or shows any particularities such as a longer recovery processes or attention spans.

Case studies

The purpose of the case studies is to explore, evaluate and illustrate the potential of the insights on passenger behaviour acquired from the new data sources to feed ATM decision-making processes. We have identified four case studies where the bidirectional nature of the cause-effect relationships between passengers' behaviour and ATM performance is considered particularly relevant: (i) passenger-centric door-to-door delay metrics, (ii) air traffic forecasting; (iii) integrated optimisation of airport landside and airside processes, and (iv) socio-economic impact of ATM disruptions. Common to the four case studies is the idea of exploring the opportunities offered by new passively collected data to find out what happened on any particular day without having to plan in advance the collection of data for that day. This will allow us to learn from "spontaneous experiments" and better understand how passenger behaviour changes as a result of disruptive events, such as ash clouds or strikes, and how such behavioural changes feed back into the ATM system and influence its performance. The specific questions to be addressed in each case study are defined with a certain degree of flexibility, so that they can be refined by taking into account the results of the data analysis work, as well as the inputs provided by relevant stakeholders.

Case study 1: passenger-centric door-to-door delay metrics

Case study 1 will focus on the estimation of passenger-centric door-to-door delay indicators and the evaluation of ATM contribution to such delay. The SESAR WP-E POEM project has shown the importance of passenger-centric metrics in fully assessing ATM system performance, and has quantified passenger value of time as a function of delay at the final destination (Cook at al., 2013b). We will investigate how to extend this work beyond the air segment of the trip, by combining information on air transport delays with the door-to-door origin-destination pairs and travel times obtained from the new geolocation data, with the purpose of providing new insights into the contribution of ATM to the European goal of at least 90% of travellers within Europe able to complete a door-to-door journey within 4 hours.

Case study 2: air traffic forecasting

In this case study, new demand forecasting models will be developed. Traffic forecasts and peak-period parameters are important for ATM systems planners to anticipate where and when congestion will occur and plan the implementation of CNS/ATM systems (ICAO, 2006). Currently, air traffic forecasts typically combine flight statistics with econometric models that relate air traffic to observed variables whose correlation with the traffic values is plausible, such as demographic, macroeconomic and tourism variables. Our hypothesis here is that the information extracted from the new data sources can bring several improvements in traffic forecasts, including:

- larger, cheaper and permanently updated behavioural data samples for model validation and calibration;
- multimodal information for door-to-door itineraries, which will help study competition and synergies between transport modes and quantify the impact of airport access/egress times on demand flows, enabling decision makers to understand the value of improved road or rail infrastructure;
- identification of new behavioural variables. Mobility and expenditure patterns, as well as other sociodemographic information inferred from the new ICT-based geolocated data sources, will be used for travellers' classification, identifying differences in sensitivity to economic and service changes and route preferences. The newly developed door-to-door delay indicators will allow us to investigate the relationship between demand and travel time reliability. Finally, we will explore the possibility of using opinion and sentiment analysis to analyse the long-term cumulative effects of delays on demand.

Case study 3: integrated optimisation of airport landside and airside processes

Central to the vision for the European aviation sector by 2050 formulated in the Flightpath 2050 report are time efficiency — meaning reduced door-to-door journey times, seamless inter-modal connections and reliability — and resilience against disruptive events. The SESAR concept of operations will drive improvements in the procedures being used by all ATM stakeholders, including airports. SESAR Project ID 06.05.04 D01 (SESAR, 2011) describes the way the AirPort Operations Centre (APOC) fits into the overall framework of the SESAR operational concept. While the main APOC responsibilities are related with the operational plan and the airside processes, these responsibilities also include "monitoring all closely linked landside/terminal processes with direct influence on the Airport and the Network Operations Plan". Recent research has shown that passengers' behaviour at the terminal and landside has a major impact on flights delays (EUROCONTROL, 2009a). The aim of this case study is to analyse the time spent by passengers in airports looking for bottlenecks and improvement areas, with the ultimate goal of grasping the impact on the ATM system. Geolocation data will be used to identify areas of congestion within the airport at given times and their correlation with special events, airline schedules and other relevant factors. Tracking algorithms will be used to characterise passenger movements inside the airport and estimate how much time travellers spend in each step of their journeys, analysing the areas where the waiting time can be reduced to achieve the 4-hours door-to-door objective. Queuing analysis will be used to estimate how much time is spent in the check-in, security, passport control and boarding queues, and how this time can be reduced (e.g., optimal number of passport/security control desks, etc.). Finally, we cannot forget that airports are business areas, each time closer to shopping centres (The Independent, 2015; Retail Week, 2014). Reducing the time the passengers spend in the airport may therefore have an economic impact on both passengers and business located in the airport. We will explore the possibility of using the credit card dataset to measure the impact of time efficiency improvements on passenger expenditures.

Case study 4: socioeconomic impact of ATM disruptions

Air travel disruptions (e.g., ash cloud, severe bad weather, controllers' strike, etc.) alter passengers' travel plans, often leaving them trapped in airports or their surroundings for hours. The objective of this case study is to explore how the information extracted from passenger-centric geolocated data can help assess the impact of ATM disruptions in a more comprehensive manner, extending existing disruption analyses to better take into account the direct and indirect effects on the passengers and society at large. For this purpose, we will compare different socioeconomic and behavioural indicators during normal days and days affected by major disruptions. Examples of the questions that will be investigated are the following:

- the modification mobility patterns, travel times and expenditure patterns in the presence of ATM disruptions;
- the impact of disruptions on the accessibility of European regions;
- the measurement of indirect economic costs of the disruptions by comparing visitors' expenditure in different cities/regions;
- passengers re-routing and transfer to alternative transport modes;
- passengers' opinions and perception of the ATM stakeholders.

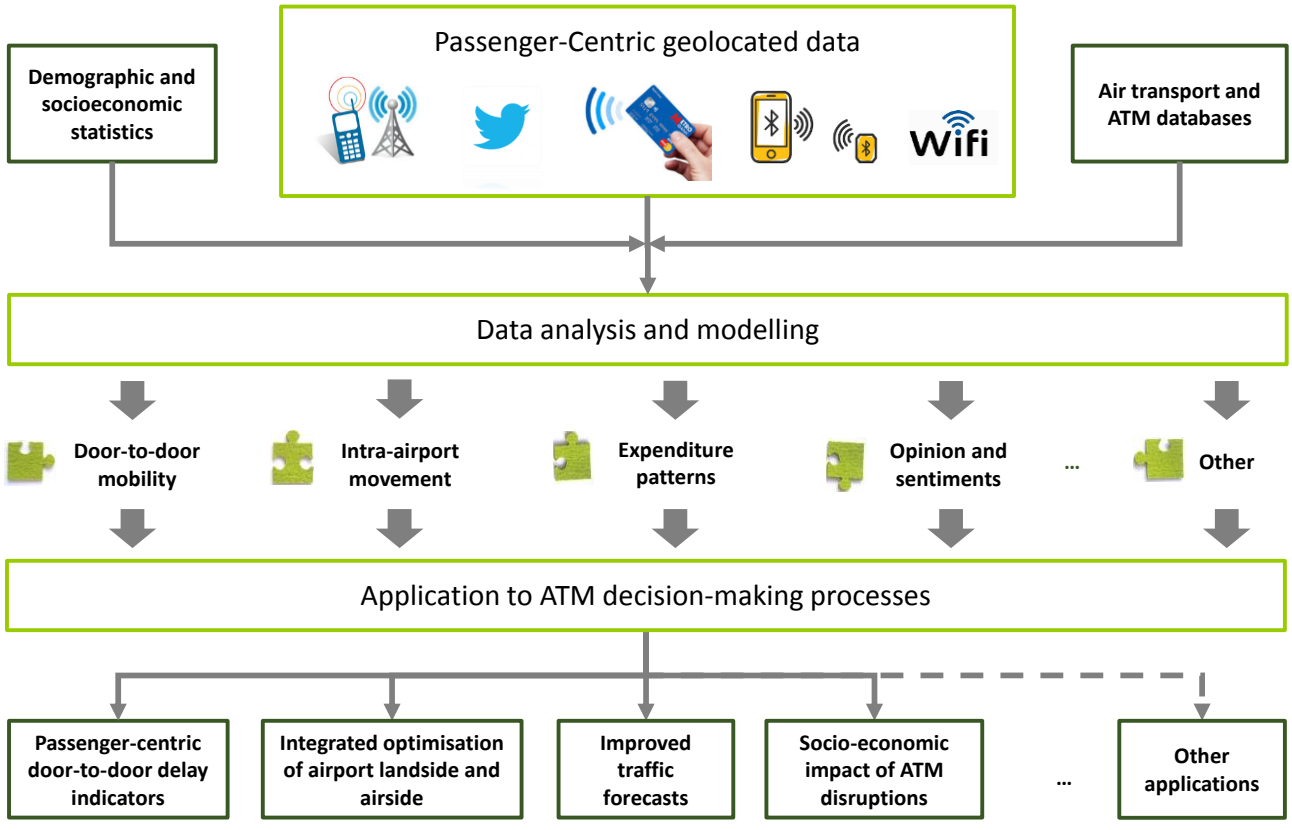

Figure 1. BigData4ATM overall concept

4.3 Target outcomes and expected impact

The project will deliver three main outcomes:

- a set of novel methodologies, algorithms and tools for translating passenger-centric geolocated data into meaningful behavioural information, such as door-to-door travel information, passenger intra-airport movements, expenditure patterns, or customer satisfaction levels;
- a set of new models and indicators built on passenger-centric data (e.g., accessibility of European regions and door-to-door travel times) providing an improved comprehension of the interrelationships between passengers' behaviour, the ATM system and European society; and
- a set of case studies demonstrating and evaluating the potential of the new data, models and indicators to provide new insights into the planning and management of the ATM system.

These outcomes are expected to render a number of benefits for the European ATM sector:

Improvement of the quality and availability of knowledge for decision-making. Information about passenger behaviour has so far been practically absent from ATM decision-making processes, largely due to the difficulties to collect accurate and updated behavioural data. The pervasive penetration of smart mobile devices opens the opportunity to fill this gap, by gathering rich data on citizens' attitudes and activities. The data sources investigated by BigData4ATM are expected to provide new insights about the relationships between passenger behaviour and ATM performance. These insights will inform the development of new theoretical models, metrics and indicators to translate the new datasets into actionable information for the planning and management of the ATM system.

More agile ATM system designs. The integration of actionable passenger-centric information may contribute to rendering the ATM system more resilient to challenges such as rapid changes in demand or disruptive events. The ability to evaluate ATM performance from the point of view of their impact on passengers and society at large, both in nominal conditions and in situations of performance degradation, is expected to provide valuable inputs for the design of agile ATM concepts and systems.

Exchanging passengers/freight information seamlessly between different transport modes. One of the results of BigData4ATM will be the extraction of modal split and travel time information for the different segments of the door-to-door itineraries. This information will serve to study the complementarity and competition between transport modes, as well as the interdependencies between airport connectivity and passengers' route choices. This knowledge will be of great value for improving the assessment of congestion externalities and defining flight prioritisation criteria that take into account the full range of ATM socioeconomic impacts. The results of these studies will contribute to identify relevant information that might be integrated in future intermodal information systems.

Performance benefits across the entire ATM system. BigData4ATM will contribute to define new passenger-centric metrics and indicators providing new angles of analysis of ATM performance, contributing to make progress towards an ATM system where traffic management decisions are ultimately driven by the passengers' needs.

References

Airports Commission (2013), Discussion Paper 01: Aviation Demand Forecasting.

Andrienko, G., N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, D. Thom (2013) Discovering Thematic Patterns in Geo-Referenced Tweets through Space-Time Visual Analytics, Computing in Science and Engineering, 2013, v.15(3), pp.72-82. http://doi.ieeecomputersociety.org/10.1109/MCSE.2013.70

Anna Aero (2014). Impact of Air France strike at French airports analysed; Paris CDG sees 12.3% drop in September traffic. Available at http://www.anna.aero/2014/10/22/impact-of-air-france-strike-at-french-airports-analysed/

Bagrow, J.P., Lin, Y.-R. (2012) Mesoscopic structure and social aspects of human mobility. PLoS ONE 7, e37676.

Borge-Holthoefer, J., Rivero, A., Garcia, I., Cauhe, E., Ferrer, A. et al. (2011) Structural and dynamical patterns on online social networks: The Spanish may 15th movement as a case study. PLoS ONE 6, e23883.

Boyd, D. (2010) Privacy and Publicity in the Context of Big Data. WWW 2010 conference, Raleigh, North Carolina, 29 April 2010.

Brockmann, D., Hufnagel, L. and Geisel, T. (2006) The scaling laws of human travel. Nature 439, 462-465.

Cheng, L. (2014) Modelling Airport Passenger Group Dynamics Using an Agent-based Method, Masters by Research thesis, Queensland University of Technology.

Conover, M.D., Davis, C., Ferrara, E., McKelvey, K., Menczer, F., et al. (2013) The geospatial characteristics of a social movement communication network. PLoS ONE 8, e55957.

Cook, A. and Tanner, G (2005) “Citizens” Study - Results of European Focus Groups Examining Public Perceptions of Air Transport Growth and ATM EEC/SEE/2005/013

Cook A., Tanner G., Cristóbal S. and Zanin M., (2012). Passenger-Oriented Enhanced Metrics, Proceedings of the 2nd SESAR Innovation Days.

Cook A., Tanner G. and Zanin M., (2013a). Towards superior air transport performance metrics – imperatives and methods, Journal of Aerospace Operations, DOI 10.3233/AOP-130032. http://dx.doi.org/10.3233/AOP-130032.

Cook A., Tanner G., Cristóbal S. and Zanin M., (2013b). New perspectives for air transport performance, in Schaefer, D. (Ed.) Proceedings of the 3rd SESAR Innovation Days.

Crane, R. and Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. PNAS 105, 15649-15653.

EUROCONTROL (2009a) Impact Study of Landside Elements on Airport Capacity and Delays.

EUROCONTROL (2009b) Challenges of Air Transport 2030 Survey of experts’ views.

EUROCONTROL (2010a) Air Transport Socio-Economic Studies at http://www.eurocontrol.int/eec/public/standard_page/proj_Strategic_soc_eco_studies.html

EUROCONTROL (2010b) Strategic and Socio-Economic Studies at EUROCONTROL Synthesis 2010

EUROCONTROL (2010c) Ash-cloud of April and May 2010: Impact on Air Traffic. EUROCONTROL/CND/STATFOR, June 2010.

EUROCONTROL (2015) Annual Network Operational Report 2014. EUROCONTROL June 2015.

European Commission (2010) Strategy for equality between women and men 2010-2015.

European Commission (2011a) White Paper: Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system. COM(2011) 144 final. Brussels, March 2011.

European Commission (2011b) Flightpath 2050. Europe's Vision for Aviation. Report of the High Level Group on Aviation Research.

European Commission (2013) Horizon 2020 Work Programme 2014-1015 - Smart, green and integrated transport.

European Union (2004) Regulation (EC) No 261/2004 establishing common rules on compensation and assistance to passengers in the event of denied boarding and of cancellation or long delay of flights. Official Journal of the European Union, February 2004.

Frias-Martinez, V., Soto, V., Hohwald, H. and Frias-Martinez, E. (2012) Characterizing urban landscapes using geolocated tweets. In IEEE SocialCom/PASSAT pages 239-248.

Grabowicz, P.A., Ramasco, J.J., Moro, E., Pujol, J.M. and Eguíluz, V.M. (2012) Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media. PLoS ONE 7, e29358.

Grabowicz, P.A., Ramasco, J.J., Gonçalves, B. and Eguiluz, V.M. (2014) Entangling Mobility and Interactions in Social Media. PLoS ONE 9, e92196.

Gonçalves, B., Perra, N. and Vespignani, A. (2011) Validation of Dunbar's number in Twitter conversations. PLoS ONE 6, e22656.

Gonzalez-Bailon, M., Borge-Holthoefer J., Rivero, A. and Moreno, Y. (2011) The dynamics of protest recruitment through an online network. Scientific Reports 1, 197.

Gonzalez, M.C., Hidalgo, C.A. and Barabasi, A.-L. (2008) Understanding individual human mobility patterns. Nature 453, 779-782.

Graham, M. (2012) Big data and the end of theory? The Guardian, 9 March 2012.

Hawelka B., Sitko I., Beinat E., Sobolevsky S., Kazakopoulos P., et al. (2013) Geolocated twitter as a proxy for global mobility patterns. ArXiv e-print arXiv:1311.0680.

IATA (2008) Air Travel Demand, Economics Briefing No 9.

IATA (2010) The impact of Eyjafjallajokull's volcanic ash plume. IATA Economic Briefing, May 2010.

ICAO (2006) Doc 8991 AT/722/3 Manual on Air Traffic Forecasting.

ICAO (2013) Passengers Protection under Cases of Flights Disruptions. Worldwide Air Transport Conference Sixth Meeting ICAO, Montreal, Canada, ACI 2013.

Jurdak, R. et al. (2015) Understanding human mobility from Twitter. http://arxiv.org/abs/1412.2154.

Kinchin B.(2004) A Synthesis of ATM Public Perception Surveys, EEC/SEE/2004/001

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., et al. (2009) Computational social science. Science 323, 721-723.

Lehmann, J., Ramasco, J.J., Gonçalves, B. and Catutto, C. (2012) Dynamical Classes of Collective Attention in Twitter. Procs. WWW, Lyon.

Lenormand, M., Picornell, M., Garcia Cantu, O., Tugores, A., Louail, T., Herranz, R., Barthelemy, M., Frias-Martinez, E. and Ramasco, J.J. (2014) Cross-checking different source of mobility information. PLoS ONE 9, e105184.

Lenormand, M., Tugores, A., Colet, P. and Ramasco, J.J. (2014b) Tweets on the road. PLoS ONE 9, e105407.

Lenormand, M., Picornell, M., Garcia Cantu, O., Tugores, A., Louail, T., Herranz, R., Barthelemy, M., Frias-Martinez, E, San Miguel, M. and Ramasco, J.J. (2015a) Comparing and modeling land use organization in cities. http://arxiv.org/abs/1503.06152.

Lenormand, M., Tugores, A., Gonçalves, B. and Ramasco, J.J. (2015b) Human diffusion and city influence. http://arxiv.org/abs/1501.07788.

Lenormand, M., Louail, T., Garcia Cantu, O., Picornell, M., Herranz, R., Murillo Arias, J., Barthelemy, M., San Miguel, M. and Ramasco, J.J. (2015c) Influence of sociodemographic characteristics on human mobility. Scientific Reports (In press),

Liben-Nowell D., Novak J., Kumar R., Raghavan P., Tomkins A. (2005) Geographic routing in social networks. PNAS 102, 11623-11628.

Louail, T., Lenormand, M., Garcia-Cantu, O., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J.J. and Barthelemy, M. (2014) From mobile phone data to the spatial structure of cities. Scientific Reports 4, 5276 (2014).

Louail, T., Lenormand, M., Garcia-Cantu, O., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J.J. and Barthelemy, M. (2015) Uncovering the spatial structure of mobility networks. Nature Communications 6, 6007.

Mahaud, P. and Arrighi de Casanova, D. (2004) What Image of ATM? An Analysis of 2002-2003 European Press EEC/SEE/2004/002

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B. and Vespignani, A. (2013) The Twitter of Babel: Mapping World Languages through Microblogging Platforms, PLoS One 8, E61981.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M. and Mascolo, C. (2012) A tale of many cities: Universal patterns in human urban mobility. PloS one 7, e37027.

OECD ITF (2014) Contemporary Airport Demand Forecasting: Choice Models and Air Transport Forecasting, Discussion Paper 2014-7.

Onnela J.-P., Saramaki J., Hyvonen J., Szabo G., Lazer D., et al. (2007) Structure and tie strengths in mobile communication networks. Proc. Natl. Acad. Sci. U.S.A. 104, 7332-7336.

Onnela, J.-P. and Reed-Tsochas, F. (2009) Spontaneous emergence of social influence in online systems. PNAS 107, 18375-18380.

Picornell M., T. Ruiz, M. Lenormand, J.J. Ramasco, T. Dubernet and E. Frías-Martínez (2015). Exploring the potential of phone call data to characterize the relationship between social network and travel behavior. Transportation, DOI: 10.1007/s11116-015-9594-1

Retail Week (2014) Analysis: How travel retail went from waste of time to strategically crucial -at: http://www.retail-week.com/property/analysis-how-travel-retail-went-from-waste-of-time-to-strategically-crucial/5058704.article.

Rose, N (2014). Passengers First, Re-thinking irregular operations. Amadeus IT Group.

Schultz, M. and Fricke, H. (2011) Managing Passenger Handling at Airport Terminals: Individual-based Approach for Modeling the Stochastic Passenger Behavior. 9th USA/Europe ATM Research and Development Seminar (ATM2011), Berlin, Germany.

SESAR (2007) Definition Phase Deliverable 3: The ATM Target Concept DLM-0612-001-02-00

SESAR (2011) Project ID 06.05.04 AirPort Operations Centre (APOC) definition, D01 Initial Operational Concept.

Song C., Qu Z., Blumm N., Barabasi A.-L. (2010) Limits of predictability in human mobility, Science 327, 1018.

Soto, V. and E. Frías-Martinez (2011) Automated land use identification using cell-phone records. Procs. of the ACM conference HotPlanet'11 pp. 17-22, New York.

The Independent (2015), Cover Story: Have airports become shopping centres with runways attached? -at: http://www.independent.co.uk/news/cover-story-have-airports-become-shopping-centres-with-runways-attached-1344426.html

The New York Times (2014). Air France Puts Cost of Pilots' Strike at More Than $400 Million Available at http://www.nytimes.com/2014/10/09/business/international/air-france-puts-cost-of-pilots-strike-at-more-than-400-million.html

Tizzoni, M. et al. (2014) On the use of human mobility proxy for the modeling of epidemics. PLoS Computational Biology 10, e1003716.

Wilson, A. (2010) Entropy in Urban and Regional Modelling: Retrospect and Prospect. Geographical Analysis 42 (2010) 364–394.

BigData4ATM

Passenger-centric Big Data Sources for Socioeconomic and Behavioural Research in ATM

This project has received funding from the SESAR Joint Undertaking under grant agreement No 699260 under European Union's Horizon 2020 research and innovation programme. Opinions expressed in this work reflect only the authors' view; the SJU shall not be considered responsible for any use that may be made of the information contained herein.

www.bigdata4atm-sesar.eu