

FEASIBILITY STUDY ON THE USE OF MOBILE POSITIONING DATA FOR TOURISM STATISTICS

Margus Tiru, Positium LBS

Rein Ahas, University of Tartu

Ulf Sonntag, The Institute for Tourism Research in Northern Europe

Maiki Ilves, Statistics Estonia

Ossi Nurmi, Statistics Finland

Jimmy Armoogum, French Institute of Science and Technology for Transport, Development and Networks

Christophe Demunter, Eurostat¹

ABSTRACT

An overview of the project “Feasibility study on the use of mobile positioning data for tourism statistics” conducted under the Eurostat contract no. 30501.2012.001-2012.452 is presented. The project investigated the feasibility of using mobile positioning data in compiling official tourism statistics and statistics in other domains. Mobile positioning is considered a new big data source and is generating interest in statistics and research community in tourism, transportation, population and other areas. The possibilities of creating supporting statistics, new indicators and beneficial aspects like timeliness, time and spatial accuracy and ubiquity, etc. are creating expectations towards this new data source. However there are many barriers and limitations that have to be taken into account. During the feasibility study the consortium studied the current status of the use of mobile positioning data (research, applications, official statistics), conducted a survey and interviews with stakeholders, analysed the legal situation on EU and EU Member States, analysed technological, methodological and economical (business, financial) barriers and opportunities of accessing and using the data. This paper reflects the most important outcomes of the study. Full reports of the study can be found on Eurostat’s webpage².

¹ The views expressed in this study do not necessarily reflect the official position of the European Commission.

² http://epp.eurostat.ec.europa.eu/portal/page/portal/tourism/methodology/projects_and_studies

TABLE OF CONTENTS

Abstract.....	1
Table of Contents.....	2
1 Introduction.....	3
2 Results of the Tasks.....	5
2.1 Stock-taking.....	5
2.2 Feasibility of Access.....	7
2.3 Feasibility of Use: Methodological Issues.....	12
2.4 Feasibility of Use: Coherence.....	17
2.5 Opportunities and Benefits.....	21
3 Conclusions.....	23
4 References.....	26

1 INTRODUCTION

A key output of the system of tourism statistics is information on tourism flows, i.e. the number of trips taken and the nights spent away by visitors outside their usual environment. Traditional data sources include surveys that are completed by accommodation establishments (reporting on guests who stayed at their establishments) and surveys filled in by households or individuals (reporting on trips made during a recent reference period). The European legal framework for tourism statistics (Regulation 692/2011) currently in place was developed against the background of these more traditional data sources (Eurostat 2012).

However, the European Statistical System (ESS) is influenced by changes in the business environment (new data requirements, need to simplify the collection process and reduce the ensuing burden, use of ICT tools), and changes in the ESS business architecture (integration of data sets, re-use of existing data or administrative data) (European Commission 2009). The fast-changing tourism market ‘has created new user needs or has changed existing user needs to which the legal framework needs to adapt by introducing new variables or breakdowns, whilst dropping some existing requirements that are no longer essential, and by improving the timeliness of the data,’ (Eurostat 2012). Therefore, Eurostat needs to play a role in developing more efficient methods for collecting data, which is also relevant for the field of tourism statistics.

‘Mobile positioning data can be a source for monitoring flows of persons, inside or outside their usual environment. In the short or mid-term, it can be an additional source of information for the system of tourism statistics (quick indicators, additional indicators to cover existing gaps – for instance flows of non-residents not staying at rented accommodation, quality checks and calibration of sample surveys), in the long term it could possibly replace part of the existing data collection work. It goes without saying that this would lead to a significant reduction in the reporting burden and to a significant reduction in processing and compilation cost for the Member States' statistical authorities’ (Eurostat 2012).

In this context, Eurostat initiated a supra-national feasibility study on the use of mobile positioning data for tourism statistics to explore the possibilities – and limits – of this new data source and methodology (Eurostat 2012). From December 2012 to March 2014, a consortium that consisted of six partners – Positium LBS (Estonia, project lead), Institute for Tourism and Research in Northern Europe (Germany), Statistics Finland (Finland), The

French Institute of Science and Technology for Transport, Development and Networks (France), Statistics Estonia (Estonia), and the University of Tartu (Estonia) – conducted this study.

The objectives of the study as underlined by Euristat were:

- a) the feasibility to access databases with mobile positioning data (including continuity of access).
- b) the feasibility to use mobile positioning data for tourism statistics – domestic, outbound as well as inbound tourism - in the European context (translating methodological concepts, evaluating possible systematic bias or other quality related issues, evaluation of the available data in relation to the requirements for tourism statistics) – this shall include tests comparing existing official tourism statistics with statistics produced using the mobile positioning data.
- c) to identify, discuss and address the main challenges (including the aspect of trust – users as well as producers).
- d) the potential impact on cost-efficiency of the data production.
- e) the possibility of re-using the methodology for other domains and define joint algorithms, in particular the integration of data from mobile positioning into the compilation of the travel statistics (i.e. travel item of Balance of Payments).

During the study 5 separate tasks were conducted, each resulting in a detailed report on the subject:

1. Stock-taking. This task made an inventory of all of the research that has been conducted to date, along with applications and experience, in EU Member States, EFTA, candidate countries and around the world, and provides an up-to-date description of the state-of-the-art in using mobile positioning data in research and applications in tourism statistics and related domains.
2. Feasibility of Access. The task reflected the potential opportunities and obstacles in terms of gaining access to passive mobile positioning data from mobile network operators. The main focus is on access to the data in order to produce official tourism statistics for the NSIs, but other usages are also considered. The task concentrated on regulatory, business and technological barriers along with practical access to the data.

3. Feasibility of Use: Methodological Issues. The task describes the methodology of tourism statistics that cover processes from obtaining raw data until the results are published. The described methodology is evaluated by using different quality aspects.
4. Feasibility of Use: Coherence. The main focus of the task was on carrying out a quantitative and qualitative comparison of mobile-phone-based tourism statistics with reference statistics that include official tourism statistics and other available indicators related to tourism activities. The qualitative element of the report is to describe the reasons for deviations in these data sources.
5. Opportunities and Benefits. This task evaluated the opportunities and benefits that can arise from using mobile positioning data for tourism statistics. Data sources for this assessment include actual usage cases, quality, synergy and coherence assessments that are based upon real data, and evaluations on data access, both with regard to technology and privacy. The assessment of opportunities and benefits in this study touches five different perspectives: quality, cost, access to new indicators, synergies and transmission.

This paper is based on the individual reports and the consolidated report of the study. Authors kindly suggest reading the original consolidated report and reports on the tasks to get more thorough insights on the results of the study.

2 RESULTS OF THE TASKS

2.1 Stock-taking

The report emphasises existing problems and solutions in technology, methodology, and regulations and other aspects of accessibility in terms of mobile positioning data. The information covered in the report is based on the circumstances that existed in summer 2013 regarding the following areas:

- Publicly available information (applications and scientific research);
- Knowledge of the consortium partners;
- Information from surveys and interviews.

The consortium collected and reviewed more than a hundred cases that had included the use of mobile positioning or other ICT-based data that bore any relevance to the study. The above

priorities and referenced usage cases were selected because of their direct or indirect pertinence when it came to using mobile data for tourism studies or as a statistical source. A total of 31 significant and representative cases are presented in the report along with an overview of the situation in 36 EU Member States, EFTA countries and EU Candidate States concerning mobile positioning data, research or applications.

The results of the usage case analysis highlight the fact that there is an increasing number of mobile data based studies, research papers, projects, applications and businesses being created. It can be assumed that in the next two to five years, such datasets will probably become a common source for studies and statistics in most European countries, and that alongside this process, the data will become available for generating tourism statistics. All successful usage cases were developed step-by-step and success was guaranteed by user persistence.

Society in general is getting used to the idea of implementing sensitive data along with the widespread progression of other privacy-sensitive developments, especially in the social media sphere. We are seeing progress in the data protection sphere, where the last decade has shown a degree of turbulence in relation to the newer aspects of privacy protection, and a good many countries are reviewing their national legislation in order to cope with social issues as well as security threats in the area of data protection. Though tourism statistics neither require nor are intended to track the activities of a specific person, this is something that is closely related to and is perceived as being part of data protection. MNOs are overcoming privacy and business confidentiality-related concerns as they see the appearance of new revenue possibilities, and also the value that can be gained from the internal use of such data. The development of the technology and the principles for processing sensitive data safely and anonymously should allow legislation to adopt and develop the practices for using such data. The analysis of the available research also demonstrated the most valued strengths of mobile data: relatively simple, quick and cheap collection of data whilst including a large sample. The time (including the level of individual events) and spatial (the accuracy of a mobile cell, potentially GPS precision) accuracies are also better than previously. The problems related to the dataset include a lack of qualitative information on the user, such as the purpose of the visits and the means of transport; as well as privacy issues and the problems that can accompany the processing of large amounts of data.

Most of the time, access to data results in long-term, trust building cooperation between all of the parties involved, in which projects grow from small-scale testing projects to wider collaborations. This has led to a step-by-step process of obtaining larger amounts of data with better quality.

A second group of companies have direct business solutions focused on providing technology for MNOs. New business projects are started in order to promote the operator through services beneficial to society. There are also cases with clear profit models.

The third group of cases involved in obtaining data consists of short-term projects, such as Real-Time Rome or sporting competitions in Milan. MNOs agree to reveal ‘a bit’ of their data for certain events or projects for the purpose of publicity or for testing potential business opportunities. Several large-scale projects have started from such small ventures.

Based on findings from the usage cases that have already been examined, it can be concluded that mobile data is being used increasingly in a number of different fields. Most of the active usage is within academia, with some already established applications on a state. There are a few direct usage cases and examples that cover the use of mobile positioning data in generating tourism statistics. The preliminary results from this task’s questionnaire show, however, that many of the statistical bodies have already considered the use of such data or have even contacted MNOs with data requests. It was discovered that a major problem for statistical bodies was not methodological but concerned access to data, privacy concerns, and the relatively high ‘entry cost’ of using new ICT-based data sources.

Business-oriented projects that are the focus of the mobile operators concentrate on geomarketing; non-MNO models focus on providing technology and/or brokerage of the results data for usage in transportation, traffic, urban studies, regional development, and tourism applications.

2.2 Feasibility of Access

This task concentrates on the questions concerning the access to the data. Legal, business-related and technological aspects were investigated during this task.

2.2.1 Regulations and Privacy Protection

A legal analysis was conducted, which covered regulation and privacy protection aspects and which consisted of an examination of the legislation at EU and national levels (in Estonia, Finland, France and Germany) relevant to the subject of the study. The aim of the analysis was to assess the current situation and to provide insight in relation to current and future options when it comes to accessing the data.

Legal restrictions are considered to be the most important barrier when it comes to accessing the data. The legal acts do not specifically concentrate on the use of mobile positioning data in statistics. The relevant areas of the legislation on the EU level include:

- the Data Protection Directive (Directive 1995/46/EC and its successor, the General Data Protection Regulation);
- the Electronic Privacy Directive (Directive 2002/58/EC);
- the Data Retention Directive³ (Directive 2006/24/EC);
- the European Statistics Regulation (Regulation 223/2009/EC);
- the European Statistics Regulation on tourism statistics (Regulation 692/2011/EU);
- and the opinions of Article 29 Data Protection Working Party⁴.

These legal acts stipulate the following main general principles for the use of mobile positioning data:

- Directly or indirectly identifiable mobile positioning data can be used and processed for statistics if the following is true:
 - a) The subscriber has given their consent for their data to be processed for such specific purpose; or
 - b) The official statistical organisation (the NSI) needs the data in order to perform an official task imposed on them by the law.

³ The European Court of Justice has declared the Data Retention Directive to be invalid in its judgement of 8 April 2014, i.e. after the date of this Report. It remains to be seen when and to what extent the member states' relevant local laws that were enacted under the Data Retention Directive will be reviewed.

Link: <http://curia.europa.eu/jcms/upload/docs/application/pdf/2014-04/cp140054en.pdf>

⁴ On 10 April 2014, i.e. after the completion of this Report, the Article 29 Data Protection Working Party adopted their Opinion 05/2014 on anonymisation techniques analysing the effectiveness and limits of existing anonymisation techniques and giving guidance on how to handle these in light of securing the privacy of individuals.

Link: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

- Fully anonymous mobile positioning data can be processed and used without restriction given that subscribers cannot be directly or indirectly identifiable in any given processing stage.

The two options set out the clear limitations on the methodology used for processing the data in theory. The full methodological opportunities open up when it comes to the use of the longitudinal data, which makes it possible to calculate the crucial aspects of tourism activities (and other domains), such as defining the usual environment, the duration of visits, movement paths during the trips, etc. However, the longitudinal data on its own is a factor that can allow the identification of individual persons. The anonymisation methods currently being used exclude any options available in relation to long-time analysis (e.g. the aggregation of the initial data before it is processed), and therefore seriously limit the available options in relation to methodology. There is a need for research programmes to develop the process of anonymisation so that it is able to preserve the longevity of the data, as no conclusive methods have been produced so far.

An alternative interpretation can be formulated, which decrees that, because the end result of processing is, by itself, anonymous (involving an aggregated number, duration, and the travel patterns for collective trips), the processing of personal data for such a purpose can be interpreted as not being illegal. However, such an interpretation can be considered as a grey area when it concerns privacy protection legislation as the legislation states that any processing of personal data, irrespective of the results, falls under the legal constraints of data protection legislation.

The current European tourism statistics regulation concentrates on data sources that are used today, which to some extent results in statistical indicators that are driven by the specific data sources that enable to compile these statistics (e.g. accommodation statistics) rather than be driven by the full set of user needs. During the course of the study, several respondents from the survey proposed the development of a central European framework for NSIs in order to enable them the access to mobile positioning data and simplify the process of obtaining data. This would mean either the inclusion of such data sources, the corresponding statistical indicators and classes in the list of official and mandatory statistics or the drafting of more specific and clearer articles within data protection and electronic communications legislation that would make it easier for MNOs to provide the data for NSIs or other users based on the assumption that the purpose behind its processing is purely the collection of statistical

information and not any intrusion into the privacy of the individuals concerned. However, this is a lengthy process at best, and it requires in-depth legislative work.

2.2.2 Business-related Barriers

Business barriers of the accessibility to MNO's data include different aspects besides the legal questions discussed above. From the MNOs' point of view, mainly three topics play a vital role:

1. Public opinion, the publicity of the MNOs, and consequences that might be caused by processing the data of the subscribers. MNOs are afraid that although, the data is processed anonymously and for statistical purposes, the public might perceive this as "tracking people" and might induce the decrease of the number of subscribers.
2. Burden for MNOs against benefits. In order to extract and process the data, MNOs need to make investments and allocate human and technical resources for maintenance. A question of patents is also one of the aspects of required investments as some of the data processing technology might be patented. Unless MNOs are compensated for their expenses, the willingness to provide the access is less probable. This topic also includes the potential possibility for MNOs to monetise on the data.
3. Confidentiality issues with business secrets includes the fear of competitors getting information about the number of subscribers, potential revenues, service activities, demographics and geography of customers, and technical information about the network (e.g. number of antennae and coverage areas) of the MNO. Therefore safeguards concerning the processing of the individual MNO's data and the access to the raw data is of a concern.

2.2.3 Technological Aspects

This task includes the discussion on the processing and preparation of the data for the further processing for statistical purposes. Where technical access and data processing are concerned, the main questions are connected to the specific data sources – there exists a wide range of databases and registries that can be used; however, they often differ between MNOs depending on the system architecture being used and on their technical ability to be able to store different types of data.

Three main types of data can be distinguished, based on their origin within the MNO's systems, where such data is required for the compilation of tourism statistics:

- a) (any kind of) event data (metadata) that covers subscriber activities and which is included in the MNO data stream;
- b) geographical cellular (network) referencing data;
- c) attribute data for subscribers (e.g. demographic information taken from the customer database).

Event data can be divided into internal and external network events and, furthermore, it can be broken down in the same way as for the various forms of tourism statistics:

- a) MNO internal events:
 - a. inbound roaming;
 - b. domestic;
- b) MNO external events:
 - a. outbound roaming;

The most common source for mobile data, as well as being the easiest to access, is Call Detail Records (CDR), with it being possible to take data from inbound and outbound roaming and domestic datasets as this data is held in storage and is rather easily accessible by the MNOs, although at the same time this is the most tenuous (in terms of the number of records/events per subscriber per day). The CDRs represent the active usage of mobile devices – covering incoming and outgoing calls and SMS text messaging. The biggest problems with CDRs are the frequency and the regularity of the records as they are based on the usage pattern of the subscriber. The average number of CDRs for tourists is approximately four events per subscriber per day – meaning that there is an average of four location facts for a phone for every single day. This is sufficient for some areas, but it sets limits upon domains in which better temporal accuracy is required (e.g. hourly statistics on a very small geographical space). Alternative data sources such as Data Detail Records (DDR), location updates, or others can include up to several hundred location events per user per day; however, such data is not often stored by MNOs. The geographical locations of the events are retrieved from the reference between the events and the geographical attributes of the network antennae.

The size of the data block, the number of records it covers, and the processing complexity it creates all require a sophisticated data processing system that can roughly be divided into two options (see Figure 1):

1. Data is extracted and processed within the MNOs, and the resulting statistical indicators are transmitted to the NSI, where the results from several MNOs are combined to create the final statistical indicators;
2. Data is extracted by MNOs and transmitted to the NSI, where the processing is carried out in order to produce the final indicators.

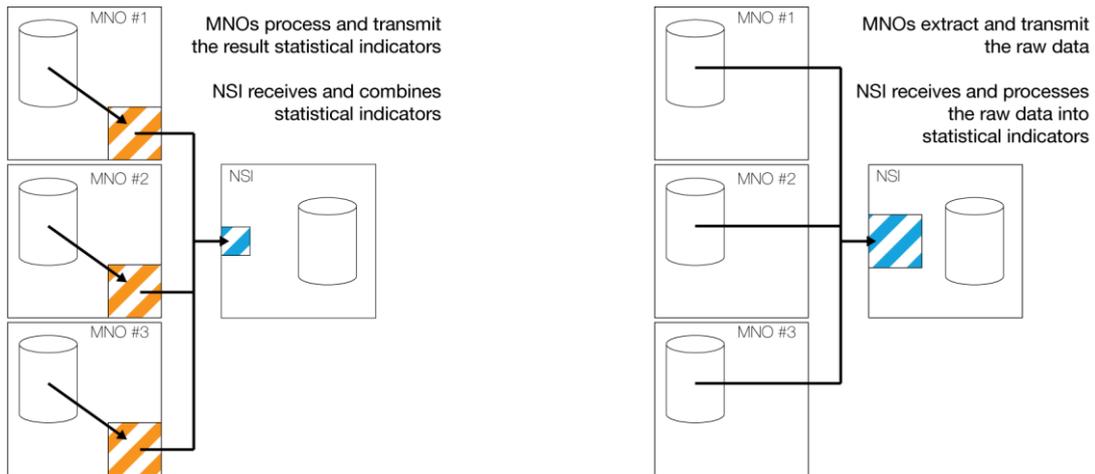


Figure 1. Two options for processing the mobile positioning data.

The cost and distribution of the burden are different for both scenarios. In addition, there is no clear preference as both these options have their clear benefits and disadvantages. The main technological challenge lies in the ability of the system to carry out periodical complex processes involving large data records within the designated timeframe with the chance of being able to recalculate the results in case any error occurs.

2.3 Feasibility of Use: Methodological Issues

2.3.1 Data Processing Methodology

In the current task, the steps following data preparation by MNOs are listed and described. During the processing of mobile positioning data, there are several important steps that enable generating the tourism statistics, such as the identification of usual environment and the country of residence, the duration of stay in specific place, differentiation between same-day and overnight visits, etc.

Depending upon the availability of the data and upon technological availability, all of those trips that are taken within the framework of the requirements can be analysed where they can be seen to correspond to the situation shown in the census or, alternatively, a subset of

observations within the framework could be selected. The sample sizes can be substantially larger in this situation when compared to traditional sample surveys, as the cost and burden of data collection is much less driven by the number of observations in the sample. The sample size can be determined from available technological capabilities and disclosure rules.

The methodology contains the following sections: the preparation of event data, frame formation, data compilation and estimation. The initial data extracted and prepared by MNOs is based upon network events that specify a specific subscriber's presence in time and space. Additional preparation may include geographical referencing, the elimination of non-human-operated mobile devices, checking the time and area coverage of the data, dealing with missing values, etc. After the data has been prepared by MNOs, the following processing steps are set out:

- Frame formation:
 - Applying trip identification algorithms – identifying each subscriber's individual trip to the destination (country of residence, foreign country) in question with the start and end times for each trip;
 - Identifying the population of interest (distinguishing tourism activities from non-tourism activities):
 - Defining roaming subscribers not actually crossing the border and entering the country (inbound, outbound);
 - Defining residents (inbound, outbound);
 - Defining the place of residence and the usual environment (domestic);
 - Identifying country-wide transit trips (inbound);
 - Identifying destination and transit countries (outbound);
- Data compilation:
 - Spatial granulation (visits at the smallest administrative level for inbound);
 - Defining variables (number of visits, duration of trips, classification, etc.);
- Estimation (from an MNO-specific sample to the whole population of interest) contains:
 - Time and space aggregation of the data (day, week, month, quarter/grid-based (one km²), LAU-2, LAU-1, country);
 - Combining data from various MNOs and computing final statistical indicators.

This methodology description in the current task could be taken as step-by-step guidelines on how to produce tourism statistics as it is rather detailed but at the same time general enough for broad use. There is an assumption that several variables relevant for the forming of new

variables exist in the dataset and that the activities of anonymous subscribers can be followed over a longer period of time to establish their residency and/or usual environment. The assumption regarding the availability of longitudinal data is crucial for the production of tourism statistics. Algorithms identifying the usual environment and country of usual residence rely upon the availability of past data. If the available data describes only a short period of time then the issue of processing errors arises and, as simulations show, the data quality can be too low to produce reliable results. Such limited data can be used as comparison indicators in some unofficial domains (e.g. the number of unique foreign subscribers on the site of attraction or concert) and for relative comparison.

2.3.2 Quality of the Methodology

The methodology is evaluated with regard to validity, accuracy, and comparability. Other evaluation aspects are covered in different tasks of this study, such as accessibility, coherence, timeliness and costs. The standard quality aspects are reviewed here to assess and improve the quality of the output. Although the same aspects are used to evaluate the estimates based upon survey data, there are differences in the relevance of these aspects to the output. For example, sampling errors and non-response errors are often measured and evaluated in traditional sample surveys as these are usually the main error sources.

There are several inherent data limitations that impact the quality of the methodology. The shortcomings that are inherent in mobile data include the lack of information on expenditure, the purpose of the trip, the method of transportation and the overall qualitative aspects of the tourism activity. It should be noted that the quality of the final outcome relies heavily on the availability of external information such as accommodation statistics, transport statistics, information about the market share for mobile operators and their subscribers and other information used for estimation. The quality of this information should be assessed before using it as auxiliary information in the estimation models.

Many quality issues that are presented by using mobile positioning data in official statistics are similar to those faced when starting to use a new administrative data source – differences in the concepts and definitions, coverage of the population frame, the representative nature of the sample, etc. In the case of administrative data or in mobile positioning data, NSIs do not have any influence over content, so an assessment of the validity is mainly qualitative (consisting of a description of the discrepancies between what is measured and what should be measured).

When comparing the definition for the target populations (on its participation in tourism activities and on the characteristics of tourism trips) against the population frame, there is one discrepancy that is immediately clear. Within the frame are all of those subscribers who used their mobile phone for calling or texting whilst the population shows all of those individuals who reside in the country. This leads to a large number of various coverage problems such as portion of the population not using mobile phones or using several devices, country-of-origin and demography-based under- and over-coverage, etc.

The accuracy is the most problematic quality aspect for this type of data, especially in terms of coverage issues. It is problematic because there are many components that contribute to the coverage bias and assessing all of them, separately or together, is a very complex task. There is no one method available at the moment that will allow easy estimation of the different biases. For several quality issues, quantitative results are given in the report based upon Estonian mobile positioning data to describe and illustrate the problem.

The quality assessment for assuring the comparability should be carried out when changes in methodology occur (e.g. due to the number and structure of MNOs), just as it should be carried out in traditional surveys. In addition, it is important to be ready to update the methodology if changes in the telecommunications technology or in the data structure occur.

There are many contributors to the coverage bias, but due to the co-effect some bias components cancel each other out (over-coverage versus under-coverage), some contribute very little, and some may contribute a lot. Many problems, however, are inherent in mobile positioning data and therefore cannot be avoided. Furthermore, their total effect, i.e. the total size of the coverage bias of an estimate of interest, needs to be evaluated or bias-corrected estimates need to be computed.

Similarly to traditional surveys, comparability can suffer for statistics based upon mobile positioning data when changes in underlying legislation occur or when bigger changes in the methodology or technology are introduced. The impact of these changes can be at best minimal or non-existent but they can also lead to a break in the series depending upon the nature of the change.

Due to the nature of passive mobile positioning data, the quality of the estimates based upon this data source depends on changes in the telecommunications market, e.g. the cost of calls and text messages and the way in which individuals use their mobile phones. Mobile phone

technology has developed very rapidly and people use mobile phones for much more than simply calling and texting. It can be seen that the increased level of options available (e.g. mobile broadband) will change people's phone usage habits, resulting in a change of the data content (e.g. less calls and messaging events and more data on internet usage – DDR). As a consequence, the use of advanced data (in addition to CDRs) is called for in order to maintain and improve the quality of the data.

Concerning the quality issues, there are several of them to take into account when processing mobile positioning data: the differences in the concepts and definitions when compared to the current official statistics; the over-coverage and under-coverage of the population frame; sampling issues; measurement and processing issues; and the comparability questions. From a statistical point of view, the main concern is the selection bias. Indeed, even if its diffusion is high all over Europe, not everybody uses a mobile phone, and many people have more than one mobile device. As the data cannot be easily obtained from all MNOs, any change in the structure of this sector would introduce a bias in the measurement of evolutions. However, our tests have shown a good fit between mobile positioning data and conventional monthly time-series on examples.

2.3.3 Relevance for Other Fields of Statistics

Although tourism statistics were the main focus of this study, several other fields of statistics may benefit from this source if they use the same or similar definitions and joint processing schemes as in tourism statistics. Based on the analysis carried out, the following areas of official statistics might benefit if they were implemented as a single system based on mobile positioning data: Balance of Payments travel item, Passenger Transport, Population, Migration and Commuting Statistics.

In economy and finance statistics, mostly the Balance of Payments (BoP) is relevant for assessing synergies with tourism statistics. Statistically, 'tourism' is a subset of 'travel' and consequently, (tourism) 'visitors' is a subset of 'travellers'. Moreover, the issues for tourism activities, namely the problem to accurately delineate the usual environment, does not apply to the travel item in the Balance of Payments. Methodologically, during the processing of data, several objectives can be achieved depending on the system's setup. For example, although domestic tourism concentrates on travel outside the usual environment, the same process can be extended to identify any trips taken within the usual environment.

For calibrating transport demand and organising transport supply, it is very important to have accurate estimates of origin-destination matrices. However, it is quite difficult and very costly to obtain these matrices through conventional survey methods. Therefore, MNOs can provide less costly and much more accurate matrices from mobile positioning data. However, the data will not show the mode of transport or the purpose of the trip.

In population statistics, the spatial distribution of population (living, working) and mobility aspects such as commuting will be relevant. As opposed to census-based statistics, mobile positioning data will always lack accuracy and will not offer the currently required level of detail in terms of socio-economic breakdowns. Nevertheless, the data is timely and can provide overall indications concerning commuting, migration and internal migration information.

The implementation of a system of statistics production based on mobile positioning data is rather expensive; however, if the system is implemented for several domains (tourism activities including BoP, transportation and population), the additional costs for adding processing components is relatively lower.

2.4 Feasibility of Use: Coherence

The focus of this task was on carrying out a quantitative and qualitative comparison of mobile-phone-based tourism statistics with reference statistics that include official tourism statistics and other available indicators related to tourism activities. The qualitative element of the report is to describe the reasons for deviations in these data sources.

The analysis is carried out with mobile phone based statistics from Estonia⁵. These statistics are compared against the official tourism statistics made available by Eurostat and against other indicators related to tourism statistics.

The key questions addressed in the report include the following:

- How exhaustively do positioning-based statistics cover physical tourism flows as measured by the reference statistics?
- How well does positioning-based data support a breakdown into the various components of tourism activities: inbound, outbound, domestic tourism and/or same-day trips, paid accommodation, and free accommodation?

⁵ Data from other countries was not available for the study due to access problems.

- How does positioning-based data compare to existing statistical indicators?
- What are the reasons behind the deviations in existing statistical indicators? Are these deviations strengths or weaknesses of mobile positioning data?
- Does positioning-based data support previously unavailable levels of detail such as geography, time, nationality, etc.?
- Are there any previously unavailable metrics that can be compiled using positioning-based statistics?

Tourism statistics from mobile positioning data were provided by a company Positium LBS, based upon datasets from two Estonian mobile network operators. The data includes Estonian inbound, outbound and domestic tourism between 2008 and 2012.

The reference statistics were provided mainly by Statistics Finland and Statistics Estonia. Reference data, available on the Eurostat website, on the outbound trips of Estonians to EU Member States was also used.

The mirror statistics method is used to compare two metrics that describe the same scope of observation. An example of mirror statistics is the number of Finnish tourists in the Estonian positioning-based inbound statistics when compared to the number of Finnish tourists in the official Finnish outbound tourism statistics. These comparisons are visually plotted and are analysed using statistical indicators that measure both coverage (CC2-value⁶) and consistency (r-value⁷). An r-value close to one indicates a high correlation with the reference statistics. Similarly, a CC2 value close to zero indicates a close match in the absolute number of trips.

A sample of these comparisons provides an overview of the consistency of mobile positioning data compared to various mirror statistics, as visualised in Figure 2.

⁶ CC2: asymmetry for mirror flows statistics – coefficient. The absolute difference between inbound and outbound flows to and from a pair of countries divided by the average of these two values.

⁷ R-value: The indicator that was chosen for measuring consistency over time is the Pearson product-moment correlation coefficient or, in short, Pearson's r. The indicator measures the linear correlation between the two variables X and Y.

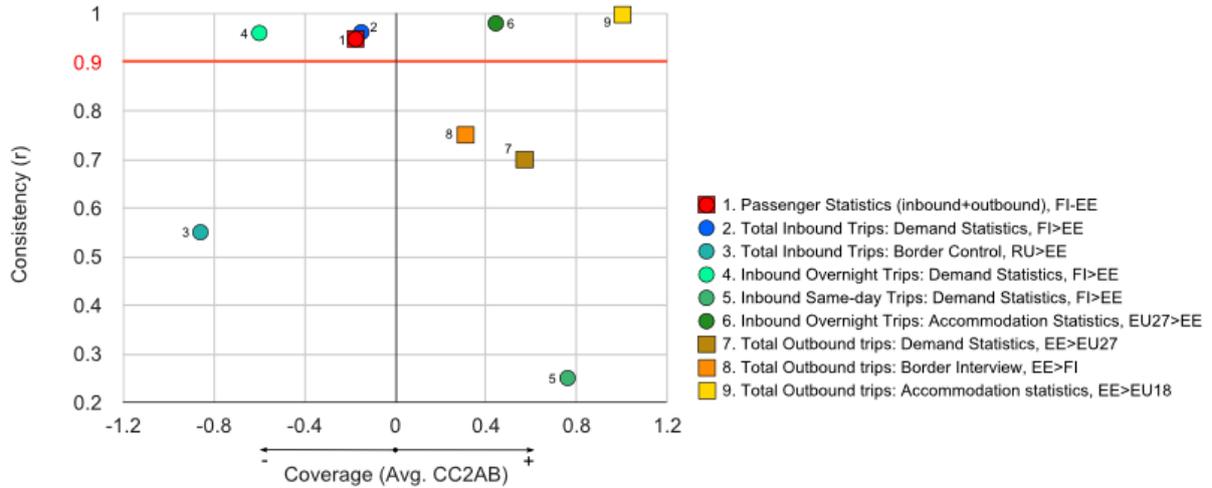


Figure 2. An overview of the two observed coherence indicators (consistency and coverage) compared against various mirror statistics (Demand, Supply, Border and Passenger Statistics).

A few relevant examples from the study are presented here to illustrate the different aspects of the coherence.

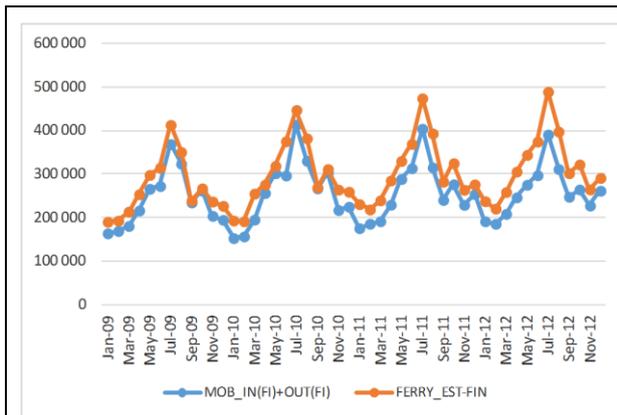


Figure 3. Monthly trips inbound (Finns to Estonia) + outbound (Estonians to Finland) in mobile positioning data when compared to the number of ferry passengers sailing between Finland and Estonia (Source: Positium, Finnish Transport Agency).

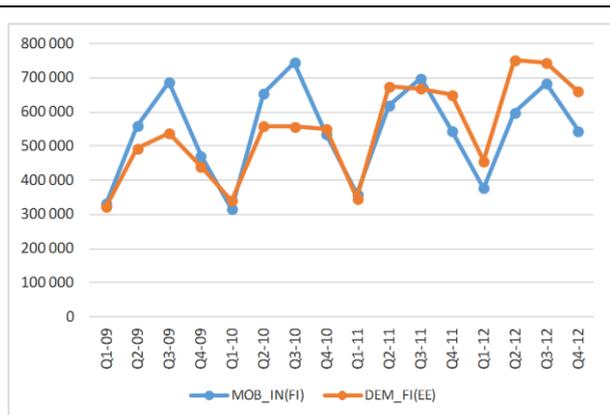


Figure 4. Quarterly inbound trips (Finns to Estonia) in mobile positioning data when compared to the total number of trips to Estonia in Finnish Demand Statistics (Source: Positium, Statistics Finland).

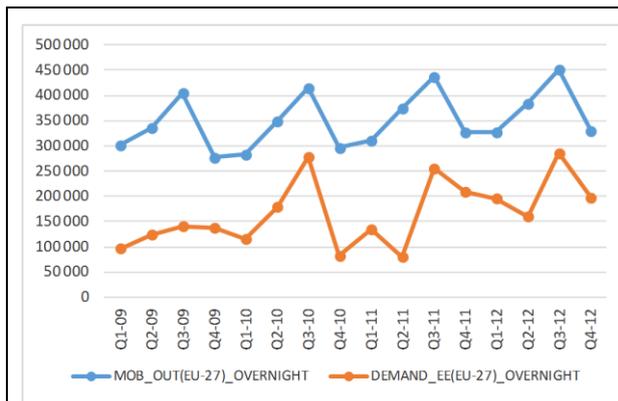


Figure 5. Quarterly outbound overnight trips (Estonians to EU27) in mobile positioning data when compared to the number of overnight trips to EU27 countries in Estonian demand statistics (Source: Positium, Statistics Estonia).

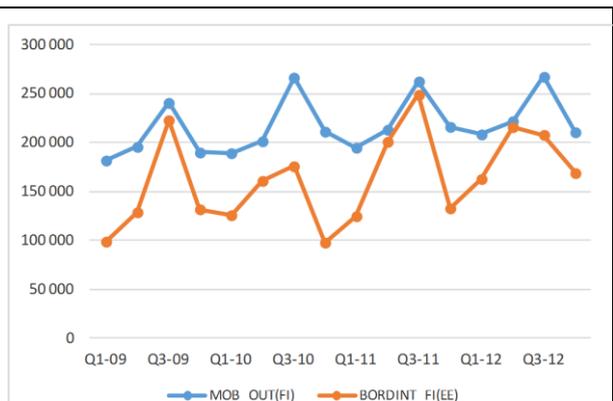


Figure 6. Quarterly outbound trips (Estonians to Finland) in mobile positioning data when compared to the number of Estonian visitors in the Finnish Border Interview Survey (Source: Positium, Statistics Finland).

Regarding overnight trips, mobile positioning data provides a consistent estimate over time and the correlation to existing tourism statistics, both in terms of demand and supply, is very good for the most part. For example, a weaker correlation to supply statistics that involves Latvian visitors to Estonia may in fact indicate one of the strengths inherent in mobile positioning data as it captures a more complete tourism flow, as compared to supply statistics that only cover stays at paid accommodation. While mobile positioning data cannot provide the split into paid and non-paid accommodation (a problem only if the split as such is relevant for the users), it is nevertheless a more comprehensive data source for overnight trips.

A potential weakness is the presence of cross-border commuters in mobile positioning data. According to definitions used in tourism statistics, these commuters should be excluded from tourism statistics since they regularly commute to another country. Many Estonians commute to Finland regularly and might appear in the Estonian outbound mobile positioning data as seen in the analysis comparing mobile positioning data to data from the Finnish Border Interview. The exclusion of these commuters requires a non-changing subscriber ID in mobile positioning data based upon which the frequent commuters could be excluded from the inbound and outbound mobile positioning data (by the definition of the usual environment). As the raw data contains this kind of user ID, such filtering of frequent visitors could be carried out at the level of raw data.

Based upon the requirements laid out in Regulation 692/2011, it is clear that mobile positioning data alone cannot fulfil these requirements. Although mobile positioning data provide fairly good estimates for the number of trips, the nights spent and the destination,

they do not produce any additional data about trips such as the purpose of the trip, the type of accommodation or the expenditure. For this information, additional data sources such as traditional surveys would still be needed.

Due to better timeliness when compared to official statistics, mobile positioning data would be particularly useful when it comes to providing quick indicators on the total number of tourism flows in order to complement the current supply statistics and their focus on paid accommodation alone. Mobile positioning data could also be used for producing more accurate regional tourism indicators. Mobile positioning data provides a source for calibrating the current surveys used for measuring tourism activities.

2.5 Opportunities and Benefits

This task evaluated the opportunities and benefits that can arise from using mobile positioning data for tourism statistics. Data sources for this assessment include actual usage cases, quality, synergy and coherence assessments that are based upon real data, and evaluations on data access, both with regard to technology and privacy.

The assessment of opportunities and benefits in this study touches five different perspectives: quality, cost, access to new indicators, synergies and transmission.

Within the **quality** perspective, it is outlined whether mobile positioning data is superior or inferior to more traditional data sources and processes in terms of completeness, timeliness, validity, accuracy, consistency and resolution. As a reference, tourism statistics as laid down in Regulation 692/2011 were chosen because this is the main framework for tourism statistics at the EU level today. It could be shown that mobile positioning data is highly consistent with reference statistics over time. At the same time mobile positioning data can be made available much more quickly than data from traditional sources, once the necessary processes are in place. As a result it can be shown that within the framework of official tourism statistics, mobile positioning data can be exploited as quick indicators and as calibration source. Of these two, quick indicators have the biggest potential to improve tourism statistics. On the other hand, it has to be stated that at present, mobile positioning data can hardly replace existing indicators within the framework of Regulation 692/2011.

Mobile positioning data can be used to potentially strengthen current tourism demand surveys through mixed-mode data collection. In such a scenario the number and duration of trips are based on mobile positioning data while tourism expenditure and ratios (purpose of trip, type

of accommodation, means of transport, etc.) still rely on demand survey. The sample size of the demand survey could be decreased considerably since the survey does not need to support breakdown by destination, thereby reducing the cost and burden of data collection. Also more countries and even sub-regions could be included in the statistics since the sample size is not an issue in mobile positioning data.

Within the framework of the official regulation, the **cost** that would be incurred within an NSI when using mobile positioning data was assessed as opposed to the sources and processes exploited today. As a result it can be shown that using mobile positioning data would require between 168 and 264 man-days per year. When compared to the workload induced traditional data sources and processes, mobile positioning data can be obtained and processed rather more efficiently. It has to be taken into account, however, that mobile positioning data is useful primarily as a calibration source or as a quick indicator, implying that at present, such data will be used in addition to other sources and processes.

Statistics as described by the Regulation 692/2011 are of course not the only tourism statistics that are being produced in EU Member States. The National Statistical Institutes (NSI) and other institutions produce tourism-related statistics that go beyond the framework of the regulation. Therefore, any possibility to gain **access to new indicators** through the use of mobile positioning data were assessed, firstly within official national tourism statistics (but not necessarily within the framework of the EU regulation), and secondly outside official tourism statistics (e.g. via destination marketing organisations). In official tourism statistics outside the EU regulation, mobile positioning data can deliver new information on those aspects, which are not required today mainly for methodological reasons. The two main fields of application are: 1) on the supply-side, overnight tourism in accommodation establishments below the threshold and 2) on the demand-side, overcoming the restriction of tourism for personal reasons. Mobile positioning data can reflect tourism in all sorts of establishments (including private accommodation) and for all purposes (including business). Spatial and timely resolution can, on average, be much more refined than required by the regulation when using mobile positioning data and therefore open up new possibilities when it comes to tourism statistics. When compared to surveys, mobile positioning data shows a number of advantages in terms of accuracy (smaller sampling error, no memory gaps), regardless of the selection bias.

Outside the field of official tourism statistics, mobile positioning data can be exploited in various contexts, such as detailed statistics in time and space, volume and structural breakdowns (nationality) for big events and other related applications, statistics describing accommodation not being covered in tourism statistics or segmentation data relying on subscriber master data.

Assessment of **synergies** also goes beyond the scope of the regulation on tourism statistics to show uses in other domains. It has been shown in pilot studies that mobile positioning data can be used in terms of official statistics specifically in the travel item of the Balance of Payments in transport and commuting statistics. Especially for transport statistics, mobile positioning data has the potential to provide new insights that are otherwise unavailable. For an NSI, the decision to exploit mobile positioning data within the regulatory framework of the specific country will be much easier when it becomes possible to share costs over different statistical domains.

Assessing the opportunities that might arise from the **transmission** of data can be viewed within the framework of the regulation (i.e. from NSIs to the European Commission), but little impact is to be expected in this area. A higher impact will probably arise from the transmission of data from MNOs to the NSI – but this only after the necessary automation processes have been carefully planned and thoroughly tested.

3 CONCLUSIONS

The aim of the current study was to assess the feasibility of using mobile positioning data for generating statistics on domestic, outbound and inbound tourism flows, and to address the strengths and weaknesses related to access, trust, cost, and the technological and methodological challenges inherent in the use of such a new data source. The international consortium that conducted the study concentrated on the various aspects involved in the use of mobile positioning data in terms of tourism statistics and other domains: an overview of the situation involving the use of mobile data; accessibility to the data from the legal, technological, financial and business aspects, including possible cost and burden implications; methodological principles of statistical data collection and compilation, including evaluation by using different quality aspects and comparing the results against existing traditional methods; opportunities offered by, as well as limitations inherent in, the

use of the data source. A more detailed discussion on each of these aspects can be found in the detailed reports, which can be consulted and downloaded from the Eurostat website.

The study demonstrated extensive interest in this data source and possibilities for a wide range of uses when it comes to tourism statistics and other domains, while at the same time acknowledging the multiple problems associated with accessing and processing the data. The number of projects and research areas that are focusing on the use of mobile positioning data is constantly increasing. Tourism statistics is one of the domains in which the opportunities are rather clear as the properties of the data correspond to the nature of the tourism activities. Inbound, outbound roaming and domestic data stored by mobile network operators (MNO) clearly corresponds to the respective inbound, outbound and domestic domains of tourism, however, not without some methodological reservations. The use of mobile positioning data (and also ‘big data’ in general) by users of tourism statistics includes the following expectations:

- Reducing both the burden on the reporting units and the cost involved in statistical processes by (fully or partially) replacing the existing, relatively expensive methods with new data sources;
- Expanding the available options in terms of measuring tourism activities through new indicators;
- Improving timeliness;
- Improving time and spatial accuracy.

The overall opinion and expectations when it comes to the use of mobile data are rather high, and it is generally believed that this data source will increasingly be implemented in the following years in terms of tourism activities and in other domains. National Statistical Institutes (NSI) perceive this data source mostly as complementary, and in some cases it is also seen as a potential replacement for existing data sources and methodologies.

Main Findings of the Study
Access to mobile positioning data is currently very limited mainly because of the regulatory limitations. There are big differences between the EU countries.
The study concludes that there is a need for a central framework for NSIs and other stakeholders in order to obtain the data legally and according to an approved methodology in

order to be able to produce comparable and reliable tourism statistics.

Longitudinal data is a must for reliable tourism statistics in order to assess the whereabouts of the subscribers over a longer period of time (e.g. usual environment, differentiation of the trips by length, identification of overnight stays, etc.).

Based on the outcomes of this study, it can be concluded that at present mobile positioning data can be used as a supplement rather than as a replacement source of data for the current official tourism indicators required in Regulation 692/2011/EU on European tourism statistics.

However, the use of mobile data as a source for tourism indicators introduces several aspects of improvement compared to the existing statistical processes, such as: timeliness (in some cases up to near-real time), access to statistical information previously not available (new indicators), calibration opportunities for existing data, better resolution, and accuracy in time and space.

Mobile positioning data can complement the currently used methods through mixed-mode data collection enabling the sample size of the conventional survey to be decreased.

Other spheres of statistics, as well as disciplines outside official statistics, can benefit from joint processes and indicators generated via mobile positioning data, making the processing and use of mobile positioning data more cost-effective.

The main strengths of mobile positioning data:	The main weaknesses of mobile positioning data:
<ol style="list-style-type: none"> 1. Fairly good consistency over time for the number of trips and nights spent compared to data based on 'traditional' methodologies; 2. Superior coverage for overnight trips when compared to accommodation statistics because mobile positioning data can also cover trips in non-paid or non-registered accommodation; 3. The option to compile more detailed breakdowns by region or country of residence; 4. The option to apply common rules and 	<ol style="list-style-type: none"> 1. Complexity of the access to MNOs' data and uncertain continuity of access to the data in the future; 2. Relative lack of information on the purpose of the trip, expenditure, type of accommodation and means of transport used; 3. Bias between some classifications due to the nature of the data, e.g. over-coverage of the same-day trips due to misclassification of overnight trips; 4. Issues related to the qualitative understanding of tourism that can be

<p>criteria for identifying or classifying specific phenomena like usual environment, definition of secondary destinations, repeat visits, frequency of visits, etc.</p> <p>5. Improved timeliness of statistics (up to near-real time) and possibility to use mobile data as unconfirmed quick indicators;</p> <p>6. Automation level of statistical production;</p> <p>7. Possibility to improve cross-border international statistics (mirror statistics) that reflects the travel network between different countries provided that those countries use mobile positioning data for inbound and outbound tourism statistics.</p>	<p>misclassified due to no understanding of the purpose of the trip (e.g. visiting relatives might not be considered as tourism if the person was asked, but it is classified as tourism according to quantitative standard international criteria);</p> <p>5. Over- and under-coverage issues related to the usage of the mobile phones, e.g. the tourists who do not appear in mobile positioning data at all, who use several mobile devices or who use the roaming service of several MNOs;</p> <p>6. Difficulty to assess the quality (especially accuracy) of statistics based on the mobile positioning data because mobile phone usage during travel is largely unknown.</p>
--	--

4 REFERENCES

European Commission (2009). COM(2009)404 – Communication from the Commission to the European Parliament and the Council on the production method of EU statistics: a vision for the next decade. Link: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF>

Eurostat (2012). Invitation to tender for the supply of statistical services 2012/S 84-137069: Feasibility study on the use of mobile positioning data for tourism statistics. 2 May 2012. Link: http://ftp.infoeuropa.euroid.pt/database/000048001-000049000/000048731_2.pdf