Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics

Eurostat Contract No. 30501.2012.001-2012.452

Report 3a. Feasibility of Use: Methodological Issues

14 April 2014

Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics

Eurostat Contract 30501.2012.001-2012.452

Report 3a. Feasibility of Use: Methodological Issues

Final report

14 April 2014

Authors of this report (alphabetically):

Rein Ahas (University of Tartu), **Jimmy Armoogum** (IFSTTAR), **Siim Esko** (University of Tartu), **Maiki Ilves** (coordinator of this report, Statistics Estonia), **Epp Karus** (Statistics Estonia), **Jean-Loup Madre** (IFSTTAR), **Ossi Nurmi** (Statistics Finland), **Françoise Potier** (IFSTTAR), **Dirk Schmücker** (NIT), **Ulf Sonntag** (NIT), **Margus Tiru** (project coordinator, Positium)

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

Table of Contents

Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics
Report 3a. Feasibility of Use: Methodological Issues

3

1. Introduction

The tourism sector accounts for a significant part of the economy in many European countries. Given the sector's potential in terms of growth and employment, as well as in terms of social and cultural integration, any appraisal of its competitiveness and position requires a good level of knowledge of official statistics regarding the volume of tourism, the characteristics of tourism trips, and the profile of the visitor, and an equally good level of knowledge about tourism-related expenditure and the benefits that can be gained for the economies of the countries in question. The adoption of Regulation 692/2011 by the European Parliament and the Council of the European Union concerning European statistics on tourism is a major step towards a harmonised system of tourism-related statistics for European countries (Eurostat, 2013a).

Two primary sources of tourism statistics are accommodation statistics and statistics about participation in tourism, tourism trips and visitors. For accommodation statistics (tourism supply), a census, cut-off sample or probability sample survey that is carried out with tourist accommodation establishments is usually used. The sample design may vary by country. The key variables in accommodation statistics are the number of persons being accommodated and the number of overnight stays by country of residence and by type of accommodation. Information about participation in tourism (known as tourism demand) is collected by surveying households or individuals. Due to the high cost of such surveys the sample is generally small and therefore estimates that are gained through detailed breakdowns may be of low reliability and the data that is collected from small countries may not be disseminated.

Other sources of tourism statistics are as follows:

- Structural business statistics (SBS) are a rich and comprehensive source of information on businesses in the EU. Data is available at the four-digit level of NACE (the statistical classification of economic activities in the European Community), and includes a wide range of economic indicators; however, most economic activities do not exclusively serve tourists and are not reliant upon them. Eurostat is currently exploring ways of integrating SBS and/or STS (short-term business statistics) data for certain so-called 'tourism characteristic activities' in the tourism statistics databases.

- In the Balance of Payments, travel is one of the main items within the services current account. It differs from other components in that it is demand-orientated, as consumers, 'the traveller', move to the location of the service provider, 'the destination being visited'.

- The Community surveys that have been carried out in terms of ICT usage (Info and Communication Technologies) by households and enterprises are tasked with collecting information on internet usage by Europeans when it comes to preparing or booking travel and accommodation, as well as in the use of ICT and e-business applications by enterprises in the tourist accommodation sector.

- The Flash Eurobarometer surveys, 'Attitudes of Europeans towards tourism', studies the different aspects of travel and visitors. The surveys are carried out in all EU Member States.

- Passenger transport plays an important role in tourism, and vice versa. While certain aspects of transport can be both tourism-related and non-tourism related, other segments (for instance air transport) can be entirely linked to tourism activity.

- Border surveys are used in several countries to collect information about inbound tourism, the characteristics of visitors and tourism trips. It is a useful source of information for countries belonging to the Schengen Area as it provides information about the number of trips/visitors entering or leaving a country by crossing the land border.

Although there is quite a large range of harmonised tourism statistics indicators, some topics are still not covered or could be improved. For example, the total volume of inbound tourism (tourism supply), including the volume of accommodation below the threshold and accommodation without charge (non-rented accommodation), border-crossing, regional tourism data, reliable detailed tourism demand data, same-day visits, efficient production and the high quality of tourism statistics, etc.

1.1. Aims, Content and Structure of Report 3a

One possible data source that could be used in the production of tourism statistics are mobile positioning data. Mobile positioning data describes mobile phone usage and as such is not directly meant to be used in official statistics. The challenges that are presented by using mobile positioning data in official statistics are similar to those that one faces when starting to

use a new administrative data source - concepts and definitions, population frame, the representative nature of the sample, etc.

The aim of this report (according to the Terms of Reference) is as follows:

- To conduct an analysis of those issues that are related to data collection and compilation (sampling design, stratification, and calibration in a sampling frame composed from a gigantic number of mobile positioning observations);
- To conduct an analysis of the quality of the data that has been obtained (on a systematic and sampling bias) when compared to traditional (i.e. currently used) data collection techniques for European statistics on tourism;
- To conduct an analysis of the data that is made available by mobile operators and the feasibility of translating tourism statistics definitions into record selection algorithms (discriminating between those flows that are relevant for tourism statistics and the rest within one's usual environment), and of the impact of such an algorithm on the quality and continuity of the series and on the coherence of data across the European Union;
- To conduct an analysis of the possible relevance of mobile positioning data for related or other fields of official statistics (e.g. the travel item of the Balance of Payments), including the possibility of being able to use joint procedures/methodologies;
- To develop a methodology that is based on technology (to provide recommendations, guidance, best practices, and standards).

The report is divided into five sections: introduction, methodology, quality, relevance for other fields of official statistics, and conclusions. In the methodology section, a description of mobile positioning data is given with an overview of the steps that have been carried out while processing such data. Then the described methodology is evaluated by using different quality aspects in the quality section. The next chapter addresses the possible relevance of mobile positioning data for other fields of statistics and provides an assessment of how well-suited the described methodology is for the usage of mobile positioning data in other fields. The report ends with conclusions.

1.2. Background of the Report

This report is part of a larger study whose goal is to assess the feasibility of using mobile positioning data for the production of tourism statistics. There are six partners

collaborating in the study. The whole study consists of five reports that address the main objective from various aspects. These reports are:

a) Report 1 - Stock-taking contains 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. Report 1 provides examples of methodological insight that have been gained from the usage cases that have been described.

b) Report 2 - Feasibility of access provides a description of the regulatory, business and technological aspects of data accessibility. Technological and some methodological aspects are provided that are relevant to the current task specifically concerning the data source, plus the characteristics and preparation of the data before tourism-specific processing of that data is carried out.

c) Report 3a (this report) - Feasibility of use, methodological issues: This task provides a methodology for the production of tourism statistics by using mobile positioning data. A detailed description of the production process is given. An evaluation of the quality of the described methodology is made.

d) Report 3b - Feasibility of use, coherence: Report 3b assesses the coherence of tourism statistics acquired from various sources (mobile positioning data, accommodation statistics, household and individual surveys, transport statistics, etc.). Tourism statistics from several countries will be analysed. An evaluation of mirror statistics will be carried out and the possible usage of mobile positioning data to increase coherence will be analysed.

e) Report 4 - Opportunities and benefits: this task concentrates upon the potential opportunities and benefits that the usage of mobile positioning data can bring to tourism statistics. In Report 4, the consortium does not collect or research new data and information, but rather integrates the results from previous tasks into a structured and coherent assessment of the potential usage of mobile positioning data in the field of tourism.

This report covers the objectives of Report 3a. The content of this report is based upon the knowledge and experience of the partners and the knowledge that is gained from surveying and interviewing organisations that have been using mobile positioning data in tourism or other domains. The current task serves as input to the subsequent Report 3b and Report 4 and also has references to Report 2.

1.3. Concepts and Definitions

The concepts used in the current report mostly follow the 'Methodological Manual for Tourism Statistics' (Eurostat 2013a). However, there are some concepts that have a slightly different interpretation here. In addition, there are some new concepts introduced. The most important concepts together with the definition that has been used when applied in this report are listed below. Figure 1 illustrates the link between the official tourism concepts and current data sources.
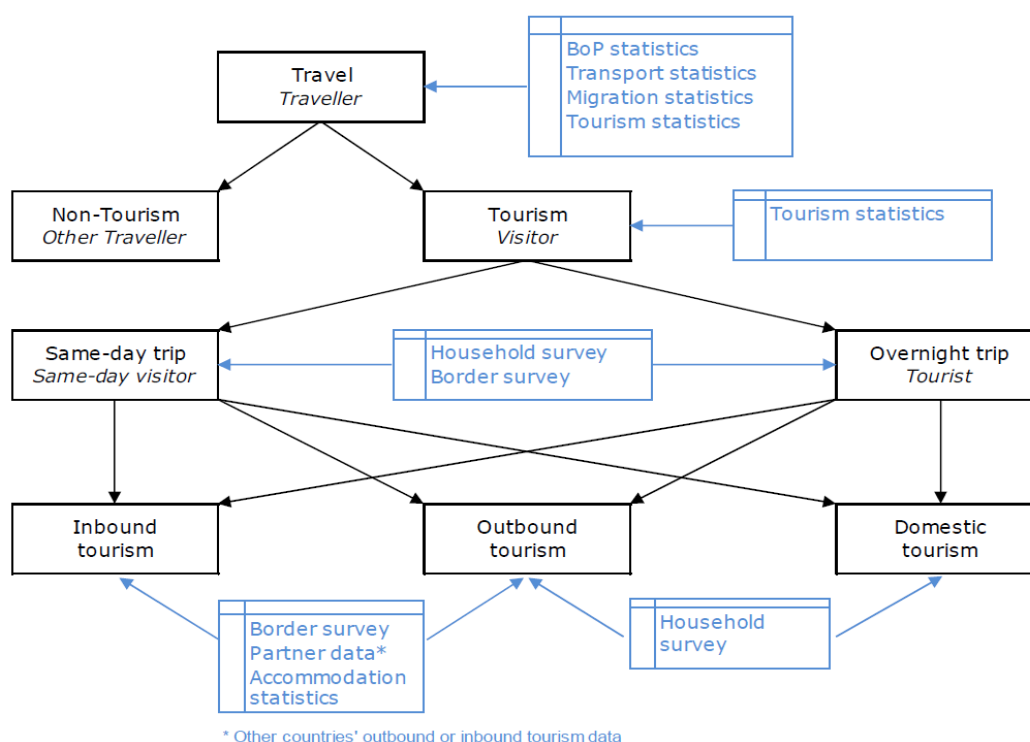

Figure 1. Scheme illustrating concepts in tourism statistics and the sources in which these concepts are used (EUROSTAT 2013a).

Travel - refers to the activity of travellers (which is similar to the official definition).

Traveller - someone who moves between different geographic locations, for any purpose and for any duration (which is similar to the official definition).

Tourism - the activity of visitors who are taking a trip to a main destination which is outside the usual environment, which lasts less than a year, and which is for any main purpose, including business, leisure or other personal purpose, other than being employed by a resident entity at the location that has been visited. The main characteristics are similar to the official definition, although persons who are employed by a resident entity in the location that has been visited cannot be excluded from mobile positioning data.

Visitor - a traveller taking a trip to a main and/or secondary destinations outside their usual environment, for less than a year (differs from the official definition - see Section 3.1).

Tourist (overnight visitor) - a visitor whose visit includes an overnight stay (which differs from the official definition - see Section 3.1).

Overnight visitor/visit/trip - as for a tourist, the term is used to specifically distinguish visits that focus on the duration of the stay at a specific place in a country (new concept).

Same-day visitor/visit/trip (excursionist) - a visitor whose visit does not include an overnight stay (similar to the official definition).

Inbound tourism (tourism form) - comprises the activities of a non-resident visitor within the country of reference on an inbound trip (similar to the official definition).

Domestic tourism (tourism form) **-** comprises the activities of a resident visitor within the country of reference either as part of a domestic trip or part of an outbound trip (similar to the official definition).

Outbound tourism (tourism form) **-** comprises the activities of a resident visitor outside the country of reference, either as part of an outbound trip or as part of a domestic trip (similar to the official definition).

Internal tourism (tourism category) **-** comprises domestic tourism and inbound tourism, that is, the activities of resident and non-resident visitors within the country of reference as part of domestic or international trips (similar to the official definition).

National tourism (tourism category) **-** comprises domestic tourism and outbound tourism, that is, the activities of resident visitors within and outside the country of reference either as a part of domestic or outbound trips (similar to the official definition).

International tourism (tourism category) **-** comprises inbound tourism and outbound tourism, that is, the activities of resident visitors outside the country of reference either as a part of domestic or outbound trips and the activities of non-resident visitors within the country of reference on inbound trips (similar to the official definition).

Country of reference - a country that is linked to the forms (inbound, domestic, outbound) and categories (internal, national, international) of tourism. Outbound tourism from the country of reference to a foreign country is inbound tourism for a foreign country (new concept).

Foreign country - a country outside the country of reference in respect of the forms (inbound, outbound) and categories (internal, national, international) of tourism. Outbound tourism from a foreign country to the country of reference is inbound tourism to the country of reference (new concept).

Trip - refers to the journey of an individual from the time at which that individual departs from their place of residence until they return; it therefore refers to a round trip. A trip is made up of visits to different places. Trips consist of one or more visits during the same round trip (similar to the official definition).

(Tourism) visit - refers to a stay in a place visited during a tourism trip. The stay does not need to be overnight to qualify as a tourism visit. Nevertheless, the notion of stay supposes that there is a stop. Entering a geographical area without stopping there does not qualify as a visit to that area. It is recommended that countries define the minimum duration of stops to be considered as being tourism visits. The concept of a visit depends on the level of the geography in which it is used. It can mean either the whole tourism-related trip or only a part of it, depending on the perspective (origin-based or destination-based) (similar to the official definition).

Overnight stay - the criterion to distinguish tourists (overnight visitor, overnight visits) from same-day visitors. A visitor is considered to have had an overnight stay/visit in a place if the visitor is believed to have stayed there during a change of calendar dates (a place in which a night is spent regardless of the actual rest/resting place). If during a change of dates, a visitor is in the middle of moving between Points A and B within a country of reference, a night might be assigned (depending on the national criteria of the specific country) to Point A, Point B, or it might not be assigned at all. However, from the perspective of country (the place is the country of reference), a visitor spent a night within a country (which differs from the official definition - see Section 3.1).

Trip section - a trip consists of the stay and movement sections. All sections (stay and movement) are aggregated into stay sections if viewed from a higher geographical level. Stay section in place A; movement section between A and B; and a stay section in place B in country X combine a single stay section when viewed from the country level (new concept).

Duration of the trip/visit/stay - mobile positioning provides a means to measure the duration of the visit in total hours, days present, nights spent. Duration of travelling to and from the destination can be identified and excluded. Total hours and nights spent per trip can be

summarised for all aggregation levels; however, days present cannot be summarised (which differs from the official definition - see Section 3.1).

Country of (usual) residence - the country in which a person spends the majority of the year.

Place of (usual) residence - the geographical location of the person's place of residence. In cases in which this is a foreign country, the place of residence is the country of residence as it is not possible to determine a more accurate/specific location of the residence within a foreign country when using mobile positioning data. In the case of the country of reference, the place of residence is a specific location within the country of reference with accuracy depending upon the method of identifying the location's actual point (e.g. the smallest administrative level, the smallest identifiable grid unit, or geographical point).

Usual environment - each form of tourism has a specific definition and method of defining the place of residence and usual environment. By default the place of residence and usual environment for subscribers of inbound data is the foreign country of the subscriber unless identified differently. For domestic and outbound subscribers, usual environment can be defined with precision of country of reference, county, municipality or some other geographical areas or administrative units. The level of detail used depends on the data available and producers' needs (which differs from the official definition - see Section 3.1).

Main destination - the main destination of a tourism trip is defined as the place visited that is central to the decision to take the trip. However, if no such place can be identified by the visitor, the main destination is defined as the place at which they spent most of their time during the trip. Again, if no such place can be identified by the visitor, then the main destination is defined as the place that is the farthest from the place of residence. In mobile positioning data, a distinction between the main destination for a trip (which is similar to the official definition) and a secondary destination (a new concept) has to be made. The main destination for the trip can be identified using the official criteria; however, during overnight trips, each day might have a different main destination (which is usually the one at which the night is spent). On a same-day trip, the main destination for a trip is the place in which most of the time was spent. A visitor can visit one main destination and several secondary destinations during one day.

Secondary destination/visit - as opposed to the main destination, a secondary visit is a place to which a visitor makes a visit (stays) in addition to the main destination for a period longer than the minimum duration of stops to be considered as being tourism visits (new concept).

Transit pass-through - as opposed to main destination and secondary destination, a transit pass-through is the place that visitors pass through or stop during a period of time that is less than the minimum duration of stop to be considered as being tourism visits. A transit pass through does not count as a tourism visit. At a country level, transit pass-through or transit trips/visits are considered as being trips for which the purpose is passing through that country on one's way to or from the country that is their main destination (similar to the official definition).

2. Methodology

The purpose of this section is to provide an overview of what is considered as being 'mobile positioning data' in the context of this study and to describe the methodology for data collection and processing. When it comes to the current study, the technological section (Section 4) of Report 2 concentrates upon initial data sources, the technologies that can be used to extract the data from Mobile Network Operators (MNOs), and the preparation of the data for further processing, specifically for tourism statistics (see Figure 2). The full chain of data processing for tourism statistics from data prepared by MNOs to the calculated tourism estimates is carried out in several steps and is described in the current report.

The population of interest for tourism statistics in the country of reference is all tourism visitors who are travelling within, into or out of this country (domestic, inbound and outbound tourism visitors). The sampling frame that corresponds to this population is different for each statistical domain (inbound, domestic, outbound) and therefore needs to be defined separately according to the domain that is under consideration.

The basis for all sampling frames is a registry of activities of subscribers of MNOs with the country of reference. This registry includes all activities of all subscribers. For the purposes of tourism statistics the observation unit needs to be redefined and only subscribers of interest need to be included in the sampling frame. The sampling frame is formed by translating events into trips and eliminating subscribers who are not of interest (see Sections 2.2.1.1, 2.3.1.1 and 2.4.1.1 for a more thorough description).

Depending upon the availability of the data and upon technological availability, all trips within the frame can be analysed where they correspond to the situation shown in the census or, alternatively, a subset of observations in the frame 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 does not depend upon the number of observations in the sample. The sample size can be determined from available technological capabilities and disclosure rules. The aspects of cost and burden are discussed in the respective chapter of Report 4.

If a subset is used, it is recommended to choose this subset according to probability sampling techniques, as non-probability samples introduce more complexity into the estimation phase later on. When a sample is used, and not a census, then sampling weights need to be taken into account when carrying out estimation. All standard sampling books cover the estimation under various sampling designs, so in this report computation of weights and estimation with weights is not described. Similarly to frame formation, data processing steps are described separately for three forms of tourism.

The specifics of methodology might depend upon the characteristics and origin of the data (time and space frequency, the geographical accuracy of the events, and the available attributes of the events or the subscribers). The processes described here (trip calculation, identification of usual environment, non-residents, transit trips, duration of stays, etc.) assume that longitudinal calculations for single subscribers are possible.

There are several steps, the complete list depending upon the type of statistics needed, that need to be carried out to transform data that is prepared by one or multiple MNOs so that understandable and usable aggregated results for tourism statistics can be computed.

Frame formation comprises the following:

- Application of trip identification algorithms - identify each subscriber's individual trip to the country with the start and end time of the trip;
- Identifying the population of interest (distinguishing tourism from non-tourism):
    a. Defining roaming subscribers not actually crossing the border and entering the country (inbound, outbound);
    b. Defining residents (inbound, outbound);
    c. Defining the place of residence and the usual environment (domestic);
    d. Identifying country-wide transit trips (inbound);
    e. Identifying destination and transit countries (outbound);

Data compilation comprises the following:

- 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);
- Time and space aggregation of the data (day, week, month, quarter/grid-based (one km^2), LAU-2, LAU-1, country);
- Combining data from various MNOs and computing final statistical indicators.
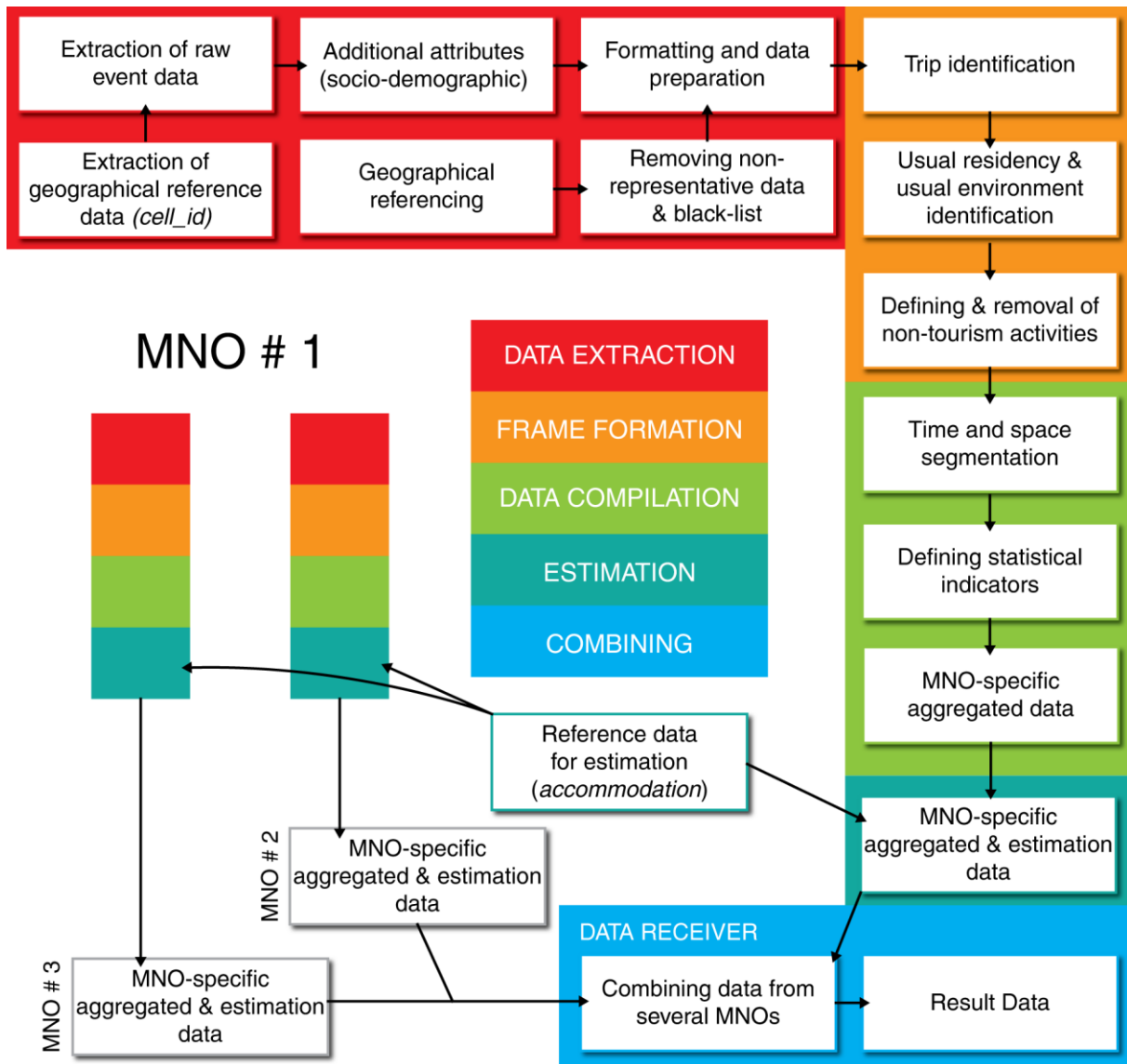

Figure 2. Data processing steps.

Figure 2 illustrates the data processing steps. Each step is explained more thoroughly in the following part of the report.

2.1. Data from MNO

2.1.1. Event Data

The initial data that is extracted and prepared by MNOs is based upon network events that specify a specific subscriber's presence in time and space. The initial processing of the data (its extraction from internal sources, formatting, etc.) is carried out by MNOs and further tourism-specific processing can be handled either within the MNOs' infrastructure or outside it, depending upon the specific situation. A detailed description of the steps carried out by MNOs is given in Report 2 Section 4 and the most relevant points are summarised here.

MNOs can provide either event-based or aggregated data. For the purpose of producing statistics, the use of event-based, unprocessed data as the initial data source is necessary. Aggregated raw data significantly reduces the options and lowers the quality, and a longitudinal analysis cannot be carried out. However, this might be the easiest option when it comes to obtaining the data from MNOs due to privacy restrictions.

With event data as the initial data source, MNOs could either provide the event data for further processing to the NSI or carry out processing according to the predefined methodology internally and provide the aggregated results.

Table 1. The required elements of the events data from mobile positioning data. This is the initial data for processing either internally within MNOs or externally after the MNOs have delivered the data to the NSI.

	Unique identity of the subscriber	**Initial time of the event**	**Country of origin**	**Geographical reference of the event**
Inbound	*subscriber_id* Permanent or temporary, anonymous or not anonymous depending upon the regulatory restrictions	*event_time* Precision of a second	*country_code* Country of subscriber's home MNO	*cell_id/coordinates* CGI-based or other accuracy
Domestic	*subscriber_id* Permanent or temporary, anonymous or not anonymous depending upon the regulatory restrictions	*event_time* Precision of a second	-	*cell_id/coordinates* CGI-based or other accuracy
Outbound	*subscriber_id* Permanent or temporary, anonymous or not anonymous depending upon the regulatory restrictions	*event_time* Precision of a second	-	*country_code* Country in which the event took place

The prepared event data from MNOs should include elements that are presented in Table 1. Geographical reference data for the CGI (Cell Global Identity) or similar antenna-

based reference data should be included with event data for inbound and domestic data in order to be able to assign a location to the event.

Alternatively, MNOs could conduct geographic referencing of events themselves (see Section 2.1.1.1). Additional attributes might be included with event records (e.g. a description/type of the event, duration, respondent ID code, etc.), additional geographical references (e.g. type, direction, power output, antenna height, etc.), and also with subscribers (e.g. socio-demographical attributes of the subscriber, invoice address, etc.). Additional attributes may benefit the quality of the results but cannot be assumed to be available in all MNOs. A description of such attributes is presented mostly in Report 2 Section 4.1. of the current study and is referenced in this report where it is deemed necessary.

In addition, MNOs should, if possible, either exclude or provide options for the data that covers non-human and other obviously biased mobile devices such as machine-to-machine (M2M) data exchange devices that do not represent human mobility and instead create potential bias. If such elimination is not possible (e.g. it turns out to be very difficult to eliminate M2M devices from inbound roaming data), estimations that are based upon elimination algorithms should be applied in order to identify them.

Although there might be differences in the specifics of the data (such as the frequency of the events, spatial accuracy, etc.), the dataset has to contain an identifier for the subscriber, and a variable indicating the time and location of the event. Here we assume that the identifier used is a permanent unique code and that the event time is expressed in the time zone of the country of reference. The location of the event variable may need more processing and the details of this are given in Section 2.1.1.1. The following methodology is based upon these assumptions.

2.1.1.1. Using a Sample

Analysis of coherence in Report 3b has shown that mobile positioning data can be specifically useful when reflecting change over time in terms of relatively high aggregate levels. Such data can be produced with better timeliness when compared to traditional data sources, therefore preparing excellent grounds for quick indicators for these high levels of aggregation (national level, monthly statistics, a few domains). The use of a sample instead of the complete dataset requires the use of fewer processing resources and therefore could even improve its timeliness. Using sampling also improves the privacy aspect as random sampling

reduces the chances of subscribers being identified. The use of sampling might be considered in the following cases:

- There are insufficient resources to process all of the data taken from MNOs or there is a need to reduce the timeliness of processing. For example, processing 15 billion records per month (with an MNO with 10 million subscribers in a model example provided in Report 2 Section 5.1.) requires sophisticated and resourceful processing capabilities. The use of sampling might reduce the costs of processing significantly while preserving the quality of the outcome and increasing its timeliness;

- An extra layer for privacy protection is needed that makes personal data more difficult to identify.

An example of the use of three random samples in Estonia with higher and lower aggregation levels is presented in Figure 3 and Figure 4.
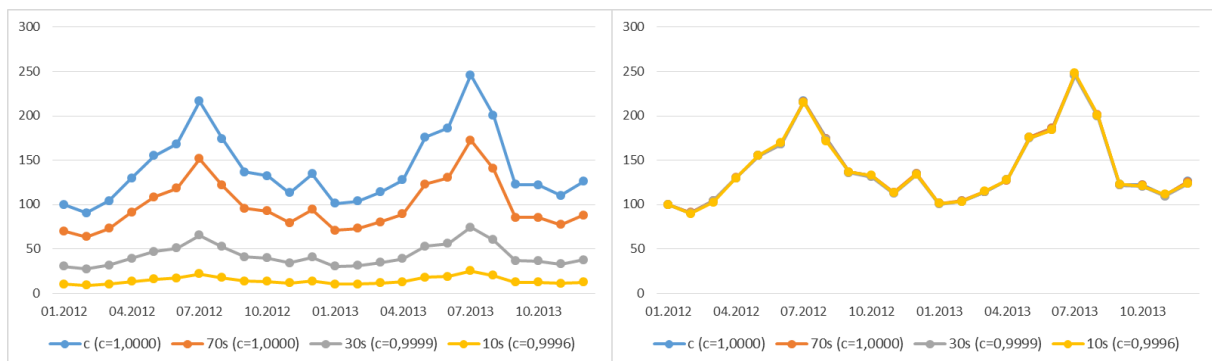

Figure 3. Uncorrected inbound data representing the indexed number of trips made by a single MNO for all nationalities at the national level (Estonia). Left-hand chart presents the indexed values based on the census series (census month 01.2012=100); right-hand chart presents the indexed values based upon individual series values (series month 01.2012=100). Sample size: c=census (100%); 70s=70% random sample; 30s=30% random sample; 10s=10% random sample of all inbound subscribers.
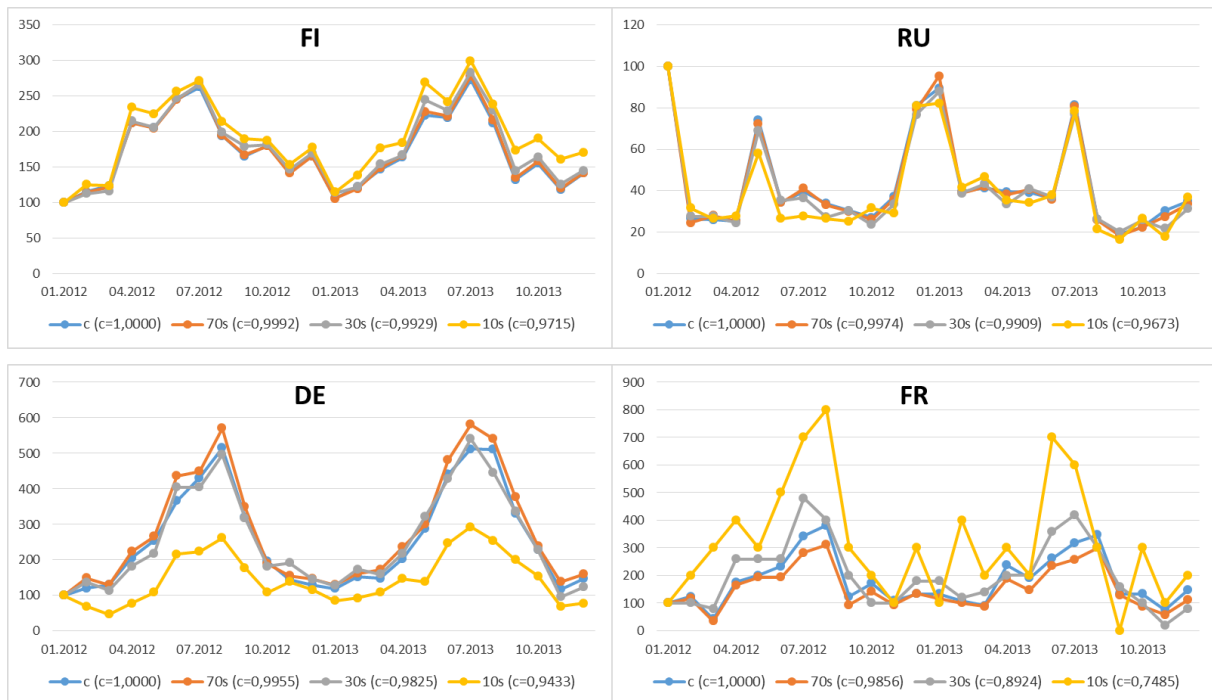
Figure 4. Uncorrected inbound data representing the indexed number of trips made by a single MNO for four different nationalities at the municipality level (town of Tartu, Estonia). Indexed values based on individual series values (series month 01.2012=100). Sample size: c=census (100%); 70s=70% random sample; 30s=30% random sample; 10s=10% random sample of all inbound subscribers.

Mobile positioning provides a wide range of opportunities when it comes to assessing tourism at smaller regional levels and over shorter periods of time (such as in monitoring the number of visitors attending an event or concert), which is not available if one relies only on traditional data sources, because either this data is not available at all or it suffers from a lack of reliability (Figure 4). In this case, however, sampling MNO data is usually not an option due to the very small fractions of segments being focused upon.

2.1.2. Location of the Event (Geographical Referencing)

Unless there is a specific technological advantage in the MNO that assigns GPS or other precision (non-antenna coordinates) location data to each event, an event-antenna data reference has to be produced to get a location attribute for each event. This will result in a dataset in which each event is referenced to a specific antenna (*cell_id* or CGI) with spatial coordinates and/or coverage area of the antenna. Such accuracy is not always available as antennae distribution varies throughout a country - urban areas are densely covered by short working-range network antennae whilst rural areas are covered by few antennae with long working ranges. This causes an effect in which smaller spatial units (the smallest national local administrative units, and smaller grid-based spatial networks of a size such as 500 x 500

18

m) might have no antennae (and therefore no events assigned to it) and other units have very high numbers of events (on account of neighbouring units with no antennae).

Such problems occur with a requirement to present the resulting aggregated data at such small regional levels. This is not a requirement in Regulation (EU) No. 692/2011. According to the Regulation, estimates carried out at the NUTS Level 2 and by type of locality are the most detailed level required for supply-side statistics and at the country level for demand-side statistics. However, in cases in which tourism statistics are required at a more detailed destination level, an estimation that is based upon the probabilistic geographic distribution of events should be used when it comes to deciding which events are assigned to a specific antenna based upon land coverage, the coverage area of the antenna, and other information as described below. This might also be necessary as the place of residence and usual environment calculation might require detailed geographical referencing.
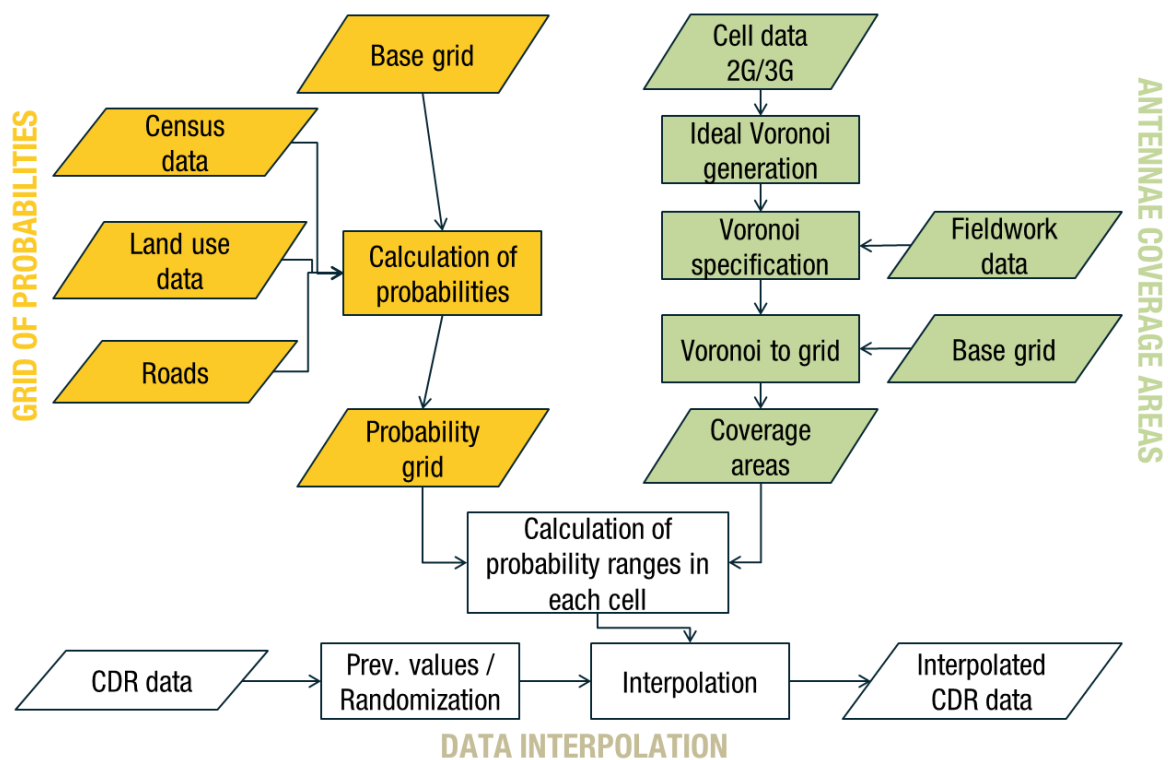

Figure 5. Probabilistic distribution processes of CDR (call detail record) events over the space using base map data and Voronoi polygons. (Positium 2012).

The probabilistic geographic distribution, if carried out correctly, should distribute the events over the coverage area of the antennae so that events are no longer assigned to the location of the antennae but rather distributed throughout the most probable locations (roads, urban area, tourism objects, etc, see Figure 5 and Figure 6). This will not provide the correct location for each individual event, but the overall aggregated result will produce a more

realistic picture (people tend to move around using roads, with such a movement largely taking place within urban areas, and less in fields and in forests).
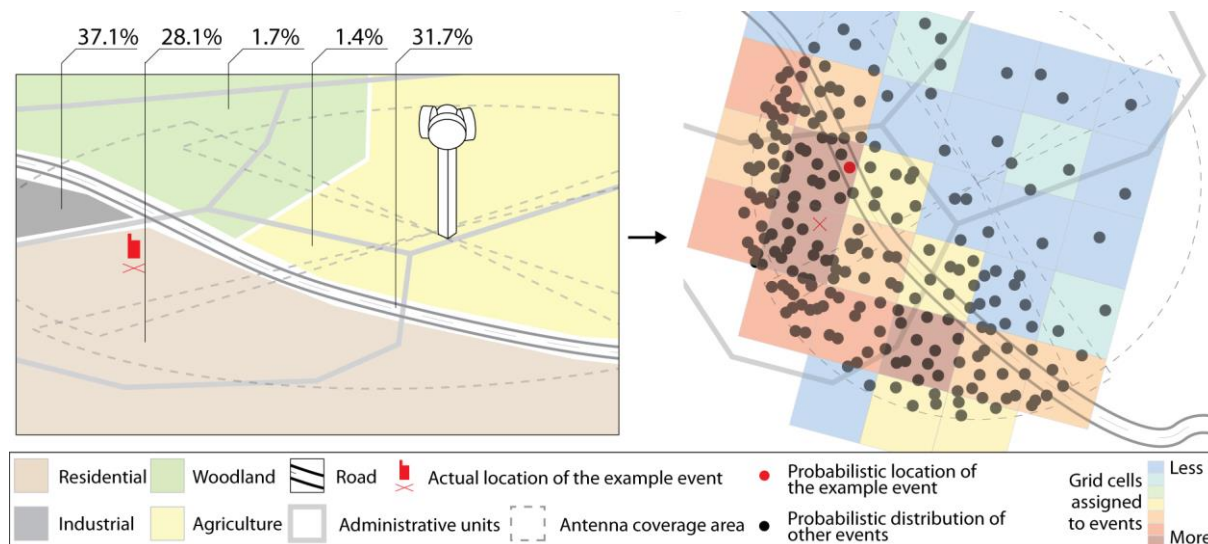

Figure 6. Distribution of events in space based upon land cover and the probability of the human presence in the specific area (Positium 2012).

Probabilistic geographic distribution is a rather complicated procedure with a number of possible bias threats and complicated algorithms. If data results are not actually required to be presented at a level at which such problems occur (in national level statistics), then this process can be revoked because of its complexity, and simple antennae-based geo-referencing can be used where the locations of the antennae are assigned to the events as location attributes.

Specific location is important in inbound and domestic tourism. In the case of outbound tourism the identification of the country level is sufficient. For the identification of the country, the location attribute's value should be retrieved from the code of the MNO providing the roaming service. For example, the code for the roaming partner in Germany (T-Mobile) should be used to extract the country code of Germany. There is the possibility that the geographical location of the phone may be more precisely determined if the roaming partner also provides the CGI of the events and any network coverage information. However, this is usually not important for the specifics of outbound tourism statistics in which a single country is a sufficient location attribute, and in addition Regulation (EU) No. 692/2011 requires only country-level statistics for demand-side statistics.

Figure 7 describes the data that has been prepared by MNOs. In some cases inbound and domestic data can be combined with domestic events that are marked as being local

country in the *country_code* field (e.g. DE for the domestic data from Germany). Such data should be accessible by the consecutive processing system.
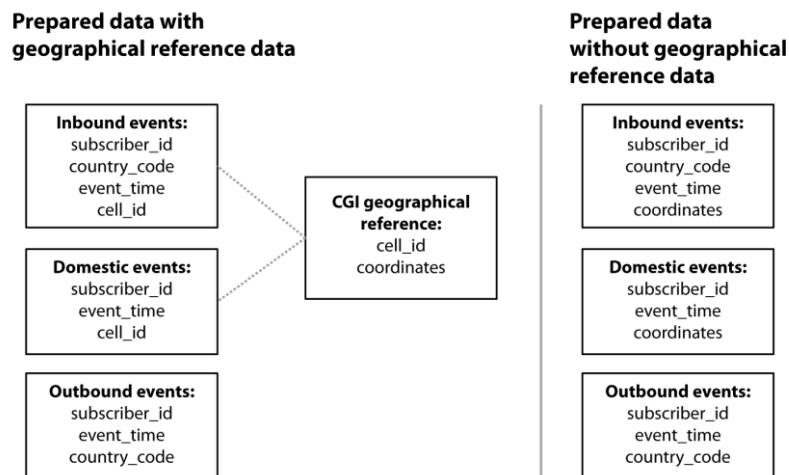

Figure 7. Description of two options of prepared data for inbound, domestic and outbound data.

2.2. Inbound Tourism

The main source for inbound tourism data is inbound roaming data that comes from MNOs. Roaming is the ability to use cellular services while travelling away from a subscriber's home network. This service allows for connecting to the mobile network of the location you are visiting without buying another (local) SIM (subscriber identification module) card. Subscribers are billed by their home network based upon the charges due to roaming. MNOs must establish roaming agreements to govern the exchange of customer billing data for their customers who 'roam' on the visited network. If the visited roaming network is in the same country as the home network, this is known as National Roaming. If the visited network is outside the home country, this is known as International Roaming (or Global Roaming).

While roaming, the roaming network handles subscribers similarly to its domestic (home) subscribers except for identifying the subscriber as the roaming client of their roaming partner MNO and with some possible limitations on services. Therefore, the event data for roaming subscribers is stored the same way as domestic data is stored. Except for national roaming, most of the roaming subscribers in MNOs are foreigners using their home mobile devices within the country of reference (e.g. Finnish visitors using mobile phones during their trip to Estonia). For this reason, inbound roaming data represents a significant amount of actual foreign visitors to the country (inbound tourism).

Over-coverage and under-coverage issues in relation to inbound roaming data are described in Section 3.2.1. The exclusion of parts of the data (specifically the identification of over-coverage in border areas, residence, and the usual environment that can be involved in non-tourism trips and transit trips) is carried out during the frame formation process.

2.2.1. Frame Formation

As presented in Section 2.1 (see Figure 7), the initial inbound roaming data consists of at least four attributes per event:

- Subscriber's identifier (*subscriber_id*);
- Country of origin identifier (*country_code*);
- Time of the event (*event_time*);
- Geographical attribute (either *cell_id* reference or actual location *coordinates*);

If a geographical attribute is not presented as coordinates, then the process of creating the location per event is required. See Section 2.1.2 (Location of the Event) for specific details.

Table 2. An illustrative example of data for one Finnish subscriber (five events).

subscriber_id	country_id	event_time	coordinates
27436527823	FI	28.10.2012 12:53:38	55.92112 18.09451
27436527823	FI	28.10.2012 16:52:38	56.11548 18.58742
27436527823	FI	15.02.2013 22:03:08	55.92112 18.09451
27436527823	FI	17.02.2013 08:33:08	56.12313 18.49385
27436527823	FI	17.02.2013 19:45:18	56.33231 18.25124

For inbound tourism, the frame formation consists of the trip identification process for all inbound roaming subscribers and the elimination of the trips/subscribers whose trips do not correspond with the definition of tourism.

2.2.1.1. Trip Identification

Inbound tourism indicators are based upon trips that have been made by visitors to the country of reference where these consist of visits to specific locations within the country (with these being classified as visits). However, MNOs do not hold any trip-specific information on subscribers. Therefore, being able to identify single trips to the country (from entry to departure) based upon the intervals between the events is carried out instead (see Table 3).

Table 3. An illustrative example of trips that can be identified from data that has been taken from one Finnish subscriber (five events, two trips).

subscriber_id	country_id	event_time	coordinates	trip_id
27436527823	FI	28.10.2012 12:53:38	55.92112 18.09451	1
27436527823	FI	28.10.2012 16:52:38	56.11548 18.58742	1
27436527823	FI	15.02.2013 22:03:08	55.92112 18.09451	2
27436527823	FI	17.02.2013 08:33:08	56.12313 18.49385	2
27436527823	FI	17.02.2013 19:45:18	56.33231 18.25124	2

The same continuous identity of the subscriber throughout the trip is used, so if there is privacy limiting the longevity of *subscriber_id*, then the trip identification algorithm is unable to identify full trips.

As there is no factual information about a person leaving the country, the trip identification algorithm should identify possible entry and departure times based upon the sequential event patterns of the subscribers. The process should identify the time 'gaps' between events and consider longer gaps as the subscriber's absence from the country. The length of the gap that is considered as being the time between trips should be set based upon the characteristics of the data (the more frequent the data, the smaller the gap should be). There is no clear and simple rule, and this has to be set either by local specialists or based upon the initial analysis of the data.

The difficulty lies in finding the most probable length of the gap between events to describe an absence that is realistic. In most cases people go on a few trips during a year and there are long gaps between the trips. The problem arises with frequent travellers and long-term visitors. They have a large number of events over a long period of time with short time gaps between the events. As there is no evidence of their presence outside the country (from the inbound service data for an MNO), nothing is known about the location of the subscribers if there are no events within the network. There are two options: they actually left the country or the device they used was idle (phone turned off or no call activity) while still within the country. The former means that new events after the gap will be considered to be new trips, and the latter means all events before and after the gap should be considered as being the same trips. In both cases the event pattern can be very similar and based upon this, definitive assumptions cannot be made.

A more sophisticated trip identification algorithm can be introduced that analyses each subscriber's individual event pattern (both time and spatial) and assumes, based upon this behaviour, what should be the maximum gap between the events within a singular trip.

The country-level trip identification presented assumes a timeframe for a subscriber's entry to and departure from a country. For lower level visits a stay identification algorithm should be used that is based upon the time and geographical segmentation of the trip. The result of this process is a cross-section of each individual trip with place and time duration information (the duration of visits to each individual geographical unit). Such sectioning should be carried out at the minimum geographical level of the resulting aggregation (e.g. EU former NUTS 5, LAU 2) or lower. Such a result provides the possibility that statistics can be presented at the regional level in the same way that it can be presented at a country level. However, the accuracy of this algorithm depends largely upon the quality of the source data (mainly the geographical distribution of the antennae). The denser the data (in terms of time and space), the better the result will be. The sparser the data, the more gaps will exist between different stays and, as there is no information on the activities, the nearest activities are extended to fill these gaps.

2.2.1.2. Eliminating Subscribers Not Crossing the Border

The network coverage of MNOs usually extends beyond the borders of a country. This means that there is always inbound roaming data for foreign subscribers whose events might be registered by a local MNO, but who actually do not physically enter the country (no crossing of the border is made).

For inbound roaming such cases can be eliminated by applying the non-entry algorithm that identifies the events of single trips that only occur within the coverage of the border antennae reaching beyond the border and therefore assumedly not entering the country. Such cells can be identified by the following:

- geographic antennae coverage analysis;
- semi-automatic recognition of antennae with a large amount of potential non-entry subscribers. This works only in specific countries for which it is possible to recognise a specific group (in terms of nationalities) of subscribers whose events mainly occur only in such antennae (in the case of Estonia this means Filipinos and Indonesians as frequently-arriving seamen aboard cargo ships in

territorial waters which lie close to the Estonian coast and who, in most cases, do not enter Estonian ports or cross the border);

- measuring the antennae coverage areas on the field.

2.2.1.3. Identifying the Country of Residence and the Usual Environment

For inbound data, the foreign country of the subscriber's home MNO is considered as the country of residence unless, based upon the visitation pattern (wherein the total duration of stays within the country of reference exceeds the threshold that is considered for visitors - majority of time for 12 months), the country of reference is determined as being the subscriber's residence, in which case the subscriber's activities in the country of reference cannot be classified as being inbound tourism activities but are instead classed within the scope of domestic tourism. The subscriber's usual environment is closely related to their place of residence.

If the subscriber is considered as being a resident of the country of reference and is processed within the scope of domestic tourism, the question of the subscriber's usual environment is not a subject of the processes that are involved in inbound. If the foreign country is considered as being the subscriber's place of residence, a question of whether part of the country of reference can be considered as being the subscriber's usual environment based upon the frequency of the visits to a specific location within the country (e.g. cross-border shopping). If such a determination can be established, then that part of the country to which any such frequent visits are made should be considered as being part of the subscriber's usual environment and can therefore be excluded from inbound tourism. However, irregular trips outside this environment (to other parts of the country) should be considered as being inbound tourism trips.

The criteria for determining residency and the frequency of trips can be set differently depending upon the domain, purpose, legislation or guidelines. For example, the criterion for the residence can include being present in all months of the previous twelve months or for 183 days for the period or for a 'running' twelve months (the majority of the year - the case for determining residency for the Balance of Payments). Because the data is quantitative, various calculation criteria can be used for different purposes.

It is important to notice that in order to make such assumptions, a long-term unique continuous ID has to be used as the identifier of the subscriber. With short-term changing IDs, it is not possible to identify long-term presence or frequent trips to the country of reference.

Based upon Estonian pilot data, the percentage of long-term or frequent visitors is under 1% of all subscribers in the frame.

2.2.1.4. Transit Trips

Transit visitors enter the country for the purpose of proceeding to the next country without any delay. Defining transit visits is rather complicated as their behaviour is often similar to short-term visitors' behaviour. Two types of transit visits can be defined based upon the spatial level:

- Country-level transit visits;
- Inside country transit visits - passing through places that are not destinations during the trip within the country.

Forms of transit at the country level may consist of one or more of the following:

- Travelling transit (airport and other means of transit through a country);
- Logistical transit with cargo (truck drivers, train engineers);
- Sailors and seamen dealing with freight on ships (within the port area).

Transit visits to a country are calculated by firstly identifying the transit corridors where transit may occur. Based upon these corridors, all trips which remain exclusively within the corridors and within the minimum time needed to pass through the corridor are considered as being transit trips.

2.2.2. Data Compilation

2.2.2.1. The Spatial Segmentation of Trips

Inbound tourism trips to the country consist of one or more visits to certain places. At the country level, a visit is equal to a trip and consists of one visit (the destination is the country of reference). Unless it is a very small country, tourism statistics also require the identification of tourism at smaller spatial levels (e.g. NUTS4/LAU2), based upon the geographical coordinates of the events. This process consists of the segmenting sections of a trip into stay and move sections, based upon the time and place of the events. Obviously, the more events that are recorded per person per time, the more accurate this trip segmentation is going to be when compared to the real-life situation. However, the movements of visitors in between the events are not known and may only be guessed or assumed. In order to identify the main destinations and time of stays (per day) and places at which nights were spent,

segmentation also has to be carried out for time (i.e. stay segments are discontinued at zero hundred hours, i.e. midnight). One main destination can be assigned during a single day. The main destination of a day is the place at which the last event was located (presumably the place at which the night was spent). Other visits to places during that day are considered as being secondary stays (excursions or same-day visits at a smaller scale). For the last day of the trip and for same-day trips, the main destination is the last place at which the subscriber was present (the location of the last event that took place on the trip).

2.2.2.2. Aggregation

The resulting inbound tourism indicators can be aggregated to a specific geographical level or different reference periods. The usual aggregation levels of geography are the country level and the number of sub-country administrative levels. Reference periods can be day, week, month, quarter and year. Because some of the indicators cannot be simply summarised from lower aggregation data, basic visit-specific datasets have to be used to produce aggregation tables for specific levels.

Table 4. Expected end results of inbound tourism indicators from mobile positioning data.

Country level aggregation	Smaller administrative unit level aggregation
Statistical indicators	
Number of trips starting within the country (foreigners' border-crossing inbound).	Number of visits starting to the specific admin. unit.
Number of trips ending within the country (foreigners' border-crossing outbound).	Number of visits ending to the specific admin. unit.
Number of unique visitors to the country.	Number of unique visitors to the specific admin. unit.
Number of days spent in the country.	Number of days present in the specific admin. unit.
Number of nights spent in the country.	Number of nights spent in the specific admin. unit.
Total duration of trips within the country in terms of hours or other time units.	Total duration of visits in the specific admin. unit in hours or other time units.
Average duration of trips within the country in hours or other time units.	Average duration of visits in the specific admin. unit in hours or other time units.
Classification of the above indicators	
Country of origin.	Country of origin.
Duration of stay: same-day/overnight trips to the country (based upon the number of days present).	Duration of stay: same-day/overnight visits to the specific admin. unit (based upon the number of days present).
Duration of stay: number of days present in the country.	Duration of stay: number of days present in the specific admin. unit.
Travel stage: country as main destination or transit pass-through.	Administrative unit as main destination, secondary visit or transit pass-through to the specific admin. unit.
Visit occasion: number of first-time, repeating visits to the country of reference since the beginning of the repeating visitation calculation point in the data/during the specified period of time.	Visit occasion: number of first-time, repeating visits to the specific admin. unit since the beginning of the repeating visitation calculation point in the data/during the specified period of time.

The previously described processes result in a dataset that can be queried for the indicators presented in Table 4. The complete list of variables required by the Regulation (EU) 692/2011 together with comments regarding the possibility of obtaining these variables by using mobile positioning data is given in Annex 1. After the aggregation of data from a single MNO, or from the data provided by all of them, the resultant indicators represent the subscribers and not the target population due to the discrepancies between the frame and population (e.g. not everybody is using mobile phones, some tourists are duplicated, etc. - see Section 3.2.1). Some issues can be resolved or minimised during the frame formation process, but some issues have to be resolved at the estimation stage. Estimation is discussed in Section 2.6.

2.3. Domestic Tourism

The main sources for domestic tourism data include domestic data from the subscribers of a given MNO. A basic assumption is made that domestic data represents a significant volume of the actual trips that have been conducted by the residents of a country within the country of reference (this being tourism and non-tourism trips). This means that under-coverage is assumed not to be a major problem. This is similar to sample surveys in which the best match between population and frame is sought, and the under-coverage problems are generally of a minor nature. However, the other sources of errors are quite different in nature, both in sample surveys and in mobile positioning data. Over-coverage and under-coverage issues that arise in relation to domestic data are described in Section 3.2.1. An exclusion of sections of the data (specifically the identification of non-tourism trips and the residents of other countries) is carried out during the frame formation process.

2.3.1. Frame Formation

As presented in Section 2.1 (see Figure 7), initial domestic data consists of at least three attributes per event:

- Subscriber's identifier (*subscriber_id*);
- Time of the event (*event_time*);
- Geographical attribute (either *cell_id* reference or actual location *coordinates*);

If a geographical attribute is not presented as coordinates, then the process of creating the location per event is required. See Section 2.1.1.1 (Location of the Event) for specific details.

For domestic tourism, frame formation consists of defining the place of residence and usual environment before identifying tourism trips.

2.3.1.1. Identifying the Country of Residence and the Usual Environment

The nature of domestic data is similar to the nature of inbound roaming data. However, the usual environment and place of residence are implicitly identified within the country of reference (unlike the case with inbound data) unless the following applies:

a) the residency of a domestic subscriber can be assigned to a foreign country based upon the outbound roaming activity of that subscriber. When taking into account the outbound data, if a subscriber spends more time abroad than is defined (one year), their data should be excluded from domestic tourism data and should instead be handled as inbound data;

b) it is possible to extend a subscriber's usual environment to a foreign country if the subscriber takes frequent trips to this country (in the form of weekly trips), in which case trips to such a foreign country should not be considered as being tourism trips.

If a person spends more time abroad than in their country of reference, this can be used as a basis for assuming that this subscriber represents the coverage issues described in rows 18 and 21 in Table 12. From an Estonian example, 2.9% of the combined domestic/outbound subscribers spend more time abroad than in Estonia. The events for these subscribers in their country of reference should be handled within the inbound dataset. However, the question arises of whether to follow the somewhat official definition that a resident has to spend one year in the country to be considered a resident.

In order to define the place of residence and usual environment within the country of reference, one approach is to use the anchor point identification model (also called meaningful places identification), which is based upon the long-term time-space patterns of the subscribers. Such a model identifies the most probable home, work-time and other important locations over time. Alternative methods for identifying residence and usual environment can be used; however, all of the methods require a longer data period before meaningful locations can be detected. In addition, a periodic recalculation should be conducted in order to be able to detect changes and carry out the reclassification of meaningful locations (e.g. in the case of historical migration detection, it is possible to detect

a change of home only after a longer period of data collection has become available for subscribers; see 2.8.1).

Based upon anchors or meaningful locations that are calculated using the anchor point model, the place of residence and usual environment can be identified for a specific period of time (subject to being recalculated and reclassified with new data updates as mentioned previously). The specific algorithms that are used can vary depending upon the methodology and definitions that are employed. Anchor points can be identified based upon the duration and/or frequency of the stay in the specific location. For example, the home anchor could be the location at which the subscriber spends most of their workday evenings and mornings; and weekends over a period of time (one month). The combination of such home anchors over a longer period of time (six months) can be used to define differences between permanent and temporary homes (the latter perhaps being holiday or secondary homes). Anchor points can change during time - good practice is to update anchors monthly and also apply changes to historical data. This means that the usual environment is a dynamic description of space over history (people might change their residence and usual environment over time) and this can 'change' the historical data. For example, based upon the data, a new home location (migration) can only be detected from the data after a significant amount of time has elapsed (at the start the new home seems only to be a new travel destination and might not even be significant). After the change has been detected, some of the historical data has to be changed.

The results of the anchor point calculation provide the geographical area (based upon the administrative units or geographical areas), though not necessarily a contiguous one, that describes a usual environment within which an individual conducts one's regular life routines. Based upon the anchor point model (Figure 8), home and work-time anchor points with secondary frequently visited anchor points should form one's usual environment. Holiday homes and secondary homes (summer houses) can be identified from the anchor point model; however, their treatment (within or outside of the usual environment) is a subject of interpretation (e.g. how to treat the migration to a summer house during all three summer months). The analysis of outbound data that can be used in order to identify any frequent trips to foreign countries is also necessary and, based upon that information, the usual environment can be extended. However, as it is not possible to identify the limited geographical area to which subscribers are travelling in foreign countries, if a subscriber travels frequently, a whole foreign country is extended as a usual environment even if the destination locations for such outbound trips are not the same (shopping in a border town inside a foreign country's

border, City A, when compared to a holiday in foreign country's City B in consecutive weeks).
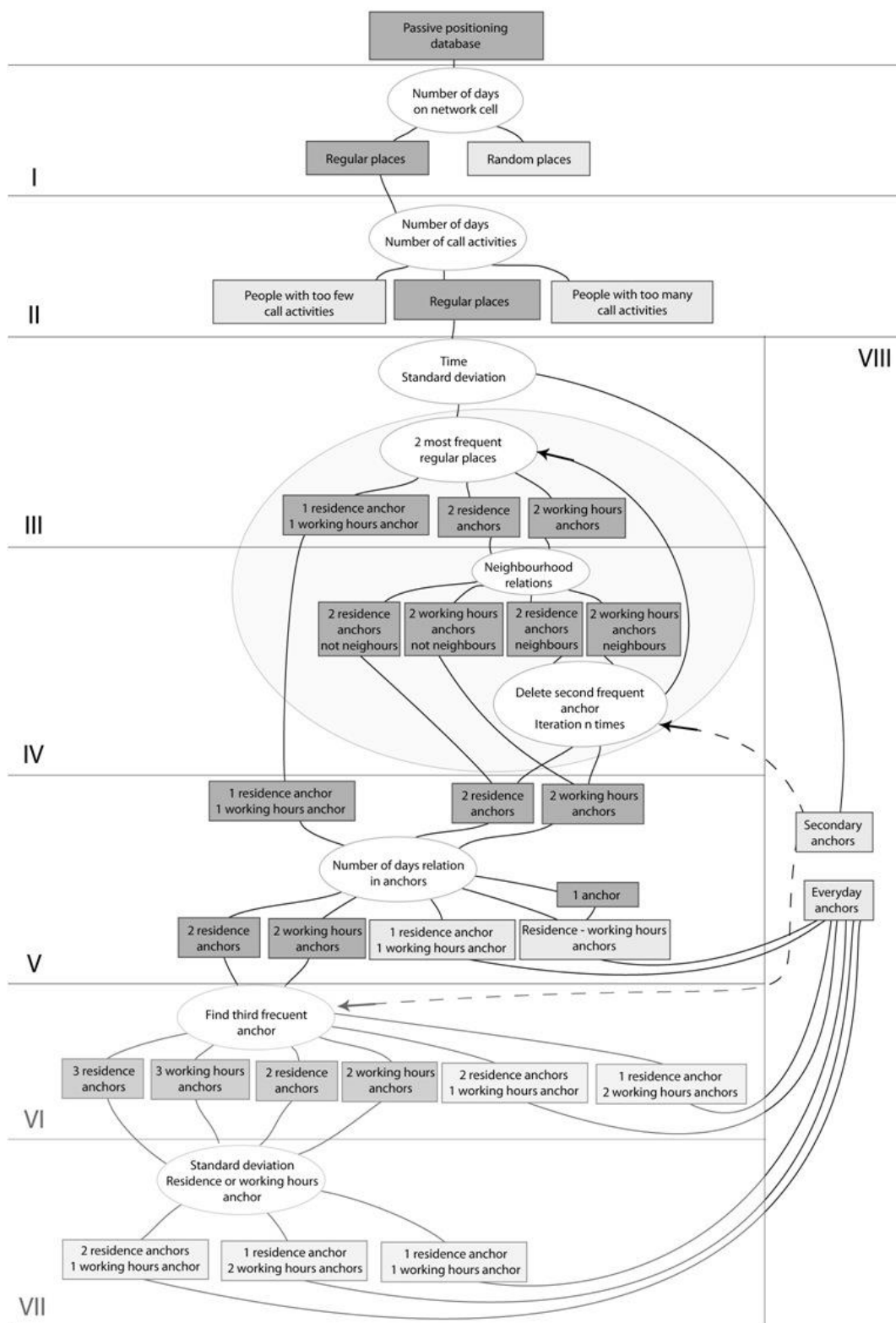

Figure 8. Anchor point determining model (Ahas et al 2010).

The resulting dataset for subscribers' usual environments should be calculated based upon specific periods and should also take into account the change of the place of residence (migration). The following calculations which involve domestic tourism are based upon those trips which were taken outside the usual environment.

The data representation for the usual environment can be limited to an administrative unit or a specific geometry (a buffer zone around the usual travel routes). The current description utilises the administrative unit approach as this is usually the common practice, although it must be noted that tourism-specific trips within this unit will be unaccounted for (large administrative units). When compared to accommodation statistics, in which domestic visitors who spend nights in hotels within their usual environment are still considered as being domestic visitors, mobile data will provide a different methodological outcome as it is not possible at a small scale to define whether the person concerned spent the night at home or in a hotel - the night was spent within the subscriber's usual environment and therefore this is not considered as being tourism activity.

The meaningful places identification should also identify regular trips to places that might be at a substantial distance away or in a different administrative area but are regularly and frequently visited. It is recommended that each country should define the precise definition of what is deemed to be regular and frequent in the context of its tourism statistics and should specify the parameters and criteria that are to be applied in the usual environment identification algorithm.

Because of the peculiarities of the data, it may be impossible to identify the place of residence or usual environment for some subscribers. Because of this reason, many temporary, short-term or otherwise defunct data has to be eliminated as there is no information about such subscribers. From the data contained in the sample pilot, home and regular anchors can be identified for 74% of the subscribers and the rest of the data is useless in terms of tourism.

There are other methodological possibilities that are available when it comes to identifying the usual environment for subscribers as the anchor point model is mainly data-driven and differs from the methodology used for identifying the place of residence and usual environment as proposed by Eurostat.

The Regulation 692/2011 (Regulation 692/2011) and the methodological description by Eurostat (Eurostat 2013a) identifies the usual environment as 'the geographical area, though not necessarily a contiguous one, within which an individual conducts his regular life

routines and shall be determined on the basis of the following criteria: the crossing of administrative borders or the distance from the place of residence, the duration of the visit, the frequency of the visit, the purpose of the visit'. The main interpretation of the usual environment should be defined based upon the subjective feeling of the respondent which is not available in the case of mobile positioning data. Therefore the proposed determination of the usual environment should be based upon the criteria like frequency of the trips, duration of the trip, crossing of the administrative border and distance from the place of residence.

Usual environment could be defined by administrative unit such as, for example, at the second or third level NUTS classification. Also it could be defined by geographical area as a distance or a radius from a residence such as, for example, fifty kilometres from a residence.

The retrospective time window used for determining usual environment could be set for a number of days; for example 45, sixty, or ninety days, for at least four weeks during the time window, with at least one week having more than one day present during the week within the time window.

If the parameters of the criteria are universally agreed upon, it will lead to results which have better comparability between regions and will improve the understanding of the concept. With mobile positioning data, all of the criteria can be implemented. However, this will not be without some issues.

Concerning the identification of the usual environment within the country of reference, either by using the anchor point model or other methods, the success of the model depends heavily upon the specific conditions used in the calculations. As the specific criteria for defining the usual environment is left for the authorities of individual countries, the outcome in different countries can vary largely. Mobile data is very quantitative, and the criteria applied to the identification of the usual environment can be very extensive, despite this following the official guidelines. The criteria suggested firstly for the identification of the usual environment and subsequently for the identification of trips outside this area can be successfully based upon the suggested 'cascade system'. The following criteria for the calculation of the usual environment can be applied:

a) Administrative unit level or other geographical representation (polygons, radius from residence, buffer zone). If the administrative unit or level or geographical area is very large, many trips that are subjectively considered as being tourism trips can be identified as being trips within the usual

environment and therefore not a part of the tourism category. With smaller areas, there is a threat that too many trips are being considered as tourism trips.

b) The frequency and duration of visits to a specific place. The underlying criteria can include any of the following:

- Time window for measuring the presence of the subscribers in specific locations. If longer periods are used as time window, the results include the 'future to be' usual environments while the subscribers were not so connected to the place. Shorter periods can result in trips home being identified as tourism trips during the period of time in which the subscriber was on a longer holiday.

- Retrospective or prospective time window. With continuous data updates, the logical solution is to analyse the data up to the point of its creation (looking only back). However, when re-processing historical data, the results can improve as the retrospective frame can 'ignore' the new usual environments for subscribers who re-locate to new places.

- Measure of frequency. The number of days, weeks the subscriber visited the administrative unit. The definition provided by Eurostat is difficult to implement in terms of quantitative data - 'less than once a week' (i.e. not every week) is rather limited when considering the fact that, during a longer holiday, people tend to visit their homes on less than a weekly basis.

- Number of days a week. The criterion can be ignored, but it can show 'stronger connection' to the place if the person spends more than a day in such a place.

- Length of stay in the places - usually is overridden by the frequency (spends small amount of time, but every day - e.g. visiting kindergarten to pick up the children).

c) The duration over time that the usual environment holds to be in effect during the period of time in which the criteria for the usual environment are not fulfilled. It is very difficult to exactly estimate the moment at which the usual environment ceases to be the usual environment. If the 'effect time' is fairly short, then longer holidays can provide an influence so that the usual environment is not in effect when the person in question returns for a short home visit during their holiday. With a longer impact, a person's usual

environment might actually cease to exist because they have migrated elsewhere but the old usual environment remains in effect (see Figure 9).
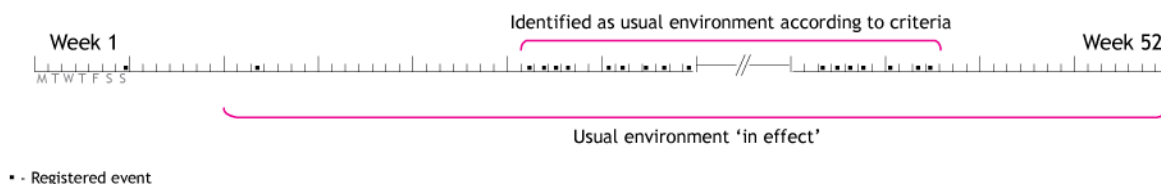

Figure 9. An example of measuring and assigning usual environment to a specific administrative unit for a period of one year. If the 'measured' usual environment (weeks 6 – week 50) applies to an extended period of time (week 3 – week 52), then the registered event or trip to this administrative unit on Wednesday on week 3 is not a tourism trip, and the trip associated with the registered event on Sunday on week 1 is a tourism trip.

These criteria have to be agreed and specified according to local circumstances, and correlated with local reference data, but at the same time remain comparable with other countries (for comparability purposes). For example, the criteria can be phrased in this way: the usual environment of the subscriber comprises the second level administrative units in which a subscriber has been present at least once a week for at least ten weeks during the ninety day retrospective time window. The time influence that has been identified is extended for ninety days prior to and following the identification day.

Because the usual environment can be extended beyond international borders, it is important to include international trips into the usual environment identification. Because it is not possible to identify the specific locations that the subscriber visited in the foreign country, the whole foreign country is considered equal to an administrative unit. If the visits to the foreign country fall under the criteria of the usual environment (i.e. subscriber visits the foreign country on a weekly basis for a specific amount of time - time window), then such a foreign country is considered as a part of the usual environment (see Section 2.5).

Figure 10, Figure 11, and Figure 12 describe the differences resulting in applying different criteria for identifying the usual environment.
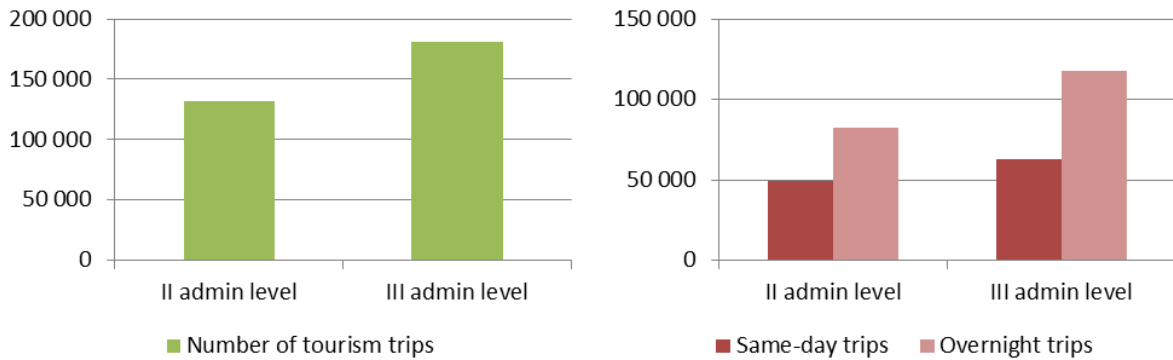
Figure 10. Different results for calculating the number of trips outside the usual environment and distribution of same-day and overnight visits based upon the 50,000 subscriber simulation during 2 years (Nov 2011-Oct 2013) with the usual environment calculated based upon the second and third administrative levels. Other criteria for defining the usual environment (UE): ninety days retrospective time window; at least four weeks during the time window; at least one week with more than one day present during the week within the time window. Criteria for tourism trip: 360 days of the impact of the UE; presence in outside the administrative borders of the UE with duration of at least three hours; trips with a continuation abroad are excluded.
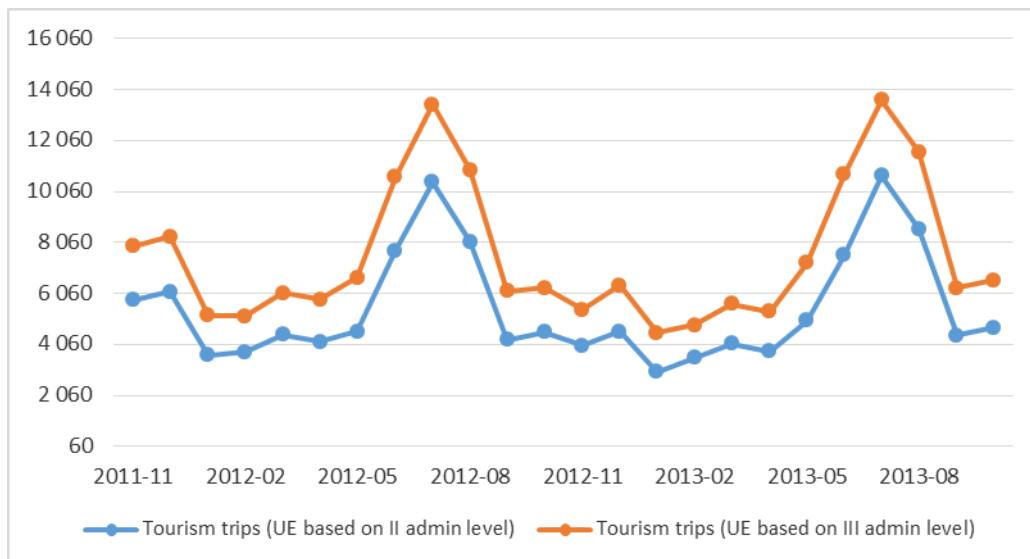
Figure 11. Monthly domestic trips outside the usual environment by 50,000 sample subscribers (Nov 2011-Oct 2013). The number of trips is based upon the different criterion being used for the usual environment (two different administrative levels). Other criteria are the same as in Figure 10.
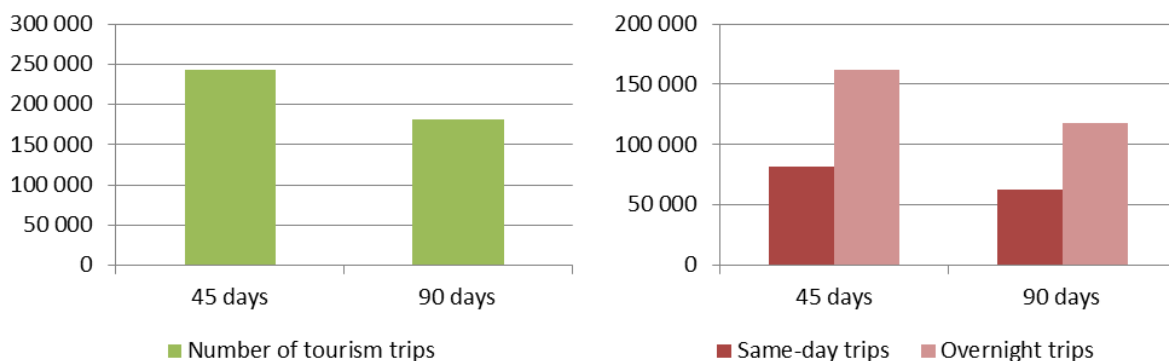
Figure 12. Different results for calculating the number of trips made outside the usual environment and the distribution of same-day and overnight visits based upon the 50,000 subscriber simulation during a period of two years (Nov 2011-Oct 2013) with the usual environment (UE) being calculated by using 45 and ninety days as a retrospective time frame. Other criteria for defining UE: third administrative level; at least four weeks during the time window; at least one week with more than one day present during the week within the time window. Criteria for tourism trip: a full 360 days covering the impact of the UE; presence outside the administrative borders of the UE with a duration of at least three hours; trips with a continuation abroad are excluded.

Because identification of the usual environment should take into consideration longer periods of time than is usually shown in the update period for the data, it is highly recommended that the data is recalculated at least for the period included in such criterion (see 2.8.1).

2.3.1.2. Trip Identification

Domestic tourism trips represent the residents of the country of reference travelling within the country of reference but outside the area of their usual environment. A domestic trip made by a resident begins when that person leaves their usual environment and it ends when the same person returns to their usual environment. Crossing the border of the usual environment is considered to be the beginning of the trip. The trip ends when the subscriber crosses the border upon their return.

Domestic and outbound tourism trips are considered to be national tourism as they represent tourism trips made by the residents of the country of reference. If subscriber IDs for domestic subscribers matches the IDs of outbound subscribers, then these two forms of tourism can be combined and compared.

Though the product of the domestic trip identification algorithms is a trip and visitation dataset that is similar to the inbound trip and visitation dataset, the domestic trip identification algorithms and concepts differ from the inbound process. Inbound data assumes a short periodic presence of the events and refers to the presence of the visitor, while the absence (signified by long gaps between the events) refers to the subscriber not participating

in inbound tourism activities. Domestic trips are calculated by using as a basis the away time from the usual environment and any missing simultaneous outbound data for a specific subscriber. Therefore, the trip identification algorithm mainly has to identify those domestic trips that are taken within the country of reference, outside the usual environment and during the period of time in which no outbound trips are conducted by the subscriber. Parts of domestic travelling outside the usual environment which are made for the purpose of conducting a foreign trip should be concatenated to outbound trips, although during such trips overnight stays in the country of reference might occur. Technically such a combination is rather challenging, but it is possible given that the subscriber IDs for domestic and outbound data are identical. Based upon this, domestic trips that have continuity in foreign countries (outbound data) should either be excluded from domestic tourism and merged with foreign trips or should instead be labelled accordingly for the purpose of using them for other reasons.

During trip identification, domestic events that occur outside the usual environment are isolated and combined into trips. The beginning and end of the trip are assigned to the first and last events of the trip.

2.3.2. Data Compilation

2.3.2.1. Spatial Segmentation of Trips

The spatial segmentation of trips involves exactly the same procedure as with inbound roaming data. Domestic trips consist of one or more visits to places outside the usual environment. At a country level, a visit is equal to a trip and consists of one visit (the destination is the country of reference). Unless it is a very small country, in which case domestic tourism is very difficult or impossible to calculate based upon mobile data, tourism statistics also require the identification of tourism on smaller spatial levels (e.g. NUTS4/LAU2) based upon the geographical coordinates of the events. This process consists of the segmenting parts of the trip into stay and move sections based upon the time and place of the events. Obviously, the more events per person per time, the more accurate this trip segmentation is compared to the real situation. In order to identify the main destinations and time stays (per day) and places at which nights were spent, the segmentation also has to be carried out for time (i.e. stay segments are discontinued at zero hundred hours, or midnight). One main destination can be assigned during a single day. The main destination of a day is the place at which the last event was located (presumably the place at which the night was spent). Other visits to places during that day are considered as being secondary stays (excursions or

same-day visits at a smaller scale). For the last day of the trip and for same-day trips, the main destination is the last place at which the subscriber was present (the location of the last event to take place on the trip).

2.3.2.2. Aggregation

Resulting domestic tourism indicators can be aggregated to specific geographical and different reference periods. The usual aggregation levels of geography are country level and a small number of sub-country administrative levels. Reference periods can be day, week, month, quarter and year. Because some of the indicators cannot be simply summarised from lower aggregation data, basic visit-specific datasets have to be used to produce aggregation tables for specific levels.

Previously described processes result in a dataset that can be queried for the indicators presented in Table 5. The complete list of variables required by the Regulation (EU) 692/2011 together with comments regarding the possibility of obtaining these variables by using mobile positioning data is given in Annex 1.

Table 5. Expected end results of the domestic tourism indicators from mobile positioning data.

Country level aggregation	Smaller administrative unit level aggregation
Statistical indicators	
Number of domestic trips starting outside the usual environment.	Number of domestic visits starting in the specific admin. unit.
Number of domestic trips ending outside the usual environment.	Number of domestic visits ending in the specific admin. unit.
Number of unique domestic visitors in the country.	Number of unique domestic visitors to the admin. unit.
Number of days spent by domestic visitors outside their usual environment.	Number of days spent by domestic visitors in the specific admin. unit.
Number of nights spent by domestic visitors outside their usual environment.	Number of nights spent by domestic visitors in the specific admin. unit.
Total duration of domestic trips in the country by domestic visitors outside their usual environment in hours or other time units.	Total duration of visits by domestic visitors in the admin. unit in terms of hours or other time units.
Average duration of domestic trips in the country by domestic visitors outside their usual environment in hours or other time units.	Average duration of visits by domestic visitors in the admin. unit in hours or other time units.
Classification of the above indicators	
Home admin. units (place of residence) of domestic visitors.	Home admin. units (place of residence) of domestic visitors.
Duration of trip for domestic visitors outside their usual environment: same-day/overnight trips (based upon the number of days away from the usual environment).	Duration of stay for domestic visitors in the specific admin. unit: same-day/overnight visits (based upon the number of days spent in the specific admin. unit).
Duration of domestic visitor trips outside their usual environment: number of days spent outside the usual	Duration of domestic visitor visits in the specific admin. units: number of days spent in the specific

environment.	admin units.
-	Administrative unit as main destination, secondary visit or transit pass-through by domestic visitors.
Visit occasion: number of first-time or repeat visits to the country of reference since the beginning of the repeating visitation calculation point in the data/during the specified period of time.	Visit occasion: number of first-time or repeat visits to the specific admin. unit since the beginning of the repeating visitation calculation point in the data/during the specified period of time.

After the aggregation of the data from a single MNO or from all of them, the resulting indicators represent the subscribers and not the target population due to the discrepancies between the frame and population (e.g. not everybody is using mobile phones, some tourists are duplicated, etc. - see Section 3.2.1). Some issues can be resolved or minimised during the frame formation process, but some issues have to be resolved at the estimation stage. Estimation is discussed in Section 2.6.

2.4. Outbound Tourism

The main source for outbound tourism data is the outbound roaming data from MNOs. Outbound roaming works on the same principles as inbound roaming (see Section 2.2) - outbound roaming for one MNO in a country of reference is inbound roaming for another MNO in a foreign roaming country. The conceptual differentiation is that the geographical reference in outbound roaming is limited to the foreign countries and not smaller geographical regions within the foreign countries. The data exchange between MNOs is based upon roaming agreements and usually represents those billable events that occurred during the trip using the roaming service(s) for partner MNOs in the foreign countries. The subscriber's home MNO receives data on the subscriber's events from partner MNOs.

Outbound roaming data represents a significant volume of actual trips to foreign countries. The over-coverage and under-coverage issues that are related to outbound roaming data are described in Section 3.2.1. The exclusion of sections of the data (specifically exclusion based upon place of residence and usual environment, over-coverage in border areas, and in-transit passage through foreign countries) is carried out during the frame formation process.

2.4.1. Frame Formation

As presented in Section 2.1 (see Figure 7), initial outbound roaming data consists of at least three attributes per event:

- Subscriber's identifier (*subscriber_id*);

- Time of the event (*event_time*);
- Country of destination identifier (*country_code*);

2.4.1.1. Trip Identification

Outbound tourism indicators are based upon trips by visitors to foreign countries from the country of reference. Trips consist of individual visits to specific foreign countries. As MNOs do not hold any trip-specific information on trips made by subscribers, individual trips abroad must be identified instead (from departure to re-entry).

The option to compare outbound roaming events to domestic events is of great benefit as there are factual events inherent in a subscriber's presence in their own country of residence. However, this might not be possible if domestic data is unavailable or the IDs for outbound and domestic data do not match. Similar gap-based trip segmentation as used in inbound roaming trip identification (see Section 2.2.1.1) should be used with additional comparison to events in the home country of reference (domestic event data for the subscriber, if possible).

2.4.1.2. Border Bias

As inbound roaming data includes subscribers who use the roaming service of the MNOs for the country of reference without actually entering the country, outbound roaming might include the usage of foreign roaming services near the borders of the country whilst not actually crossing the border. If possible, such trips should be excluded from the dataset. The problem lies in the similarity of such accidental roaming data to actual short same-day visits to places near the borders of the country for the reference in the foreign country. It is possible to identify a number of same-day trips when only one event, or only a few of them, were registered in a neighbouring country and if there are also some domestic events that have been registered on the same day, it might be assumed that some of these trips were instead accidental use of the foreign roaming service and not an actual trip. Therefore, an estimation of such cases might be used when it comes to assigning a number of such trips to accidental roaming usage so that they can be eliminated from outbound tourism data.

2.4.1.3. Identifying the Country of Residence and the Usual Environment

If a foreign country can be identified as a place of residence or as part of the subscriber's usual environment, data concerning subscriber's trips to such a country should be

excluded from outbound tourism statistics. Such processes and the criteria of exclusion is closely related to domestic data and should be conducted in parallel. By default the place of residence of the outbound subscriber is their country of reference unless the total amount of time they spend abroad exceeds a predetermined number of days during a certain period of time (one year). In such a case, outbound data for this subscriber should be excluded from the outbound tourism processes and the subscriber's domestic data should be processed within the inbound tourism processes. An open question to consider is whether the total duration of foreign visits or the total duration of such visits to a specific foreign country should be used to calculate the residence. A person might spend a long enough time abroad to be considered to be a non-resident of the country of reference; however, at the same time, no specific foreign country might be considered as the actual country of residence.

If a subscriber travels frequently to a specific foreign country, it should be considered whether their usual environment should be extended to such a foreign country and therefore trips to such a country should not be considered as being tourism-related trips. However, as the outbound data does not carry the information about the whereabouts of the subscriber within the foreign countries, the whole country will be considered to be part of the usual environment and actual tourism trips to the country are also excluded (e.g. frequent shopping trips to City A near the border of a foreign country when compared to irregular holiday trips to City B in the same foreign country).

These calculations, however, are only possible if a long-term unique continuous ID is used. With a short-term changing ID, it is not possible to identify the long-term presence or frequency of trips abroad.

2.4.1.4. Defining Destination and Transit Countries

During the visit abroad, there has to be at least one main destination country. When using mobile positioning data, several destination countries can actually be identified based upon different criteria. In some cases it is not possible to identify a single main destination because several countries might present the same quantitative parameters (e.g. the same number of days in different foreign countries). At least one main destination still has to be assigned with other countries being destination or transit countries. In such a case the main destination of the outbound trip is identified by:

- the country with the longest period of stay;

- the furthest country from the home country (if several countries have the same number of days spent);

Countries that are not considered to be main destinations, but that can be considered as being destinations (opposing to transit pass-through countries), can have different criteria assigned. For example, a country might be a (secondary) destination if the subscriber spends at least one night (two days) in the country during a trip to the main destination.

Transit countries are considered as being those that are neither main nor secondary destination countries during the trip. The number of transit countries from outbound roaming is underestimated as phones are most probably used more in destination countries than in transit countries. While the transit country classification should be used for elimination, unless it is totally deleted from the database, compensation should be used to raise the number of transit country visits that might be useful for purposes in other domains (transportation).

2.4.2. Data Compilation

2.4.2.1. Aggregation

Resulting outbound tourism indicators can be aggregated to specific time bases. The usual aggregation levels are day, week, month, quarter and year. Because some of the indicators cannot be simply summarised from lower aggregation data, basic visit-specific datasets have to be used to produce aggregation tables for specific levels.

Previously described processes result in a dataset that can be queried for indicators presented in Table 6.

After the aggregation of the data from a single MNO or from all of them, the resultant indicators represent the subscribers and not the target population due to the discrepancies between the frame and population (e.g. not everybody is using mobile phones, some tourists are duplicated, etc. - see Section 3.2.1). Some issues can be resolved or minimised during the frame formation process, but some issues have to be resolved at the estimation stage. Estimation is discussed in Section 2.6.

Table 6. Expected end results of outbound tourism indicators from mobile positioning data.

All trips outside the country of reference	Specific foreign country
Statistical indicators	
Number of starting outbound trips from the country (the number of outbound border-crossings for residents of the country of reference).	Number of starting outbound visits to a specific foreign country. For a single outbound trip, all visits to a specific foreign country are considered as being one and the first entry is considered to be the starting

All trips outside the country of reference	Specific foreign country
Statistical indicators	
	point for the visit.
Number of ending outbound trips from the country (the number of inbound border-crossings for residents of the country of reference).	Number of ending outbound visits to a specific foreign country. For a single outbound trip, all visits to a specific foreign country are considered as being one and the last departure is considered to be the end point for the visit.
Number of unique outbound visitors from the country of reference.	Number of unique outbound visitors from the country of reference to the specific foreign country.
Number of days spent by outbound visitors from the country of reference.	Number of days present by outbound visitors from the country of reference in the specific foreign country.
Number of nights spent by outbound visitors from the country of reference.	Number of nights spent by outbound visitors from the country of reference to the specific foreign country.
Total duration of outbound trips from the country of reference in hours or other time units.	Total duration of outbound visits from the country of reference to the specific foreign country in hours or other time units.
Average duration of outbound trips from the country of reference in hours or other time units.	Average duration of outbound visits from the country of reference to the specific foreign country in hours or other time units.
Classification of the above indicators	
-	Foreign country
Duration of outbound trip: same-day/overnight trips (based upon the number of days away from the country of reference).	Classification by number of days present during the outbound trip during the visit to a specific foreign country: same-day/overnight visits (based upon number of days spent in the specific foreign country).
Duration of outbound trip: number of days away from the country of reference.	Classification by number of days present during the outbound trip during the visit to a specific foreign country: number of days spent in the specific foreign country.
-	Travel stage: main destination, transit visit to the specific foreign country.
Visit occasion: number of first-time, repeating outbound visits since the beginning of the repeating visitation calculation point in the data/during the specified period of time.	Visit occasion: number of first-time, repeating outbound visits to the specific foreign country since the beginning of the repeating visitation calculation point in the data/during the specified period of time.

The complete list of variables required by the Regulation (EU) 692/2011 together with comments regarding the possibility of obtaining these variables by using mobile positioning data is given in Annex 1.

2.5. Combined Approach to Country & Place of Residence, and Usual Environment

As previously mentioned, there is an internal logic of identifying residence and the usual environment for subscribers in each data set (inbound, domestic and outbound). However, due to international mobility of the people, a combined aspect of the identification

of residence should be used as in many cases the trans-border activity affects the geographical concept of dwelling of the subscriber.

From the point of view of the mobile data at hand, three concepts of the country of residence can be identified:

- A foreign country is a country of residence for the subscriber (see Figure 13 and Figure 14 - I.1 and D.1);
- The country of reference is the country of residence for the subscriber (see Figure 13 and Figure 14 - I.2 and D.2);
- The residence of the subscriber is unknown. This, however, has two sub-options:
  - It is known that the country of residence is a foreign country, but there is no sufficient information about what specific foreign country (domestic subscriber is only active in foreign countries, but does not stay in any specific foreign country long enough that it can be considered as a proof for residence, see Figure 14 - D.3);
  - There is absolutely no information about the residence of the subscriber (e.g. very short-time domestic pre-paid card owners, see Figure 14 - D.4).

Concerning the place of residence of the subscriber, there can be three options:

- The subscriber's place of residence is within a foreign country and therefore the place of residence equals the country of residence (because it is not possible to identify the place of residence within a foreign country, see Figure 13 and Figure 14 - I.1, D.1 and following);
- The subscriber's country of residence is the country of reference and the place of residence at some more detailed level can be identified;
- The subscriber's place of residence is unknown.

From the point of view of tourism statistics, the usual environment is more important than the place of residence. The place of residence is part of the usual environment. Concerning the identification of usual environment there are several aspects depending upon the nature of the mobile positioning data and the criteria used for identifying it. There are five options for identification of subscriber's usual environment:

- The country of residence is a foreign country and the usual environment is fully located outside the country of reference (see Figure 13 and Figure 14 - I.1.1 and D.1.1);

- The country of residence is a foreign country but the usual environment extends to some parts in the country of reference (see Figure 13 and Figure 14 - I.1.2 and D.1.2);

- The country of residence is a country of reference and the usual environment is only within the country of reference (see Figure 13 and Figure 14 - I.2.1 and D.2.1);

- The country of residence is a country of reference but the usual environment extends from the country of reference to a foreign country (see Figure 14 - D.2.2);

- It is not possible to identify the usual environment and therefore the subscriber is excluded from any tourism dataset (see Figure 13 and Figure 14 - I.2.2 and D.2.3).

Figure 13 and Figure 14 illustrate the identification options for the country of reference, usual environment and potential tourism destinations of the inbound roaming, domestic and outbound roaming data. Independent of the specific methodology being used to identify the country of reference, usual environment and tourism destinations, the original data can be divided between seven different systematic types of division (I.1.1.1, I.1.2.1, I.2.1.1, D.1.1.1, D.1.2.1, D.2.1.1, and D.2.2.1).
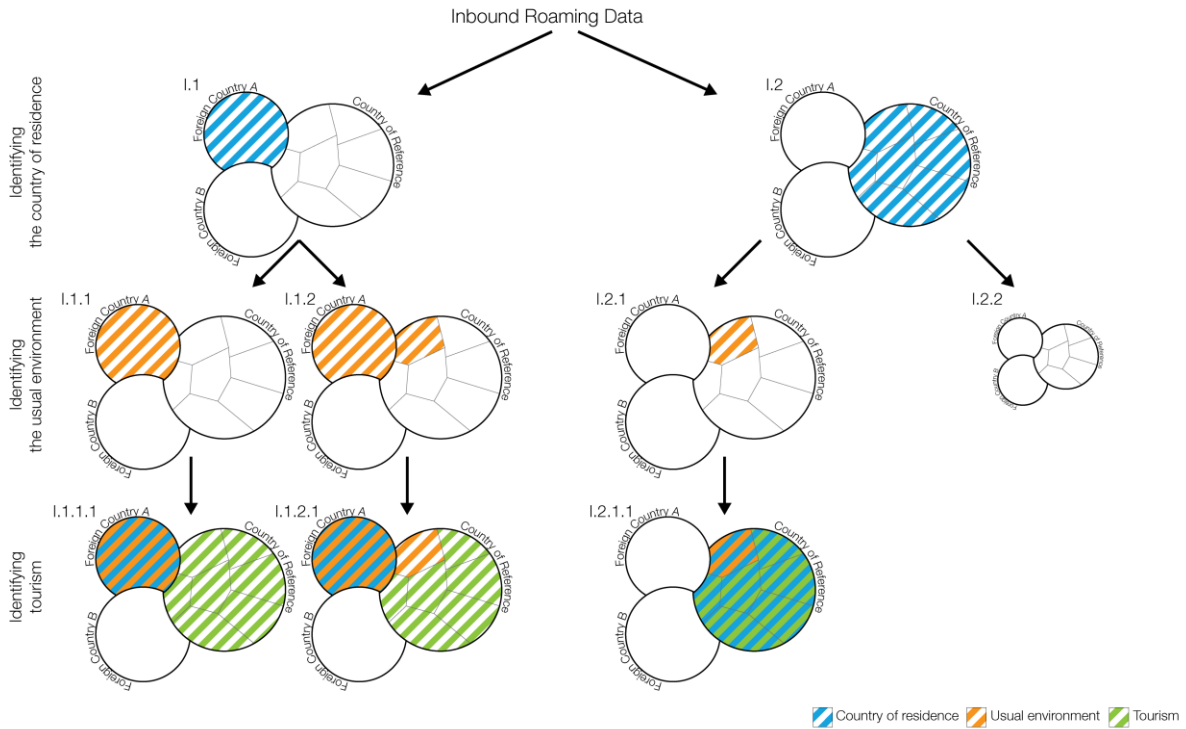
Figure 13. An illustration showing the options available when it comes to defining the country of residence, usual environment and potential tourism destinations based upon inbound roaming data taken from mobile positioning data.
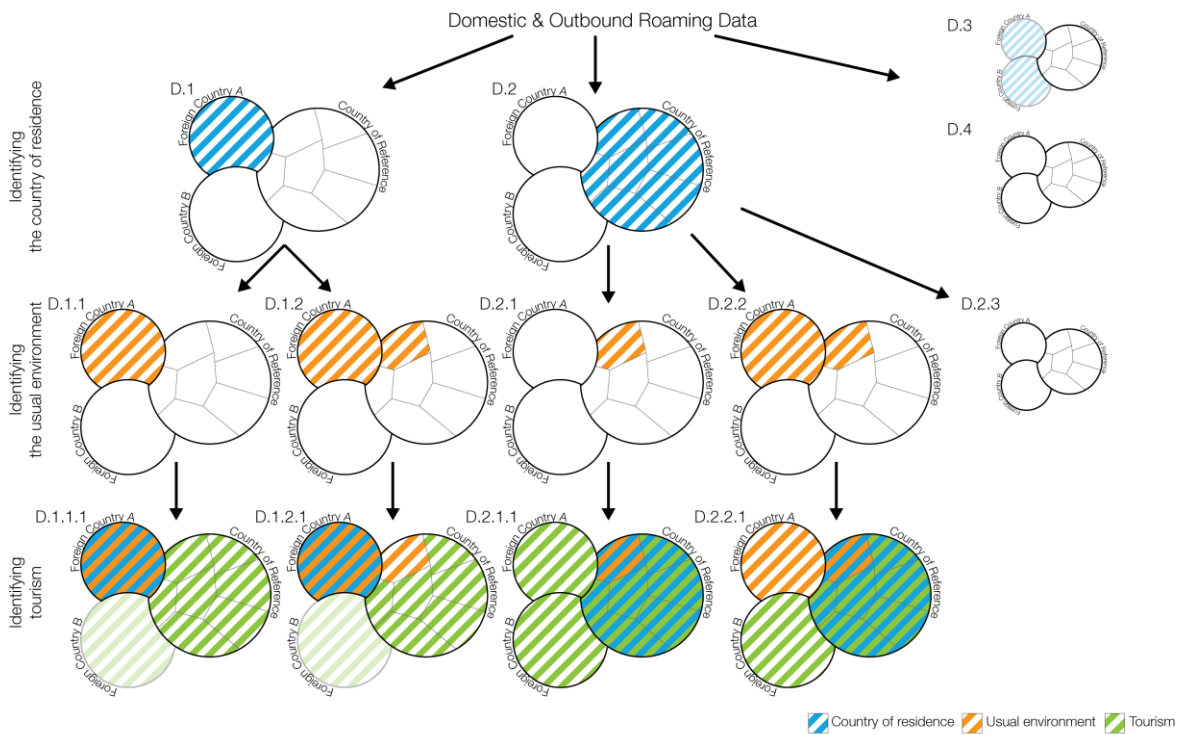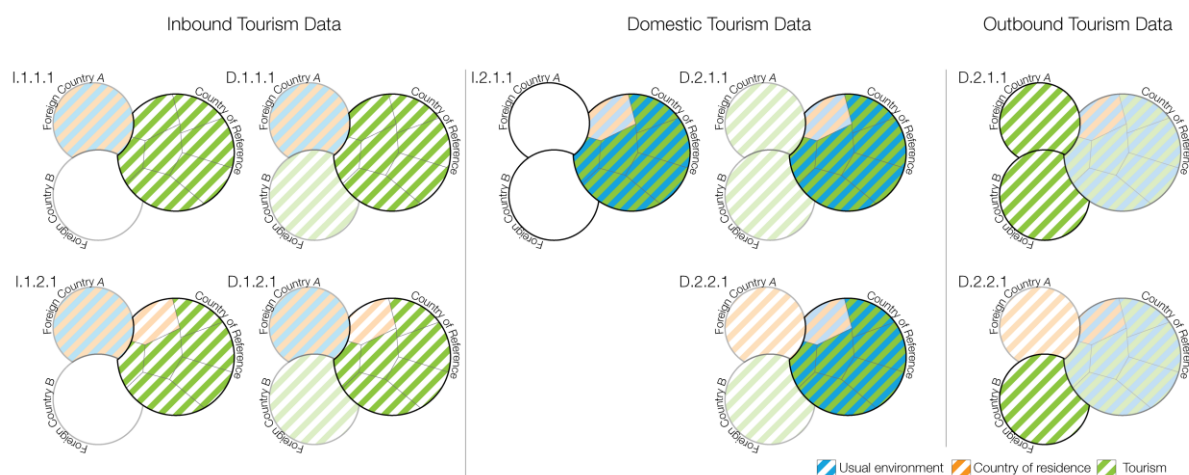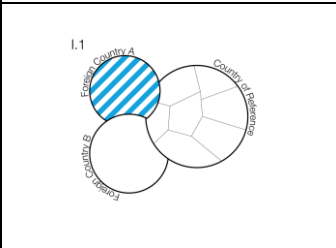

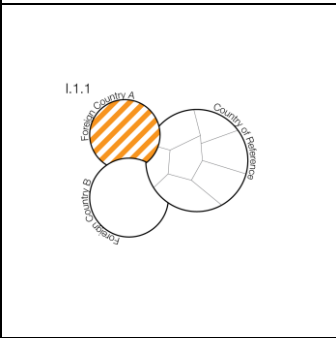Figure 14. An illustration showing the options available when it comes to defining the country of residence, usual environment and potential tourism destinations based upon domestic and outbound roaming data taken from mobile positioning data.
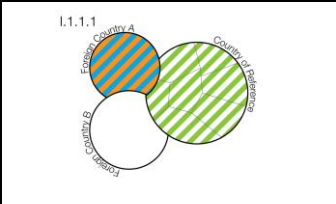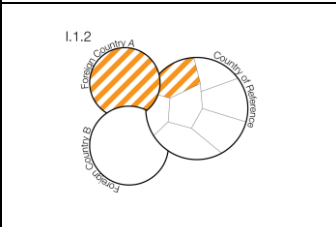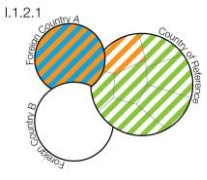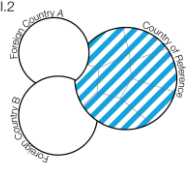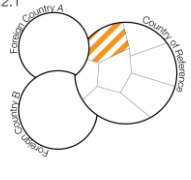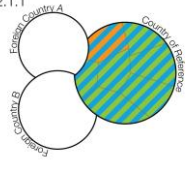
Figure 15. Potential tourism destinations based upon the systematic identification of country of residence and usual environment. The description of the system components are provided in Table 7.

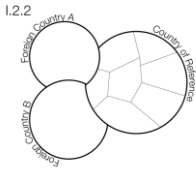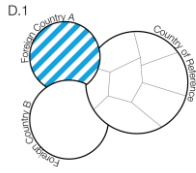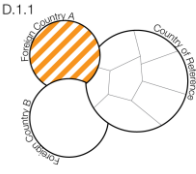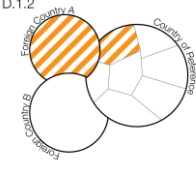Table 7 describes each possible situation in detail.

Table 7. Description of the different options and concepts used to identify the country of residence, usual environment and tourism destinations based upon the initial source data and the type of tourism.

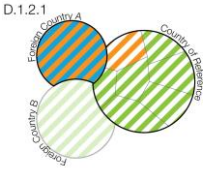|	**Explanation**	**Data source**	**Tourism form**
	Inbound roaming subscriber does not spend the majority of the reference period in the country of reference. This is the majority of the case representing 99.6% of all inbound data based trips from the Estonian example.	Inbound roaming	Inbound tourism
	The usual environment of the resident of the foreign country does not extend to the country of reference. The usual environment is fully identical to the country of residence. This is the case for majority of inbound roaming subscribers – 98.8% conducting 86.6% of the trips to Estonia.	Inbound roaming	Inbound tourism
	The country of residence and the usual environment are outside the country of reference. The country of reference is fully a potential destination for inbound tourism.	Inbound roaming	Inbound tourism
	The usual environment of the resident of the foreign country extends to the country of reference (e.g. frequent visitor to a specific part of the country). Although the number of inbound roaming subscribers who have a	Inbound roaming	Inbound tourism
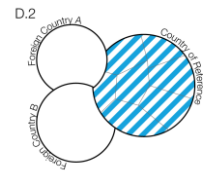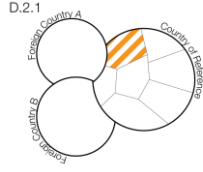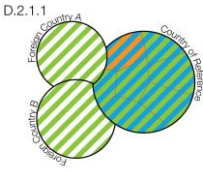
Country of residence / Usual environment / Tourism	Explanation	Data source	Tourism form
	part of the usual environment within the country of reference is low (1.2%), they are representing 13.4% of the trips (case in Estonia).		
I.1.2.1	The country of residence is a foreign country. The usual environment extends from the foreign country to some part of the country of reference. The country of reference is partly a potential destination for inbound tourism when travelling outside the usual environment.	Inbound roaming	Inbound tourism
I.2	Inbound roaming subscriber spends the majority of the reference period in the country of reference. The subscriber is considered as a resident of the country of reference and participant of domestic tourism. This is the minority of the case representing 0.4% of all inbound roaming data based trips from the Estonian example but can be much higher in countries with more international MNOs (e.g. Andorra, Luxembourg).	Inbound roaming	Domestic tourism
I.2.1	The usual environment of the subscriber is identified within the country of reference. This is valid for roughly 65% of inbound roaming subscribers whose country of residence is the country of reference.	Inbound roaming	Domestic tourism
I.2.1.1	The country of residence is the country of reference. The usual environment is a part of the country of reference. The country of reference is the potential destinations for domestic tourism when travelling outside the usual environment.	Inbound roaming	Domestic tourism
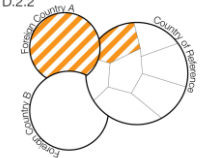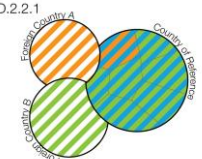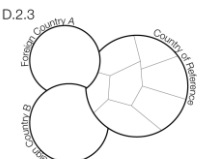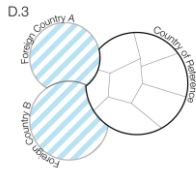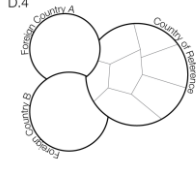
⬚ Country of residence ⬚ Usual environment ⬚ Tourism	Explanation	Data source	Tourism form
I.2.2	Although the country of residence of the inbound roaming subscriber can be identified as the country of reference, it is not possible to identify the usual environment of the subscriber and therefore the subscriber is excluded from any tourism dataset. This is valid for roughly 35% of inbound roaming subscribers whose country of residence is the country of reference.	Inbound roaming	N/A
D.1	Based upon a comparison between the domestic and outbound roaming datasets, the country of residence of the subscriber is a foreign country. From the total trips calculated based on the outbound data, 2% of the domestic subscribers spend more time abroad in a specific foreign country with 13.9% of all initial outbound trips.	Domestic and outbound roaming	Inbound tourism
D.1.1	The usual environment of the subscriber is only in the foreign country/countries. This is the case for 7% of domestic subscribers whose country of residence is the foreign country.	Domestic and outbound roaming	Inbound tourism
D.1.1.1	The country of residence and the usual environment are outside the country of reference. The country of reference is the potential destinations for domestic tourism when travelling outside the usual environment. Although the outbound data also makes is possible to identify visits to other foreign countries, it is not a part of any tourism form for a country of reference (because the residence of the subscriber is already one foreign country).	Domestic and outbound roaming	Inbound tourism
D.1.2	The usual environment of the resident of the foreign country extends to the country of reference (e.g. a frequent visitor to a specific part of the country). This is the case for 93%	Domestic and outbound roaming	Inbound tourism

	Explanation	Data source	Tourism form
	of domestic subscribers whose country of residence is the foreign country.		
D.1.2.1	The country of residence is a foreign country. The usual environment extends from that foreign country to an unspecified part of the country of reference. The country of reference is partly a potential destination for inbound tourism when travelling outside the usual environment.	Domestic and outbound roaming	Inbound tourism
D.2	Based upon a comparison between the domestic and outbound roaming datasets, the country of residence of the subscriber is the country of reference. From the total trips calculated based on the domestic/outbound data, 97.1% of the domestic subscribers are spending more time in the home country representing 75.5% of all initial outbound trips.	Domestic and outbound roaming	Domestic and outbound tourism
D.2.1	The usual environment of the subscriber is identified only within the country of reference. This is valid for 77.7% of domestic subscribers whose country of residence is the country of reference. The percentage is heavily dependent on the criteria used for identifying the usual environment.	Domestic and outbound roaming	Domestic and outbound tourism
D.2.1.1	The country of residence is the country of reference. The usual environment is part of the country of reference. That part that is outside the usual environment within the country of reference is a potential destination for domestic tourism. Foreign countries are the potential destinations for outbound tourism.	Domestic and outbound roaming	Domestic and outbound tourism

Country of residence Usual environment Tourism	Explanation	Data source	Tourism form
	The usual environment of the subscriber is identified within the country of reference and extends to a foreign country (e.g. frequent short trips to the foreign country). This is valid for 3.2% of domestic subscribers whose country of residence is the country of reference. The percentage is heavily dependent on the criteria used for identifying the usual environment.	Domestic and outbound roaming	Domestic and outbound tourism
	The country of residence is the country of reference. The usual environment is part of the country of reference and extends to a foreign country. That part that is outside the usual environment within the country of reference is a potential destination for domestic tourism. Foreign countries outside the usual environment are the potential destinations for outbound tourism. It should be noted that in such case, as there is no information on the exact extent of the usual environment within the foreign country, the whole foreign country is considered as a part of the usual environment and therefore all trips that are made there, even if they are actually tourism trips, are considered as trips within the usual environment.	Domestic and outbound roaming	Domestic and outbound tourism
	It is not possible to identify any usual environment for the subscriber. The subscriber is excluded from any tourism dataset. This is valid for 19.1% of domestic subscribers whose country of residence is the country of reference, but it is not possible to identify any usual environment. Such subscribers are usually very short-term (e.g. very short period pre-paid SIM card usage). The percentage is heavily dependent on the criteria used for identifying the usual environment.	Domestic and outbound roaming	N/A

▨ Country of residence ▨ Usual environment ▨ Tourism	**Explanation**	**Data source**	**Tourism form**
D.3	Based upon a comparison between the domestic and outbound roaming datasets, the subscriber spends more time abroad than in the country of reference, but there is not enough information to identify the specific foreign country. The subscriber is excluded from any tourism dataset. From the total trips calculated based on the outbound data, 0.9% of the domestic subscribers spend more time abroad, but no specific foreign country can be established, representing the total of 10.5% of all initial outbound trips. This proportion is shared with following D.4 option without clear distribution.	Domestic and outbound roaming	N/A
D.4	Based upon a comparison between the domestic and outbound roaming datasets, it is not possible to determine the country of residence (e.g. very short domestic pre-paid SIM card usage). The subscriber is excluded from any tourism dataset.	Domestic and outbound roaming	N/A

2.6. Estimation

As a result of previous processes a basic MNO-specific inbound, domestic and outbound tourism aggregated dataset for statistical indicators has been created with a set of classification attributes describing the nature of the trips and visits. As this dataset does not precisely describe the tourism indicators for the country, but for one of the specific MNO(s) instead, a number of adjustments have to be made in order to provide results that describe the general population of interest (the real number of tourism visits and visitors). The biggest problem is that the precise coverage of the dataset is unknown as there is no other reliable source that covers the total number of tourists in all forms, and in most cases information on the magnitude of individual coverage issues (e.g. the number of visitors not using mobile phones) does not exist. So when relying on expert opinions, statistical models and approximations, an adjustment of the estimates should be carried out in order to avoid the production of biased estimates.

Before any adjustment on estimates is carried out, mapping of data subsets is carried out for which adjustments have to be used in order to convert subscriber-specific data to represent the general number of inbound visitors, and a theoretical basis for calculations has to be established. This part of the work is closely related to the quality issues described in Section 3.

Factors that may cause the bias that need to be taken into account are:

- Coverage of the dataset - various over-coverage and under-coverage issues that influence the accuracy of the estimates described in Section 3.2.1;
- The validity of the concepts used - differences in translating the quantitative measurements and calculations when compared to official definitions (or real life situations). For example, defining non-tourism trips for exclusion based upon qualitative information (e.g. subjective feeling of the respondent concerning the usual environment or the purpose of the travel);
- Measurement problems in the dataset - possible errors or lack of data for some geographical regions or technological problems with data (e.g. systematic geographical data errors, missing data);
- Algorithms used in the data processing - the misinterpretation of concepts due to calculations, either due to incomplete data or imperfect algorithms (e.g. miscalculating the actual user environment, the shorter duration of trips, confusing frequent trips with long-term trips, etc.).

Not all factors can be taken into account for estimation in practice due to the lack of relevant reference data. The reference data that is used to compare mobile-based indicators and produce an estimation model depends upon the availability of the data (see Table 8).

For inbound statistics, an initial comparison should be carried out at least with accommodation and/or any survey statistics. This estimation should be made using the comparison of the number of overnight visits and nights spent to the number of accommodated foreigners and the number of nights spent for different countries. The number of overnight visitors should be equal to or higher than the official accommodation number because many visitors do not stay at the official accommodation during the trip.

For domestic statistics, an initial comparison should be carried out with the accommodation and/or any survey statistics. This estimation should be made using the comparison of the number of overnight visits and nights spent to the number of

accommodated domestic visitors and the number of nights spent. The number of overnight visitors should be equal to or higher than the official accommodation number because many visitors do not stay at the official accommodation during the trip.

For outbound statistics, data from foreign countries concerning the number of accommodated visitors from the country of reference should be compared to the number of visits to specific foreign countries.

For outbound and inbound visits, any reliable reference data covering international trips between specific countries (ferry, air transport) can be used to estimate the total number of international visits.

Other reference data sources can be used for such a comparison. The purpose of this reference is to establish the most realistic basis for estimations that rely upon a comparison between similar representations of tourism in different domains. There is no clear rule or proposition of what kind of reference data can be used as different countries produce different statistics that might or might not be useful for creating such estimates. Possible sources of reference data is provided in Table 8.

The estimation model for the data using various correction coefficients should result in a model for either each MNO separately or all MNOs together, depending upon the technical setup in the specific country.

Table 8. Possible reference data that can be used to calibrate the results of mobile positioning data.

INBOUND	
Reference data	**Data from mobile positioning**
Number of accommodated foreigners in accommodation facilities	Number of overnight visits
Number of nights spent in accommodation facilities	Number of nights spent
Number of border-crossings (entry and/or departure)	Number of trips/visits (starting and/or ending)
Statistics for points-of-entry (airports, seaports, etc.)	Number of trips/visits starting and ending in the specific locations
Border survey, border survey estimations	Number of trips starting from a specific place
Survey covering the number of same-day visits	Number of same-day visits
Foreign countries' outbound data (border survey, household survey)	Number of inbound trips from a specific foreign country, proportion of same-day and overnight visits
DOMESTIC	
Reference data	**Data from mobile positioning**
Number of accommodated residents in accommodation facilities	Number of overnight visits
Number of spent nights in accommodation facilities	Number of nights spent
Household survey	Number of domestic trips/visits, same-day and

	overnight visits
Census data	Statistics that relate to the location of usual residencies and workplaces
Estimation from country-wide domestic tourism surveys	Number of domestic trips/visits, same-day and overnight visits
Transportation statistics	Number of tourism trips/usual trips between places
OUTBOUND	
Reference data	**Data from mobile positioning**
Border survey	Number of outbound visits
Tourism statistical data from other countries (mirror statistics)	Number of outbound visits/nights spent in specific countries
Household survey	Number of outbound visits, proportion between the number of same-day and overnight visits
Data from travel agencies	Number of outbound visits
Cross-border transportation statistics	Comparing international tourism data (combined outbound and inbound trips)

The choice of the estimation method and necessary adjustments relies heavily upon the availability of the external data sources and depending upon the model (cross-sectional versus time-series) and on the length of the time series data. In Section 3, under each quality aspect suggestions are made, how to assess the bias and/or how to adjust estimates. Expert opinion should be used as the last resort when no other information is available to avoid subjectivity and lack of comparability from one period to another. Note that even when no additional data source is used at the estimation stage, then at the quality assessment stage additional data sources are still needed.

An example of the estimation of the inbound data is presented in Figure 16, where the raw uncorrected monthly number of trips to Estonia by tourists of one country based upon the data of one MNO is compared to accommodation statistics. The figure illustrates the fact that the adjustment of the levels is quite large, but the direction of the month-to-month changes is pretty much the same in both the corrected and uncorrected series. Therefore, in this particular case, the trends and ratio estimates that are based upon uncorrected data are likely to be not much different from trends and ratios that are computed against corrected data.
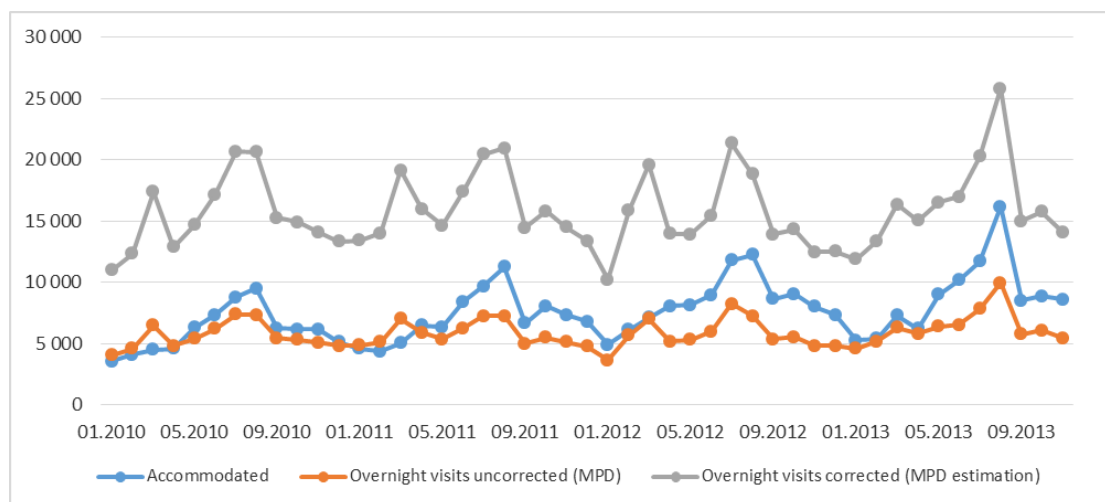
Figure 16. An example of an estimation based upon the data from one MNO concerning tourists from one country.

2.7. Combining Data from Different Operators

When using data from several MNOs, there are a few options in which data can be combined. Depending upon agreements with MNOs and legislation, there are several possible ways of combining data from different MNOs. A description of different options is presented in Report 2 Section 4.3. The current document will describe Option A from which the resulting estimates and aggregated data are combined (Table 9). Alternative option B (Table 10) does not require individual aggregation and estimates on MNOs side – the data is combined in the initial phase and aggregated and corrected as already combined dataset.

Table 9. Option A for data handling. This option presents the maximum privacy protection option with most implementation and maintenance cost.

Internal MNO					Outside
MNO No 1	Data extraction and internal preparations	Coding/anonymisation	Frame formation	Data compilation	Estimation and aggregation	Combining aggregates from different MNOs
MNO No 2	Data extraction and internal preparations	Coding/anonymisation	Frame formation	Data compilation	Estimation and aggregation	
MNO No 3	Data extraction and internal preparations	Coding/anonymisation	Frame formation	Data compilation	Estimation and aggregation	

The advantages of Option A are as follows:

- Operators' business-sensitive data is kept during the processing of that data (no raw data is shared);

- Ideal in terms of subscriber privacy since no raw data at the subscriber level is delivered outside the MNO's infrastructure.

The disadvantages of Option A are as follows:

- It is not possible to combine the same subscribers using the roaming services of various MNOs (inbound cross-roaming);

- Estimations for population of interest (coefficients) are carried out based upon each MNO's results. Results from MNOs have to be combined based upon a mathematical logic (a weighted average based upon the visitor's home country);

- Each MNO has to implement a sophisticated system for processing the data from raw data extraction up to the estimation and results. Resources have to be available to carry out quality assurance (QA) for all of the steps.

Table 10. Option B for data handling. This option presents the minimum privacy protection option with least implementation and maintenance cost.

Internal MNO		Outside				
MNO No 1	Data extraction and internal preparations	Coding/anonymisation	Frame formation	Data compilation	Estimation and aggregation	Combining aggregates from different MNOs
MNO No 2	Data extraction and internal preparations					
MNO No 3	Data extraction and internal preparations					

The advantages of Option B are as follows:

- It is possible to combine the same subscribers using the roaming services of various MNOs (eliminating inbound cross-roaming).

- Estimations for population of interest (coefficients) are carried out based upon all available data combined as one dataset and not based in individual MNO's;

- MNOs do not have to implement sophisticated system for processing, instead they just need to set up the extraction and delivery system of the data.

The disadvantages of Option B are as follows:

- Operators' business-sensitive data is exposed to the processor during the processing of that data;

- Indirectly identifiable data is exposed outside MNOs and therefore it is more sensitive from the point of view of the privacy protection.

The results of data compilation from each MNO should provide estimates for the population of interest. In an ideal situation, all estimates from MNOs should be similar and therefore the results should be the same. However, each MNO usually has various over-coverage and under-coverage in some proportions of the population that is of interest. And there is rarely a suitable for all estimation method, so the results differ. Combining aggregated results from various MNOs should result in more realistic estimates for the population of interest. One way to combine the results is to calculate the weighted average for each statistical indicator where weights take into account the MNO specific characteristics. For example, if one MNO has a very high penetration rate among one part of the population (e.g. foreigners from one country), the estimates based upon the data from that MNO concerning that population group should have greater weight when compared to other MNOs in the combined estimate. If the population group is equally represented in each MNO, the weights should be the same. Combining aggregates from two MNOs into one estimate is illustrated in Figure 17 (equal weights) and Figure 18 (specific country, unequal weights). The examples provided are based upon real data from Estonian MNOs.
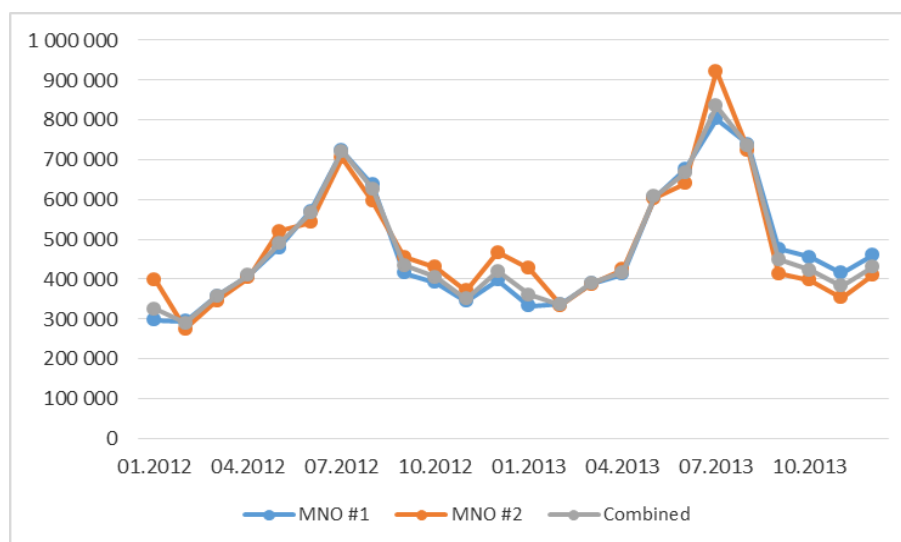
Figure 17. An example of monthly estimates of the total number of inbound trips for two MNOs separately and combined.
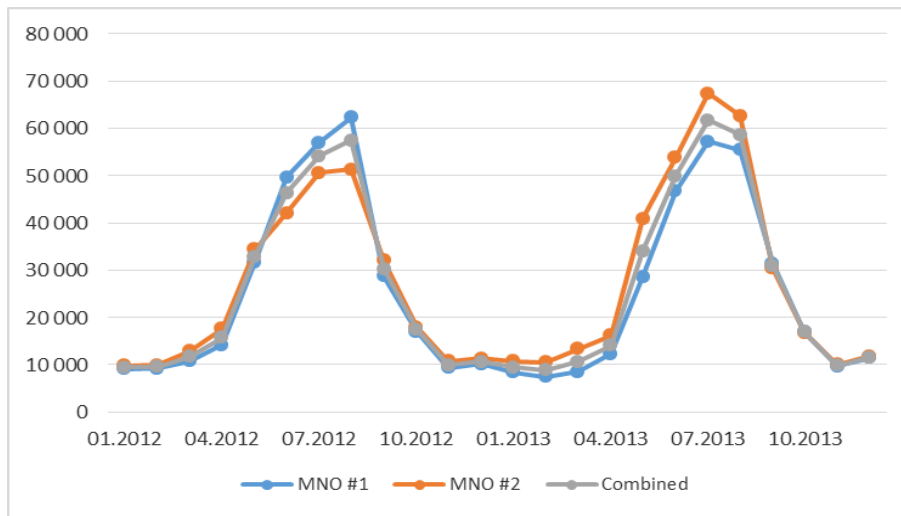
Figure 18. An example of monthly estimates of the total number of inbound trips from one country for two MNOs shown separately and combined - the weight of data from both MNOs is of equal standing.
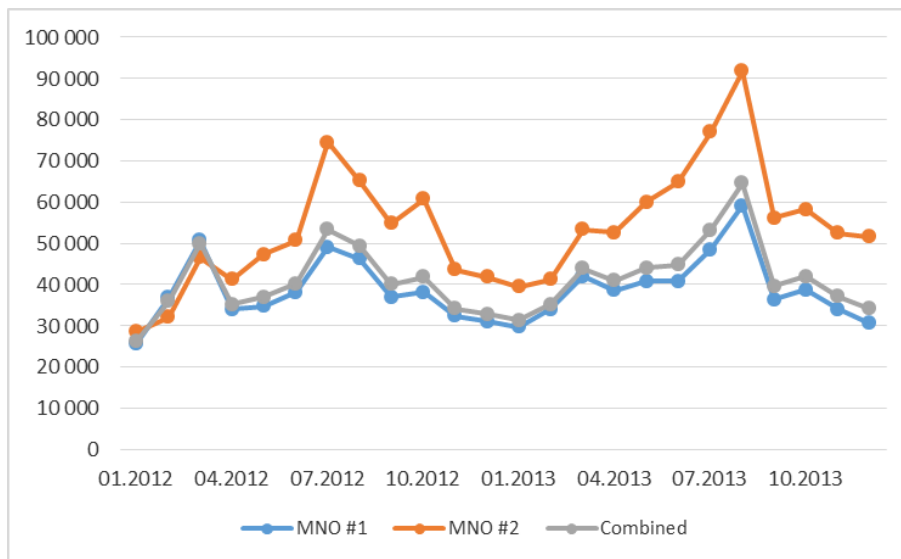
Figure 19. An example of monthly estimates of the total number of inbound trips from one country for two MNOs shown separately and combined - the weight of the data from MNOs is different, as the estimates for those trips which were based upon the data taken from MNO No 2 are less reliable.

2.8. Other Issues

2.8.1. Revisions, Preserving Historical Data

Mobile data can be updated periodically. Data that concerns the specific limited period of time describe tourism activity incompletely as the actual trips are ongoing and do not start or end accordingly with the limits of the data. Longer statistical periods describe any

phenomena better; therefore, it is useful to recalculate some particular proportion of recent historical data when data updates are available. A recalculation of the data (both frame formation and data compilation) may result in a change of previous periods' results. This should be taken into consideration and a revision policy should be adopted for the rules that are set for recalculation.

Potentially unfinished trips and usual environment/residence should be analysed during recalculation. Due to the recalculation, the following changes may occur in previous periods when compared to the outcomes of recalculated results:

- Change in place of residence and usual environment for subscribers results in a different number of tourism trips (especially in domestic case);

- Change in the duration of trips results in a different number of same-day and overnight trips;

Figure 20 presents an example that shows recalculated and initial values for a simulated monthly updates of domestic tourism trips. The sensitivity of the model depends highly on the criteria used for defining the usual environment (see also 2.3.1.1).
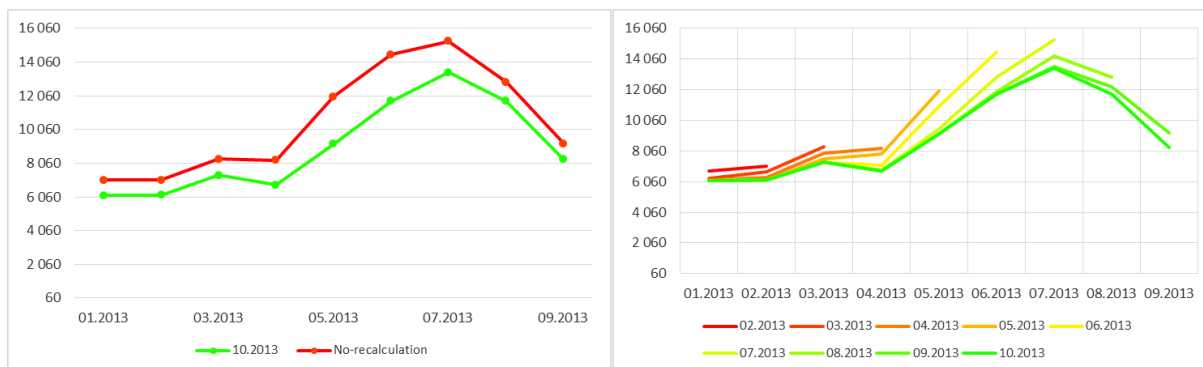
Figure 20. An example of the difference between recalculated and initial monthly domestic tourism trips for a random sample of 50,000 domestic subscribers. The left-hand chart represents a comparison between the results that have been combined from non-recalculated monthly updates and the recalculated results for the whole period. The right-hand chart represents the situation for each set of monthly updates with recalculations added. The period for defining the usual environment is ninety days, which is rather sensitive towards identifying new usual environments.

The typical case for domestic tourism can be students' arrival at the university. After the summer, new students arrive at the university town. If the algorithm to be used for identifying the usual environment is set to three months (a person must regularly be present or visit the administrative unit on a weekly basis for a period of three months), then for a period of three months those students will be identified as tourists. If the recalculation is not used, then those students will appear as residents on the third month but will remain tourists for the

previous two months. If the data is recalculated, then once the update has been carried out on the third month, the number of tourists decreases and the number of residents increases in the first month.

Due to the requirement of recalculating the data, original event data from MNOs should be preserved at least for a certain period of time in order to be able to carry out any recalculations. This is also valuable in case any changes in methodology should occur. However, a recalculation of the historical data due to methodological changes should seriously be considered when such a necessity occurs.

A change in methodology might emerge if the characteristics of the data change drastically and the current methodology may result in different outcomes. Such a change should be thoroughly analysed and changes to the methodology should not affect the ability to compare historical data. Most of the potential changes in the characteristics of the data might reside in an increase in the number of events and/or frame/sample and compensation can most probably be provided in the estimation process (by applying fewer and smaller correction coefficients for spatial under-coverage or over-coverage), thereby preserving comparability over time.

Preserving at least some part of the original event data is a requirement in order to be able to reprocess the data for revisions and in case changes in methodology occur. However, this is strongly connected to the legal obligation for MNOs to erase the historical data of subscribers.

2.8.2. Statistical Confidentiality

Original data from mobile positioning is highly sensitive and has to be processed so that the identification of specific subscribers is prevented. The results (aggregated results) should be checked and processed before publishing so that disclosure of an individual person either directly or indirectly is not possible. Therefore, the estimates that rely on a few subscribers or few trips should be left unpublished or changed (e.g. rounded).

3. Quality

The aim of this section is to evaluate the methodology described in the previous section by highlighting the differences from the standard methodology and by addressing issues that may influence the output quality.

The methodology is evaluated in this report with regard to validity, accuracy, and comparability. Other quality aspects are covered in different reports for this study, such as accessibility (Report 2), coherence (Report 3b), timeliness (Report 4) and costs (Report 4).

The standard quality aspects are viewed 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. For example, in household surveys it is quite usual to have response rates between 60 and 70%.

However, non-response errors do not appear in the context of mobile positioning data because here all units are observed. Similarly to when administrative data is used, in the case of mobile positioning data the biggest concerns from a quality point of view are the differences in definitions (i.e. what we want to measure versus what is available in the dataset), plus coverage and processing errors, as well as comparability over time. The aforementioned aspects will be described and discussed in the following part of the chapter.

3.1. Validity

Tourism statistical concepts are described in the *Methodological Manual for Tourism Statistics* (Eurostat 2013a), which is in line with *International Recommendations for Tourism Statistics* (UN 2008). The most important concepts together with their definitions were given in Section 1.3. Regardless the data source to be used for the production of official tourism statistics, these concepts need to be followed in order to guarantee comparability over regions and time and the coherence with other statistics. The aim of this subsection is to compare the concepts in the mobile positioning field and if possible, give an assessment of the size of the discrepancy.

Following the standard concepts is quite straightforward when surveys are used for data collection, although even then it is not without its own challenges.

Validity is not often assessed in official statistics, it is more common in the clinical studies and studies in the field of social sciences. In these fields, the validity refers to the degree of consistency between the desired construct and the instrument measuring that construct. In official statistics such abstract constructs that need evaluation are rarely measured. However, in surveys where questionnaires are used, the process when questionnaires are tested and developed and different versions of the questions are tested is

actually evaluation of the validity. Different statistical methods exist that allow to assess such validity (also called construct validity).

Now, when other sources than survey data are used, the concepts there are usually not in line with the standards and careful evaluation of the definitions needs to be carried out, and ways to reduce discrepancies between existing definitions and standard definitions should be sought. 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). This evaluation is shown in Table 11 for the main concepts.

The only quantitative assessment that can be carried out relies on the correlations with other measures of the same concept that are measured at the same time (this is also called concurrent validity). Such other measures here can include supply-side statistics, demand-side statistics, and transport statistics. Comparison between mobile positioning data and the aforementioned statistics is carried out in Report 3b, which also contains correlations.

Table 11. Descriptions of tourism-related definitions that are applicable to mobile positioning data, stressing the differences from official definitions.

	Inbound	Domestic	Outbound
Estimation units			
Visitor	A traveller taking a trip to a main destination outside their usual environment, for less than a year. Mobile positioning data allows the difference between visitor, trip and visit to be distinguished during the period of observation. However, mobile positioning data does not provide information on the purpose of the visit neither does it allow work-related visits to be identified in which the employer resides in the country visited.	A traveller taking a trip to a main destination outside their usual environment, for less than a year. Mobile positioning data allows the difference between visitor, trip and visit to be distinguished during the period of observation. However, mobile positioning data does not provide information on the purpose of the visit neither does it allow work-related visits to be identified in which the employer resides in the place visited.	Same as visitor in inbound.

	Inbound	Domestic	Outbound
Tourism trip	Tourism trip made by a foreign resident that begins with entry (border-crossing) to the country of reference and ends with departure from the country of reference.	Tourism trip made by a resident of the country of reference within the country of reference. A domestic trip begins when the person leaves their usual environment and ends when the person returns to their usual environment (excluding travelling within the person's usual environment).	Tourism trip made by a resident of the country of reference outside the country of reference (to one or many foreign countries). An outbound trip begins with departure (border-crossing) from the country of reference and ends with entry back to the country of reference.

	Inbound	Domestic	Outbound
Visitor characteristics			
Country of residence, usual environment	The place of residence is the same as the usual environment. Country of residence is approximated by the foreign country in which the home MNO of the subscriber is located unless the visitor's duration of visits in the country of reference exceeds a threshold (depending upon the method) of time spent and they are considered to be a resident and no longer an inbound tourism visitor.	The place of residence is located within the country within the usual environment. A usual environment is considered to be a geographical territory (administrative unit or geometrical figure - depending upon the concept of the usual environment of a specific country of reference) in which a person conducts their regular daily routines.	The place of residence can be the same as in domestic tourism, but is considered to be the country, not as the smaller scale place. A usual environment is considered to be the whole territory of the country of reference unless the visitor's duration of visits in the foreign countries exceeds a threshold (depending upon the method) of time spent and is therefore considered to be a resident of some other foreign country and is no longer an outbound tourism visitor.
Gender and age group	Not identifiable	If subscriber's data is available then it is gender and age group of the subscriber and not the visitor.	If subscriber's data is available then it is gender and age group of the subscriber and not the visitor.

Trip characteristics	
Start of the trip	Start of the trip is approximated by the first event recorded by the MNO that is classified as outside usual environment (domestic) or country of residence (outbound, inbound)
Duration of the trip	Mobile positioning provides a means to measure the duration of the trip in total hours, days present, nights spent. Total hours and nights spent per trip can be summarised on all aggregation levels; however, days present cannot be summarised.
Same-day trip	Trip that does not include overnight stay.
Overnight trip	Trip with a stay during which a change of dates occurs, a place at which the night is spent (regardless of the actual rest/resting place).
Visit	Limited to the actions of visitors, therefore representing tourism. The concept of a visit depends upon the level of the geography where it is used. It can mean either the whole tourism trip or only a part of it, depending upon the perspective (origin-based or destination-based).
Main destination	A distinction between the main destination of the trip and a main destination of the day can be made. The main destination is approximated by one of the countries/places where event was recorded by the MNO using the official criteria (longest stay, farthest place); however, during overnight trips, each day might have a different main destination (usually the one at which the night is spent). On a same-day trip, the main destination of a trip is the place at which most of the time was spent. A visitor can visit one main destination and several secondary destinations during one day.

Regarding the concept of visitor and tourism trips, the main difference from the official definition concerns visits that are work-related visits in which the employer resides in the country or place that has been visited. If these work-related visits are done regularly then it will be identified as usual environment and excluded from the frame (see Table 12 No 21). In case it is irregular activity, it is likely included in the frame. Depending upon the country and the work commuting patterns this may or may not have influential impact on the estimates. According to the Estonian Population and Housing Census 2011 about 25,000 Estonians commute to work outside Estonia and over a third of employed persons commute to work to other local government units. Population and Housing Census data does not give information about how regular this commuting is.

Regarding the concept of usual environment and country of residence then the discussion regarding the differences from the official definition can be found in Section

2.3.1.1 and the topic is tightly related with several coverage issues listed in Table 12 and processing issues discussed in Section 3.2.4.

Regarding the listed trip characteristics, several differences from official definitions are present due to how data is generated. Same-day trips and overnight trips are not easily distinguishable because the start and end of the trip are not directly available but are determined by a trip identification algorithm. See also Table 12 No 6. The same is true of the concept main destination.

When looking at the definitions, one can say that concepts that may have validity problems include the place of residence (inbound, outbound), the usual environment (domestic), same-day trips, and main destination (outbound). The impact is difficult to assess because it is closely related with the coverage issues (e.g. penetration of the mobile phones, mobile phone usage patterns in different countries and by different nationalities) and processing issues (e.g. availability of the longitudinal data and the effect to the algorithms).

Indirectly one can assess the size and the impact of the validity problem by looking at the coherence between different data sources. However, when comparing results from different sources it is not possible to distinguish the source causing the difference, it is often the co-effect of many issues e.g. coverage, processing, validity. Coherence is covered in Report 3b of this project.

3.2. Accuracy

This section provides an overview of the coverage, sampling, measurement and processing issues when using the mobile positioning data. Recommendations on how to deal with the problems will be listed and the effects on the final results will be analysed.

3.2.1. Coverage Issues

The target population was defined in Section 2 and the compilation of the sampling frame was described in three separate sections for outbound (2.4.1), inbound (2.2.1) and domestic (2.3.1) tourism. Perfect coverage means that there is one-to-one mapping between the target population units and frame units. Imperfect coverage means that there are some units missing from the frame (under-coverage), there are units that should not be in the frame (over-coverage) or there are some units more than once (duplicates). All these three occasions are called coverage errors and they lead to coverage bias.

In traditional surveys, the under-coverage is the most serious problem among the three error sources, as over-coverage and duplicates can be identified and their effect removed during data processing. In the case of mobile positioning data, both under- and over-coverage are problematic, while duplicates can be identified and removed.

In case of mobile positioning data, there is one discrepancy that is immediately clear from comparing the definition of the target populations (on participation in tourism and on the characteristics of tourism trips) and the population frame. In the frame are all of those subscribers who used their mobile phone for calling or texting, whilst the target population includes all of those individuals who reside in the country. This leads to a large number of various coverage problems, the complete list of which is given in Table 12.

Some coverage issues that are listed in the table can be avoided during frame formation or data compilation (before estimation is carried out) by identifying those observations that do not form part of the population and by excluding them from the frame.

Table 12. List of possible coverage issues with data from MNOs.

Issue	Under- or over-coverage	Possible solution
1. Inbound, domestic, outbound: visitors who do not use mobile phones (children, phone-free holidays, expensive roaming services, do not own a phone, etc.).	Under-coverage	At estimation stage use information from additional sources (statistics relating to mobile phone not users and their travelling behaviour) to make and check assumptions about the travel behaviour of the under-covered group and then apply appropriate models to adjust estimates. Both model based and model-assisted estimators can be used. In Estonian case model based approach was used and assumption that people without mobile phones behave similarly to those who have mobile phones was done.
2. Inbound, domestic, outbound: some visitors use more than one mobile device on their trips: • all devices use the same network during the trip causing a duplication of the data for a single person in one of the MNO's datasets; • devices use different networks during the trip which cause the duplication of the data for a single person from the datasets supplied by different MNOs.	Over-coverage	At estimation stage use information from additional sources if available to make and check assumptions about the number of people having several mobile phones and then apply appropriate models to adjust estimates. If the phenomenon does not occur often then no adjustment need to be done. In Estonian case this phenomenon has not been assessed.
3. Inbound: national roaming subscribers.	Over-coverage	Can be fully excluded as the country of roaming partner is the same as the country

Issue	Under- or over-coverage	Possible solution
		of reference.
4. Inbound, outbound: visitors not actually entering/exiting the country of reference (not crossing the border) but who are using the MNO roaming service of the country of reference or foreign country.	Over-coverage	Can be excluded partially based upon border bias recognition algorithms. Based upon the Estonian data the percentage of excluded trips among inbound roaming has been 10.4% of all trips. A total of 58% of such trips are classed as trips with only one event. Among outbound roaming the exclusion embraces roughly 6% of the initially-calculated outbound trips (14.4% of same-day visits).
5. Inbound, domestic, outbound: technological devices that were not removed by MNOs during the preparation process.	Over-coverage	Possible removal based upon event patterns. Based upon the Estonian inbound roaming data, the percentage has been from 0.04% to 2% depending upon the MNO and the algorithm used. Based upon Estonian outbound roaming data the number of such devices is very small, less than 0.01%.
6. Inbound, domestic, outbound: same-day visits are over-represented when compared to overnight visits. This is caused because it is likelier that visitors do not use their mobile phones on every single day of the trip.	Over-coverage of shorter visits on the account of longer visits. Under-coverage of the duration of the trips.	Model the likelihood of single trip being a same-day trip and use modelled values in the estimation. In case the reliable reference data concerning the difference between the same-day and overnight visitors exists, such data should be used to correct the mobile data.
7. Inbound, domestic, outbound: Data from selected MNOs only i.e. not all data from MNOs is available.	Under-coverage	Apply external information regarding the penetration rates and customers' profile. See also Section 2.7.
8. Inbound, outbound: visitors who use local SIMs for some reason and are not represented in the inbound/outbound roaming registry (see also Point 18).	Under-coverage	Calibration or similar approach where known totals from other sources are used to correct for the under-covered part of the population. In Estonian case, accommodation statistics and sea transport statistics of passengers was used to estimate the number of such visitors.
9. Inbound, outbound: technological limitations on the use of mobile phones from/in specific countries due to technological barriers, limitations on roaming service (no roaming agreement between MNOs), high roaming costs, or network coverage is bad.	Under-coverage	Has to be a part of different MNO-based country-specific estimations compensating the technological limitations. In Estonia, subscribers from US and Japan are in such a situation and therefore require higher correction coefficients.
10. Inbound, outbound: international roaming (travel) SIMs that are sold globally and where the country of the MNO does not	Over-coverage of the country of the MNO, under-	Exclusion of such subscribers should be possible for outbound roaming as MNOs usually know who their subscribers using

Issue	Under- or over-coverage	Possible solution
match the actual country of origin of the subscriber.	coverage of the subscribers from the countries that are using this service.	their international travel roaming service are. Possible compensation on the account of the actual residents of the country of reference might be required. The number of such roaming cards is very different in countries and it is very difficult to measure. Based upon Estonian data no such subscribers have been possible to identify from MNO prepared data.
11. Inbound, domestic: differences in phone usage patterns depending upon the location peculiarities are very difficult to understand; however, it is logical to assume that in different environments, people use their mobile phones differently; therefore generating different volumes of events. The best example is the one in which phones are more often used in urban areas when compared to rural areas.	Under-coverage in rural areas	At estimation stage use information from additional sources to make and check assumptions about the travel behaviour of the under-covered group and then apply appropriate models to adjust estimates.
12. Inbound, domestic: limitations on the use of mobile phones in specific areas within a single country in which network coverage is poor.	Under-coverage	At estimation stage use This problem is difficult to overcome as no relevant additional information is likely available. If the area is large and is relevant from additional sources to make and check assumptions about the travel behaviour of the under-covered group and then apply appropriate models to adjust estimates tourism perspective small area estimation methods e.g. synthetic estimator could be applied.
13. Inbound, domestic: how representative the set of visitors or trips is at the geographical level. Because visitors do not use a mobile phone at every place they visit during the trip, it is logical to assume that the smaller the geographical level, the less representative the mobile data will be when it concerns tourism visits to those places.	Under-coverage on lower administrative levels	Higher correction coefficients for lower administrative levels when compared to higher administrative levels.
14. Inbound: cross-roaming, i.e. a single device using several roaming services during the trip causing duplication of the data for a single subscriber in two or more MNOs' datasets.	Over-coverage	Unless data for this device (subscriber) can be identified as the same subscriber and combined, this has to be dealt with during the estimation process.
15. Inbound: a different penetration rate for MNOs among foreign subscribers of specific countries.	Under-coverage or over-coverage	Assuming that information is available that allows to create separate MNO-based models that have country-specific

Issue	Under- or over-coverage	Possible solution
		parameters then weight adjusted estimates can be computed. In Estonian case very different penetration of subscribers is used (lower to neighbouring nations FI, LV; higher for US and Japan because of technological limitations).
16. Inbound: residents of the country of reference who are using foreign phones.	Over-coverage	Can be excluded partially based upon the duration of the presence within the country of reference. Based upon the Estonian data approx. 0.2-0.4% of the total number of trips can be identified as being trips made by residents and not tourism.
17. Inbound, outbound: visitors passing through the country (transit visits).	Over-coverage	Can be excluded partially based upon identifying short trips within transit corridors. Depending upon the month, the number of transit trips in the Estonian data vary from 2% to 10% of total inbound trips and 9.2% of the stays in foreign countries can be considered as transit pass-through.
18. Domestic, outbound: non-resident subscribers whose usual environment and residence are outside the country of reference but who are using local pre-paid SIMs for various reasons (see also Points 8 and 19).	Over-coverage	Exclude subscribers with a short lifetime or if ID linkage to outbound data is possible, compare the duration of stays within the country of reference and abroad for exclusion from the domestic and outbound data and inclusion to inbound data. Based upon Estonian data 2.9% of all domestic subscribers spend more time abroad, representing 24% of all outbound trips, and therefore should be considered as foreign residents and excluded from domestic tourism dataset.
19. Domestic: resident subscribers who change their pre-paid cards very often (short longevity of the *subscriber_id*) and whose residence and/or usual environment cannot be identified (see also Point 18).	Under- or over-coverage	Exclude subscribers with a short lifetime.
20. Domestic: subscribers whose place of residence and/or usual environment is calculated incorrectly within the country of reference.	Under- or over-coverage	The incorrect calculation has a random nature and the usual residency and usual environment are 'compensated' by the similar incorrect calculations for other subscribers. The quantity of incorrect calculations depends on the accuracy (5% for second level administrative unit, 13% for third level administrative units).
21. Inbound, domestic, outbound: trips within a subscriber's usual environment not considered to be tourism trips.	Over-coverage	Identification of a subscriber's usual environment within the country of reference. Excluding trips taken within the

Issue	Under- or over-coverage	Possible solution
		usual environment which are classed as non-tourism trips.
22. Domestic: the different penetration rates for MNOs.	Under- or over-coverage	Surveys covering the penetration rates of MNOs can be used in the estimations if such information is available.
23. Domestic: the different regional and socio-demographic penetration rates for MNOs.	Under or over-coverage	Surveys covering regional and socio-demographic penetration rates of MNOs can be used in the estimations if such information is available.
24. Outbound: residents of those foreign countries who are using the SIM cards for that particular MNO in the country of reference for various reasons and should be excluded from outbound tourism data;	Over-coverage	If ID linkage to outbound data is possible, compare the duration of stays within the country of reference and abroad for exclusion from the domestic and outbound data and inclusion to inbound data. Also estimation. Based upon Estonian data roughly 3% of combined domestic-outbound subscribers spend more time abroad than in the country of reference.

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 the 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. One should note that, for example, a fairly large percentage of people in the reference country who do not possess a phone does not automatically lead to large coverage bias. A bias occurs when those people who do not have a phone travel a lot (relatively more than those who do have and use phones). Unfortunately, we do not have statistics that show the relationship between owning and using a mobile phone and a person's travelling habits. If we can assume that people who do not possess a phone do not travel, then this component does not contribute to the coverage bias of the variable the number of trips.

Some information about mobile phone usage while travelling is available for Europe. The Special Eurobarometer 414 (2014) shows that 28% of the travellers in the EU switch off their mobile phones when visiting another EU country (the corresponding specific figures are 33% for DE and 41% for FR but only 13% for EE and 15% for FI).
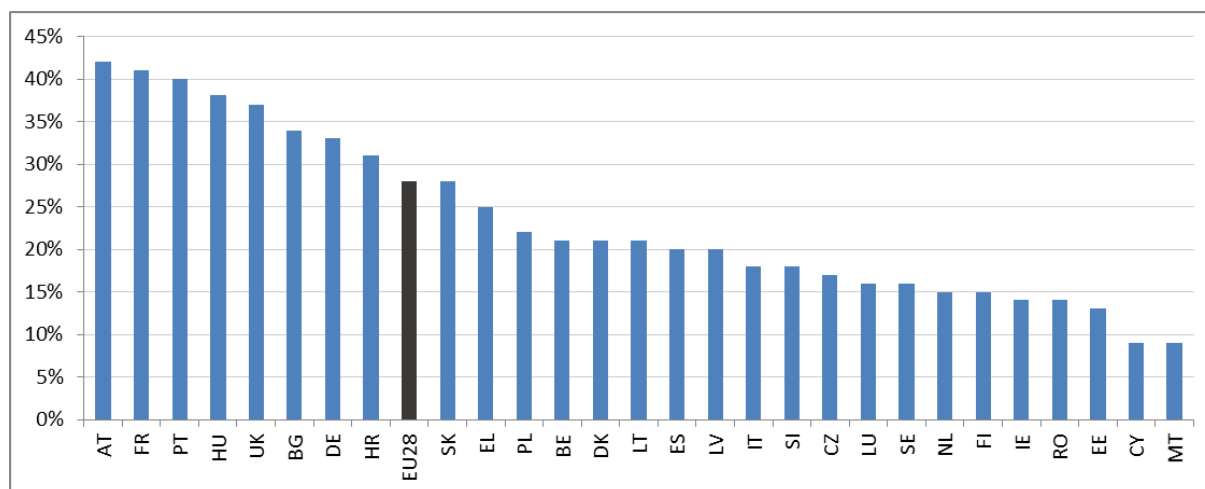
Figure 21. Percentage of the population (aged 15+, who have travelled and own a personal mobile phone) who generally switch off their mobile phone and never use it while visiting another EU country (Source: Special Eurobarometer 414).

The best but also the most costly way to evaluate the bias would be to carry out sample surveys for quality assessment purposes for each coverage issue, providing the basis for individual estimations. This survey would give the estimates for the proportion of people not having or using mobile phones, proportion of people using two or more devices etc. This information would allow to compute the total size of the bias under certain assumptions. Linking survey data with mobile positioning data at a micro level (which would reveal quality problems not only at an aggregate level but also at a record level) would allow more precise bias estimations to be carried out. However, this solution is ideal from a theoretical point of view, but it is not a very realistic solution in practice.

A less costly way to reduce the bias is to use the information from the other sources (e.g. results from the surveys covering the relation between the number of same-day trips and overnight trips to the country for inbound, the usage of mobile phone in different countries for outbound). Sample survey literature has shown that a coverage bias can be dealt with by using a method such as, for example, a calibration estimation (Särndal & Lundström 2005) if suitable auxiliary information exists. In addition, different model-based estimates using aggregate data from other sources can be constructed. Both model-assisted and model-based estimates are well covered in literature in which tools for carrying out model validation and methods of computing precision measures are also supplied.
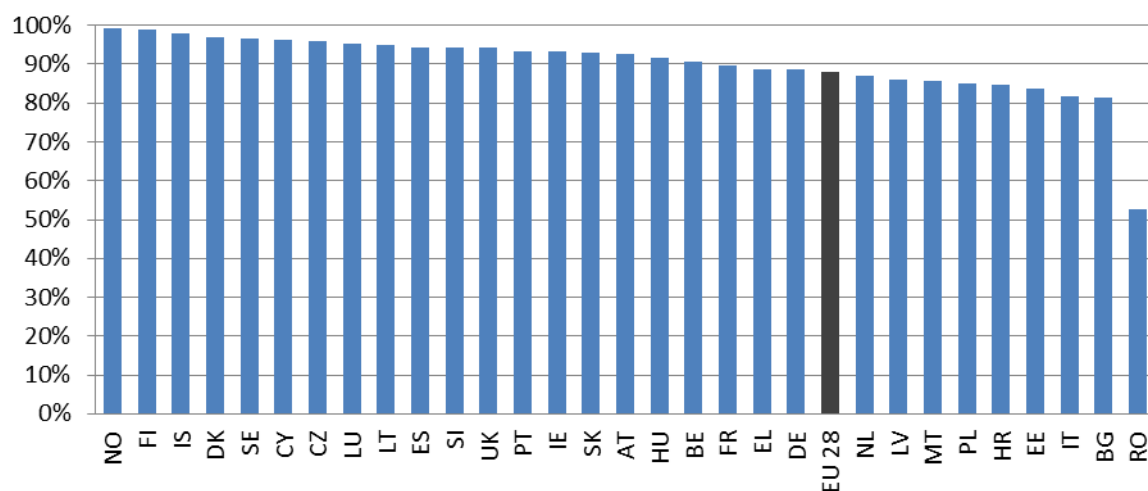
Figure 22. Percentage of the population (aged 16-74) that had used a mobile phone (first quarter of 2012) (Source: Eurostat).

Table 13. Users of mobile phones aged 16-74 in Estonia by different characteristics, first quarter 2012 (Source: Statistics Estonia)

	Percentage of mobile users, %
All individuals	83.9
Males	83.5
Females	84.2
Persons aged 16-24	89.7
Persons aged 25-34	89.5
Persons aged 35-44	85.9
Persons aged 45-54	82.6
Persons aged 55-64	79.6
Persons aged 65-74	71.6
Persons with below upper secondary education	77.1
Persons with upper secondary education	83.8
Persons with tertiary education	87.7
Employed persons	86.9
Employed persons also using their mobile phone for work-related tasks	44.9
Unemployed persons	83.6
Students (not in the labour force)	90.2
Retired (not in the labour force) and other inactive individuals	73.9

Figure 22 and Table 13 provide an idea of the size of some of the coverage problems listed in Table 12. The results are from a sample survey of individuals carried out in the second quarter of 2012. Owning a mobile phone is very common in Europe. In many EU countries the percentage of mobile phone users reaches 90% or more among 16-74 years old;

The Estonian results show, however, that in different population groups it is either more common or less common. Of course, this statistic does not say anything about how often and for what purpose the phone is being used on these occasions, therefore we do not know the impact of 16% of people not owning a mobile phone to the coverage bias of an estimate.

If no coverage problems existed and the data from MNOs represented the perfect frame for the population of interest, estimation would not be necessary. However, this is never the case and therefore using data from all MNOs from the country of reference decreases the overall coverage bias, but the estimation process is still required. However, if a single MNO represents the population of interest much better than other MNOs of the country of reference (e.g. the penetration rate for one MNO is higher than 75% among all subscribers), a comparison of the cost of implementation and maintenance with the quality of the resulting indicators from one or more MNOs should be conducted. If improvement of the quality of the results with all MNOs when compared to a single large MNO is minimal, the question of the necessity of using all MNOs should be raised.

The evaluation of the size of all the listed coverage problems needs to be carried out for each country separately, as many problems listed here depend on the environment (e.g. the price of the calls) and culture (e.g. children having mobile phones, having many mobile phones). For example, the use of foreign mobile phones by the commuters in Luxembourg (and therefore non-tourism) is probably different when compared to other countries.

3.2.2. Sampling Issues

Drawing a sample from the frame formed of mobile positioning data might be used because of privacy restrictions or because of limited technical resources for processing the large amount of data (see Section 2.1.1.1). If for some reason sampling is carried out, it is recommended to apply probability sampling techniques. In the case of a non-probability sample is drawn, the issue of sampling bias also arises. The choice of sampling designs is a wide one - simple random sampling or stratified simple random sampling are the likeliest choices as they are easy to implement. Stratified sampling design can be used only when information about subscribers or trips is available, such as demographic information about subscribers or trip characteristics.

When sampling is used, then sample among active subscribers during reference time period should be used. Subscriber as a sampling unit simplifies the estimation later on. Person rather than trip is sampling unit also in the surveys carried out for demand statistics and

border interview surveys. Sampling of events is not recommended as it complicates the identification of a place of residence in cases that involve domestic tourism and the computation of the duration of the trip in cases involving all kinds of tourism.

It has to be noted that because some of the exclusion procedures are made in the frame formation process, the sampling made by the MNO before frame formation is somewhat different from the sample produced after the frame formation. During the estimation process, sample weights have to be introduced that correspond to the sampling design. For obvious reasons, computing sampling weights and assessing sampling error is much simpler in the case of a random sample being used when compared to a non-probability sample being used.

In traditional surveys, the sampling errors vary a great deal, depending on the estimate, and can be quite large. In the Estonian Travellers survey, the coefficient for variation remained around the 4% and 5% mark for the total number of outbound and domestic trips. In the Finnish Travellers survey, the same estimates had a variation coefficient of around 16%. In the case of sampling from the frame of mobile positioning data, larger samples should be drawn, as there is no extra cost for the additional unit and it provides a smaller level of variability in the estimates.

3.2.3. Measurement Issues

The original data that has been extracted from MNO databases can involve various systematic and random measurement errors. The identification and possible correction of such faults should be carried out by MNOs before providing the data for further processing. However, the quality assurance process should be conducted every time a data package is received from MNOs. Measurement errors can be systematic or random; easily detectable or not (see Table 14).

Table 14. Possible measurement issues.

Issue	Frequency of occurrence	Opportunities for correction
Missing values for any attributes;	Common	Depending upon the nature of this issue, the records with missing attributes might just be deleted as they represent a very small fracture of the data.
Incorrect formats of the attributes (e.g. wrong time format of the *event_time*);	Rare	Difficult to correct and usually re-extraction of the data is required unless the number of such errors is very small and can be simply deleted with no significant effect to the results.
The illogical nature of the time attributes (e.g. data for January 2013 includes	Rare	Unless very specific records are identified (and deleted or requested again), the whole period of

Issue	Frequency of occurrence	Opportunities for correction
events for January 2012);		suspicious data has to be re-extracted. Usually not a serious impact, as they represent a fracture of the whole data unless this error is of systematic nature and widespread.
Incorrect country codes (inbound and outbound roaming);	Rare	If the specific country and wrong code is known, the records can be corrected. Most often occurs when a new country is established (e.g. South Sudan).
Incorrect location attributes (e.g. coordinates outside country borders for inbound roaming and domestic data);	Common	Usually a small fracture is incorrect. This should be fixed as soon as possible by the MNO.
Duplicated data;	Common	Can be corrected; however, may be costly in terms of resources (comparing millions of data points)
Missing data for a specific period of time or region (which could be missed if the time span or region is very small).	Rare	Requesting the data for the missing period. Sometimes this is technically not possible e.g. downtime of the MNO network and no data was generated nor stored during that time.

Such problems can be caused by systematic or random mechanisms. Systematic errors are in some terms easier to correct as there should be one problem causing them. Random errors are harder to correct and explain. Random errors usually represent a very small portion of the data and it is often easier to exclude them from the data rather than ask for a correction. Even when left in the dataset, they often cancel each other out and therefore have a negligible effect on the total outcome. Systematic errors are more serious and will cause a bias in the final results. Common systematic errors are caused by errors in the geographical representation (e.g. the coordinates of the antenna are wrong), but other attributes can also be erroneous because of a systematic error (e.g. the *event_time* attribute is presented in the wrong time zone). Those errors have to be communicated back to the MNO and resolved before further processing of the data.

Hidden errors are much more difficult to discover and are more damaging to the results. It is very probable that random hidden errors would not be discovered. Systematic errors, however, might be revealed only after the data has been processed and the estimation has been fully completed and the results do not seem to represent the population of interest. For example, systematic incorrect coordinates of the events result in a large number of visits to places to which there should not be so many visitors. The results might also reveal missing data (e.g. a large portion of the population of interest is missing). Such issues have to be communicated to the MNO, who should investigate the reason for the systematic error (or missing data) and the data should be resubmitted and processed again.

It is difficult to predict the impact of errors, but regular quality assessments of measurement errors help to detect the problems early on. If not regularly monitored, the effects can be very damaging up to the point at which the resulting data is unusable. There are some errors that cannot be resolved and have to be taken care of during estimation process, if possible. If not possible, the problem description together with the size of the impact should be given together with the estimates. Technical problems might cause gaps in the data that cannot be restored (downtime of the MNO). Therefore, using several MNOs for the data can also play a safety precaution role as problems usually occur only in individual MNOs at a time, and data from other MNOs can be used.

3.2.4. Processing Issues

There are two types of processing issues; technology-related and data-related. Technology-related issues are mostly linked with the ability to process a large amount of data and to handle the complexity of the algorithms. The hardware resources for this process should be able to process the data within a reasonable amount of time. However, the cost for the technological implementation for this process should also be reasonable. If the amount of time to process the data (including possible recalculation due to initial errors in the data) is nearing the data update interval, either more resources should be added or the use of a sample should be considered.

Because some of the algorithms used in the frame formation and data compilation processes are rather complicated by its nature, the issue of the ability to correctly process the data is important. The complexity of some such algorithms depends upon the national situation of the country of reference and can be simplified. However, the balance between the simplicity (time) of the process and the quality of the results should be considered. If possible, a control survey should be conducted in order to assess the accuracy of such calculations as they often depend upon national peculiarities. The survey should analyse how accurately the usual environment algorithm can define the actual usual environment for the subscriber based upon the data that has been identified for them when compared to the responses from the survey.

Another example of the data-related issue is the misclassification of inbound frequent travellers and long-term visitors when a trip identification algorithm is used, which is described in Sections 2.2.1.1, 2.3.1.1 and 2.4.1.1. As the data pattern for frequent visitors and long-term visitors (residents) is very similar and often not separable, such groups can be misclassified. Long-term visitors are also possibly residents that have to be eliminated from

tourism. A complex geographical pattern analysis can be applied to the dataset (long-term subscribers mostly stay in one place; frequent visitors mostly move between the entry and exit points and the destination), but this is a rather complicated algorithm that does not resolve a majority of these cases (e.g. when entry and departure are very close to the destination).

Figure 23 describes how a different definition of residency (with residency being defined by the number of days spent in the reference country during last 365 days) can influence the number of subscribers to be excluded from the frame. It also shows how important it is to have a longitudinal dataset for determining residency, as the impact on the number of subscribers to be excluded is quite large, especially within the frame of outbound tourism. The simulation results are based upon real data that has been supplied from one MNO in Estonia.
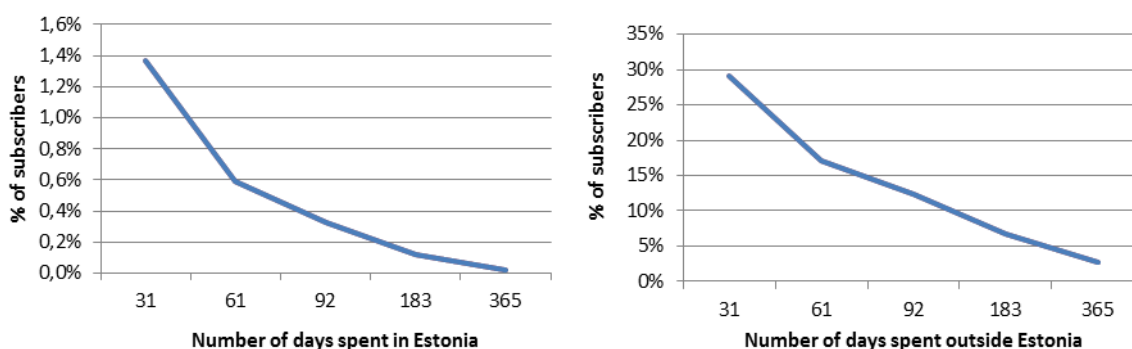
Figure 23. Simulation results for illustrating the classification into residents and non-residents based upon the number of days spent in or outside Estonia. Left-hand side figure describes inbound tourism and right-hand side outbound tourism.

Another example is identifying the exact duration of a whole trip or individual stops in places if the event data is very sparse and similar data can represent trips with different characteristics (e.g. a short visit for shopping purposes from a foreign country when compared to a whole-day visit represented by a single event). Because of a lack of better data, very often no distinction can be made.

The impact of choosing a different length between events to determine a trip is illustrated in Figure 24. Simulation results show that after a gap length of 52 hours or more the impact on the number of trips is very small. The simulation results are based upon real data supplied by one MNO in Estonia.

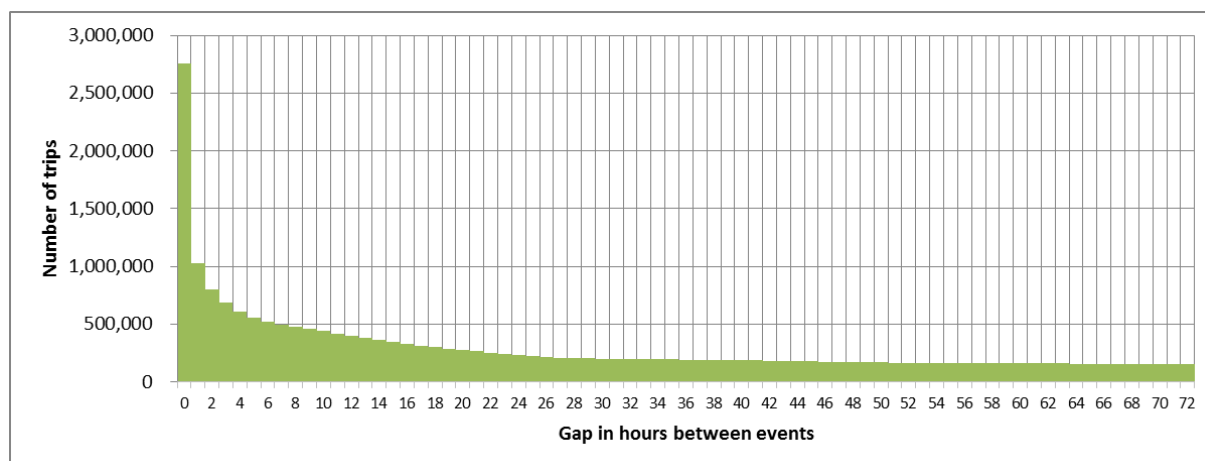Figure 24. Simulation results for illustrating how the number of trips is influenced by the determined length of the gap between two events.

As described in Section 2.3.1.1, there are several approaches when it comes to identifying the place of residence and usual environment of domestic subscribers based upon mobile data. One approach is to use the anchor point model (Figure 8). A study conducted jointly by the University of Tartu and Positium (Positium 2012) has shown that the anchor point model can identify the actual home administrative unit in 95% (median 100%) of cases for the second administrative unit level, in 87% (94% median) of the cases for the third level administrative unit level, and in 74% (87% median) of cases for the antenna coverage area level. This is for those subscribers whose anchor point is identifiable. The study was conducted in order to assess the accuracy of the anchor point methodology, involving 200 respondents and a total of 5,361 calendar months. The anchor points that have been miscalculated mainly represent coverage issue No 20 in Table 12 but also No-s 3, 19, 21. The size of the problem can be measured to a certain degree by conducting the survey or comparing the results with a nationwide registry such as the census or population register (see Figure 25 and Figure 26). The differences are based mainly upon local peculiarities (residents not registered in their real homes due to tax optimisation or other reasons). When it comes to comparing different datasets, the problem lies in the lack of accurate information about the number of residents. Census data is usually the most accurate information available. However, a comparison with census data can only be carried out when the census is up-to-date.
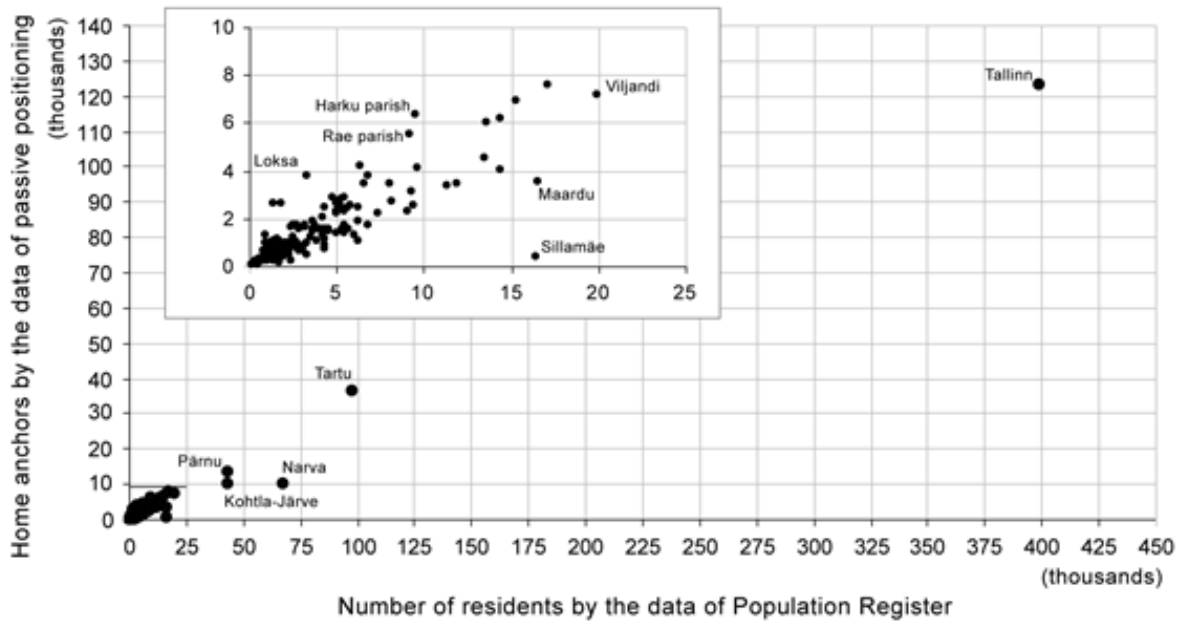
Figure 25. Correlation between modelled home anchor points and the number of registered persons in local municipalities in Estonia (Ahas et al. 2010).
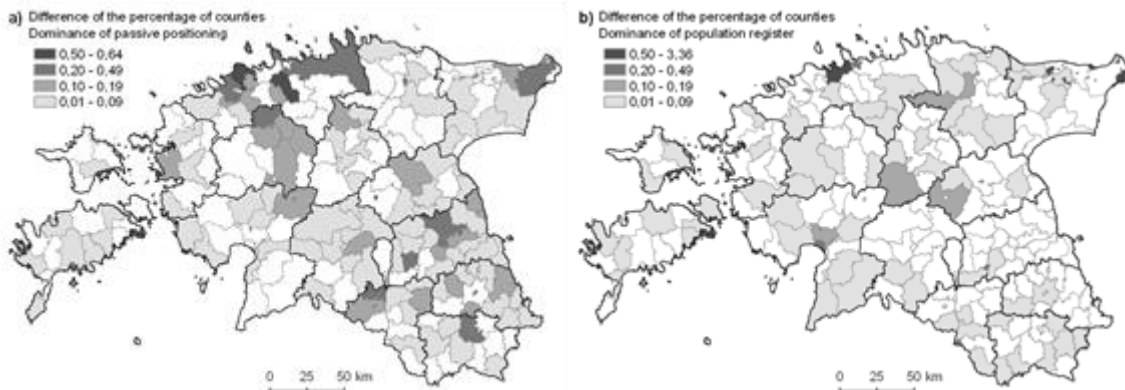
Figure 26. The difference between the number of home anchor point between the model output and data from the population register. Presented difference shown in percentage points. The total number of homes in all municipalities is 100%. a) Predominance of positioning b) Predominance of the Population Register.

Home anchor points for the average of 74% of domestic subscribers can be identified (Figure 27). This is due to the coverage issues mentioned in Table 12 (No-s 18, 19 and 20). For roughly a quarter of subscribers it is not possible to identify the place of residence, so they can be excluded from the frame as they represent only 8% of the total events produced. This is also the main reason for it not being possible to identify the home location - there are too few registered events available.
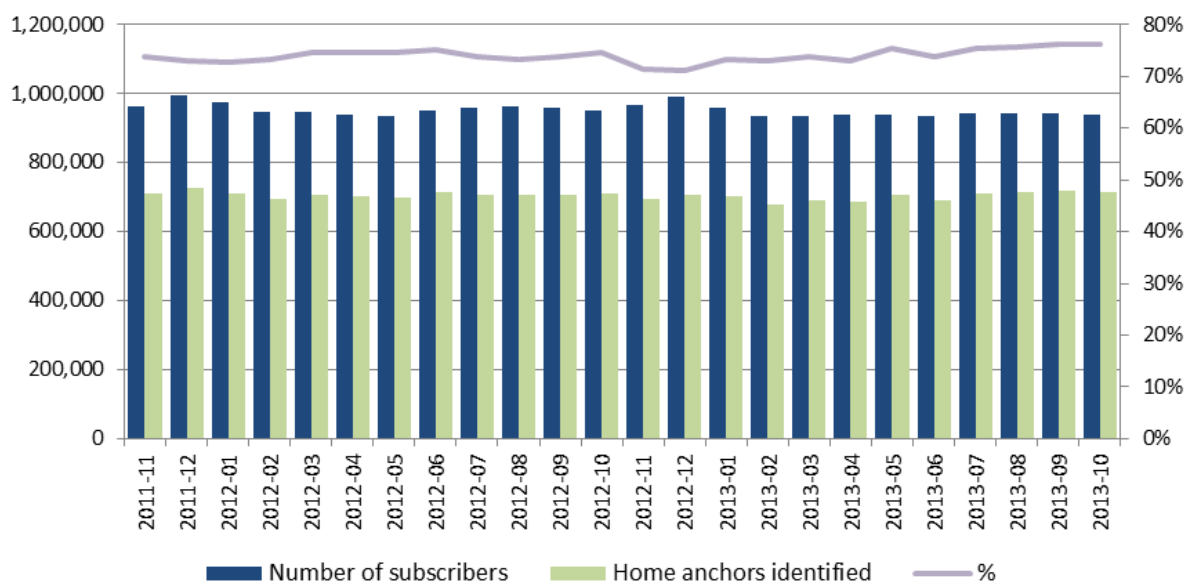
Figure 27. The proportion of subscribers per month with an identified home anchor point based upon the anchor point model used in Estonian data.

Similar problems also exist with other interpretations of data (border bias, transit trips, spatial segmentation of the visits, and interpreting the locations of nights spent). Most of them are non-standard and depend heavily upon national circumstances and the characteristics of the data supplied by specific MNOs.

3.3. Comparability

Comparability is an important quality dimension as many users of statistics are interested in changes over shorter or longer time period, as well as in comparing geographical areas, like countries, with each other. However, it is quite challenging to establish a high level of comparability because the environment, technologies and user needs change over time. In this chapter, several obstacles are discussed that may lead to breaks in the time series or cause incomparability between countries.

3.3.1. Comparability over Time

For a high level of comparability, changes in methodology, technology and legislation are not desirable (although also not very realistic), as any change may affect comparability of the time series. Guaranteeing comparability is also problematic in traditional surveys or surveys that rely upon administrative data. There, the comparability can be compromised when a new data collection method is used, or when the questionnaire is changed or different

imputation or estimation methods are applied. In addition, statistics that make use of administrative data are sensitive to changes in legislation.

Similarly to traditional surveys, comparability can suffer for statistics that are 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. Possible technological changes and their impacts are discussed in Report 2 Section 4.6. Methodological changes (changing the frame, processing algorithms, estimation methods), if necessary, need to be carried out by taking into account the effect on all quality dimensions and by avoiding changes that lower quality.

Due to the nature of passive mobile positioning data, the quality of those estimates that are based upon this data source depends upon changes in the telecommunication 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 is quite likely that the increased possibilities will change people's calling and texting habits, and as a result the content of the data also changes. Various experts in the telecommunications field foresee an increase in the traffic contained in mobile instant messaging, which is an alternative to text messaging and may lead to fewer text messages being sent in the future. Mobile instant messages cannot be observed by using passive mobile positioning data. This means that the quality of statistics that are based upon mobile positioning data are likely to change as well.

In Figure 28, the statistics that relate to the total number of text messages in European countries over time is shown. In 2003-2009, the number of text messages went up very rapidly, which from the statistician's point of view means that there is more relevant information to be used for compiling statistics based upon mobile positioning data. The raise in the number of text messages is likely related with the prices of the text messages going down. A GSMA (2011) report remarks that 'prices for a medium usage basket of mobile services declined by 11% per annum in the EU27 from 2006 to 2010; the high usage basket saw a similar price decline of 13%'. It also means that such a rapid change in the volume of the data and mobile phone usage patterns will very likely affect not only the quality but also the methodology and may possibly lead to more accurate but incomparable results.

In comparison, Estonian telecommunication statistics shows an upward trend in the time series for the total volume of text messages (Figure 29), but the volume of mobile phone calls has been fairly stable over time and may decline in the future (Figure 30).
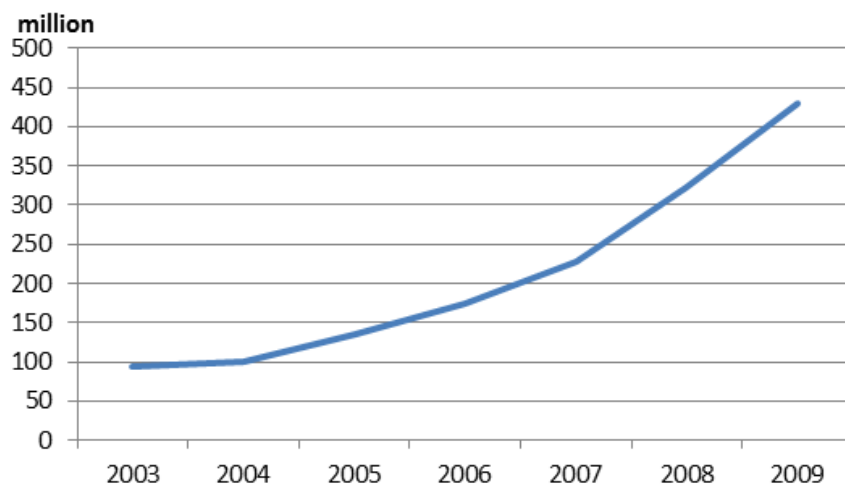

Figure 28. Volume of text messages sent on mobile phones in EU countries (excluding UK and the Netherlands), 2003-2009 (Source: Eurostat).
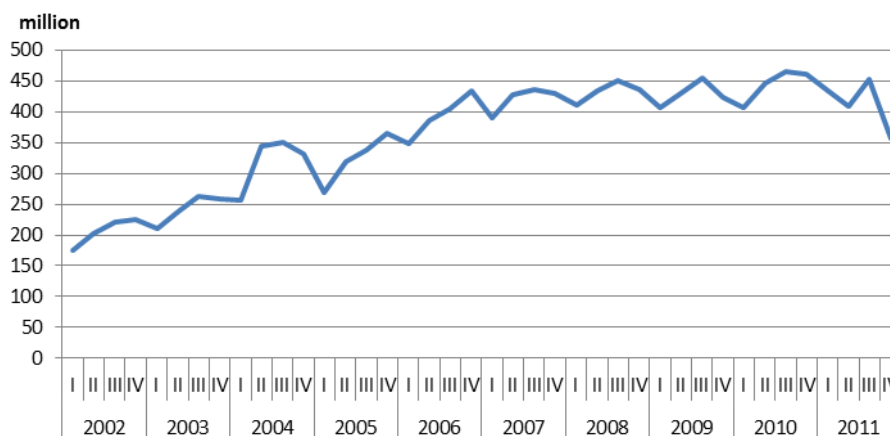

Figure 29. Total number of calls made from mobile phones in Estonia, 2002-2011 (Source: Statistics Estonia).
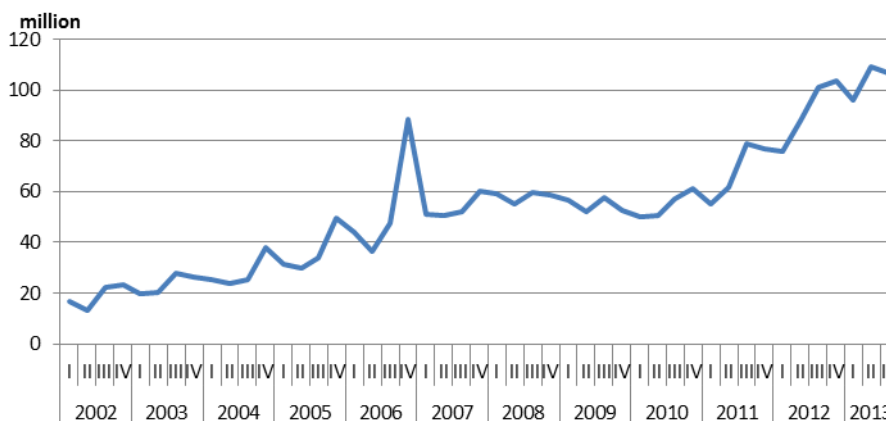
Figure 30. Volume of text messages sent on mobile phones in Estonia, 2002 - third quarter of 2013 (Source: Statistics Estonia).

The changes in the telecommunication market (e.g. emerging of new MNOs, merging of MNOs) can have an effect on comparability as well, but with careful analysis and appropriate estimation methods, the problem of decreasing the level of comparability could be avoided. The changes in the market and their impact is also discussed in Report 2 under risk assessment.

All in all, it is important to assess all the changes and evaluate their possible impact to the quality in general, including the effect to the comparability with earlier data. In the case of breaks in the time series, the explanation of the cause of the break needs to be given, and if revisions can be made, the whole series should be recalculated. Back-calculation of the series is common procedure when changes occur and back-calculations can be made.

3.3.2. Comparability between Countries

The use of the same concepts is one of the key elements for enabling comparison between different countries. For tourism statistics, the concepts are harmonised by Eurostat and countries providing statistics to Eurostat should comply with these harmonised concepts. However, the use of data sources other than data from the direct survey makes applying the same concepts difficult because the data user is not in control of the content. The same applies when it comes to mobile positioning data where some of the concepts (mentioned in Section 3.1) cannot be directly measured by using mobile positioning data. It is very difficult to quantify the difference between desired and applied concepts but, as mentioned in Section 3.1, for many concepts listed there the difference is expected to be small.

Of course, other sources for incomparability also exist, like the use of different frames or data collection modes or the application of different imputation and estimation methods, but these differences usually do not compromise comparability, or, if they do, then it is often

possible to estimate the level of the difference and correct the results. In the case of mobile positioning data, the coverage bias should be taken into account when comparing results between countries as coverage bias indicates one of the most important differences.

4. Relevance for Other Fields of Official Statistics

Tourism statistics are the main domain for which mobile positioning data could be used. Several other fields of official statistics may benefit from this source if they use the same or similar definitions to tourism statistics. The characteristics of mobile positioning data allow such data to be processed so that it meets the needs of different areas of statistics. Although there are significant conceptual differences between various fields of statistics, it is advantageous to seek synergies and common indicators that can be derived from the data and joint processes.

There are two options for using the methodology for different fields if the data processing is based upon the same source data and within one infrastructure (i.e. using the same hardware to process the data):

1. Common use of generated indicators;
2. Parallel domain-specific processes.

The option of using joint processes for calculating various indicators is a great opportunity when it comes to using mobile positioning data. Ideally, common processes should be used for all domains of interest with the separate process branches for specific outcomes where a different methodology is required (Figure 31). Common processes should involve the preparation of data and the calculation of indicators that are commonly used (although some might use different criteria, e.g. residence and usual environment).
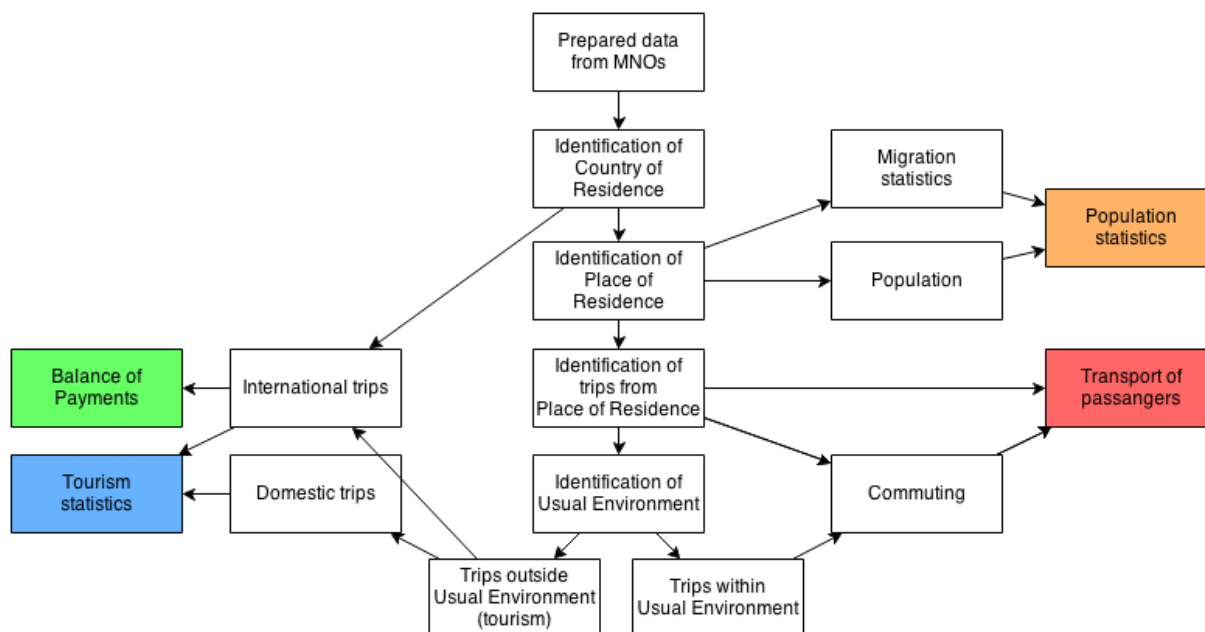
Figure 31. Simplified scheme involving joint data processing for tourism and other statistical domains.

This chapter describes the potential fields of official statistics that might benefit from using joint procedures and methodologies. By knowing the main definitions of variables in other fields, it is possible to assess the differences from definitions used in tourism statistics. Also it is possible to assess if and in what extent the procedures of tourism statistics should be amended to estimate the new variables based upon the mobile positioning data. The current chapter specifies for each field: a) the main definitions, b) currently used data sources, c) problems with current data sources and d) how mobile positioning data could benefit.

4.1. Balance of Payments, Travel Item

4.1.1. Definitions

Table 15. Concepts and definitions used in Balance of Payments

Variable	Definition
International travel	A visit of a resident of one country to another country, while the length of the visit is shorter than one year and the main purpose of the visit is not gainful activities or long-term studies in the country of destination.
Overnight visit	The traveller spends at least one night in a foreign country.
Same-day visit	The traveller arrives and leaves the foreign country in the same day, while spending at least four hours there.
Inbound travel	Same-day and overnight visits of non-residents to a reference country.
Outbound travel	same-day and overnight visits of residents to foreign countries
Resident	A person with permanent residence in a reference country.
Non-resident	A person with permanent residence outside a reference country.
Total number of visits	The number of same-day and overnight visits in total.

Total length of visits (days)	The length of overnight visits (days) and same-day visits in total.

Travel is one of the main items within the services account of the balance of payment and is closely connected to tourism statistics as it involves data on inbound and outbound tourism travel and visits. The concepts used in balance of payments have several differences when compared to the concepts used in tourism statistics. As the services account reflects those services that are sold to and purchased from non-residents by residents, these differences lie mostly in the definitions of the nature of the expenditures. In addition, for Balance of Payment purposes the main interest are visits and not trips, as it is in tourism statistics. The concepts of travellers and visitors are similar to those used in tourism statistics.

4.1.2. Data Sources

The number of travellers could be estimated using data from various different surveys, like border crossing, transport, accommodation, travel and household surveys etc. Mobile positioning data can be used to provide an estimate of the number of travellers, of visits, and of the duration of the visit to be used in models that estimate travel services in the Balance of Payments. Mobile positioning data cannot provide any information on the expenditure of visitors; a separate survey, administrative or other data source is required to estimate the expenditure of visitors.

4.1.3. Problems with Current Data Sources

Surveys are expensive and time consuming. Household surveys are often of a low reliability and low level of detail (geographical, etc.) due to the small sample size and low response rate. However, as pointed out above, surveys cannot be entirely replaced by mobile positioning data because the latter does not contain any information on visitor spending.

4.1.4. How Mobile Positioning Data Could Be Used for BoP Calculations

Mobile data presents options when it comes to identifying those indicators that are connected to trips taken by non-residents within the country of reference and by residents of the country who are travelling outside the country of reference. As the Balance of Payment travel item is tightly connected with tourism statistics, often using same data sources (albeit with several differences), then mobile data can provide indicators on both inbound travels of

non-residents and their travels abroad. The following indicators can be used from mobile positioning data:

Inbound:

- number of visits foreign countries;
- country of origin of the travellers;
- number of days in the country;
- same-day visits, overnight visits.

Outbound:

- number of visits to foreign countries;
- number of days spent in the foreign country;
- same-day visits, overnight visits.

Described statistics based upon mobile positioning data is compiled for Central Bank of Estonia and used as one data source in Balance of Payments since 2009 (see Use Case 7 in Report 1).

Mobile positioning data has several limitations and does not provide full information source for Balance of Payments travel of item. As mentioned above, the expenditure information has to be collected from other sources. Also, mobile data has the similar limitations as with tourism data in terms of qualitative information (the purpose of the visit, methodological limitations concerning the coverage issues and other quality aspects).

4.2. Tourism Satellite Account

4.2.1. Definitions

Table 16. Concepts and definitions used in the Tourism Satellite Account.

Variable	Definition
Tourism gross value added	The overall proportion of gross value added that has been generated by tourism industries and other economic industries that directly serve visitors in response to internal tourism consumption.
Net taxes on tourism products	Taxes on tourism products minus subsidies on tourism products (calculated using the ratio of output).
Tourism gross domestic product	The sum of tourism value added for all industries plus tourism taxes less subsidies on products.
Tourism gross fixed capital formation	The total value of net acquisition and/or improvements to tourism-specific and other produced fixed assets by tourism industries as well as by public and private sector producers outside them.
Domestic tourism	The activities of resident visitors of a country who are travelling to and staying in places

	only within a country but outside their usual environment
Domestic visitor	A resident of a country who travels within a country and visits places outside their usual environment.
Inbound tourism	The activities of non-resident visitors who are travelling to and staying in a country and outside their usual environment.
Internal tourism	Tourism conducted by visitors, both resident and non-resident, within the economic territory of the country of reference
Non-resident	A person who possesses citizenship of a foreign country or, in the case of citizenship being absent, who has permanent residence abroad.
Non-resident visitor	A person who travels to a country other the one in which they have their place of residence but outside their usual environment for a period not exceeding twelve consecutive months and for whom the main purpose of their visit is something other than the exercise of an activity remunerated from within the place that is being visited.
Resident	Citizens residing in a country; aliens residing in a country with a permanent residence permit or with a temporary residence permit for at least one year; citizens studying or receiving medical treatment abroad, regardless of the length of their studies or for medical treatment; diplomats, military personnel, staff from consular and other representative offices as well as their family members who are staying abroad and who enjoy immunity and diplomatic privileges; ship's crews, seasonal and border workers, regardless of the duration of their residence within the territory of a foreign country.
Tourism	The activities of persons travelling to and staying in places which are outside their usual environment for not more than twelve consecutive months for the purposes of activities involving leisure or business or for other purposes.
Tourism consumption	The total consumption expenditure made by a visitor or on behalf of a visitor for or during their trip and stay at a destination.
Visitor	A person travelling to a place other than that of their usual environment for less than twelve consecutive months and whose main purpose for visiting is something other than the exercise of an activity that is remunerated from within the place visited.

4.2.2. Data Sources

The main data sources for Tourism Satellite Accounts are: the Tourism Demand Survey, the Accommodation Survey, Structural Business Statistics, the supply and use tables, the symmetric input-output table, the balance of payments, plus administrative Income and Social Tax Declarations data, etc.

4.2.3. Problems with Current Data Sources

Household surveys are often of a low reliability and low level of detail due to the small sample size and low response rate. For that reason the estimation of domestic tourism expenditures by same-day visitors is complicated. It is also complicated to estimate the internal and external parts of the missions cost due to lack of relevant information. Business surveys are expensive and time consuming.

4.2.4. How Mobile Positioning Data Could Be Used for TSAs

For Tourism Satellite Accounts, mobile data can only present opportunities in domestic and inbound travels as somewhat better alternative mobility information. TSA requires a much more in-depth understanding of the purpose and the consumption aspect of tourism (from the demand side of things), something that mobile positioning data cannot provide. However, in terms of the quantitative numbers of domestic and inbound travellers, mobile data can provide value in terms of tourism statistics.

4.3. Transport of Passengers

4.3.1. Definitions

Table 17. Concepts and definitions used in the field of transport of passengers.

Variable	Definition
Domestic traffic	The origin and the destination point of a trip are inside of a country.
International traffic	The origin and/or destination point of a trip is outside of a country.
Passenger	Any person who makes a journey by public or private vehicle.
Passenger trip	The combination between the place of embarkation and the place of disembarkation of passengers conveyed by a vehicle.
Passenger traffic volume (in passenger-kilometres)	One passenger-kilometre is the transport of one person across a distance of one kilometre.

In transport statistics, passenger trips have a wider meaning than that used in tourism statistics as they cover any trip regardless of the purpose of the trip or whether or not it is outside the usual environment. The important aspect of transportation statistics is that they include classification by transport modes (road, rail, inland waterways, sea, or air). It is very difficult to identify a transport mode from mobile positioning data; indefinite estimations based upon movement patterns can be conducted concerning the transport modes of travellers. Because transport statistics concerning passengers includes all travellers, the processing of such data has to run parallel with tourism statistics, not including the exclusion of non-tourism activities. This can provide a total number of travellers for national trips (foreigners and residents) and trips taken outside the country.

4.3.2. Data Sources

The main sources are different business surveys, which are usually separate for each mode of transport. The data is collected either from all businesses or from sampled businesses only.

Within the European Union, passenger transport statistics are produced according to following legislative acts:

- DIRECTIVE 2009/42/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 6 May 2009 on statistical returns in respect of carriage of goods and passengers by sea;

- REGULATION (EC) No 1365/2006 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 6 September 2006 on statistics for goods transport by inland waterways and repealing Council Directive 80/1119/EEC;

- Regulation (EC) No 91/2003 of the European Parliament and of the Council of 16 December 2002 on rail transport statistics;

- Regulation (EC) No 437/2003 of the European Parliament and of the Council of 27 February 2003 on statistical returns in respect of the carriage of passengers, freight and mail by air.

4.3.3. Problems with Current Data Sources

Currently there is no legal basis for collecting Road Passenger and Passenger Mobility data at the EU level. Some countries may have national surveys but the data from those surveys is not harmonised between countries. Key Passenger Mobility indicators of interest are as follows:

1) Share of trip makers in the total reference population in one day
2) Daily trips modal share
3) Average number of trips/person/day
4) Average travel distance (km) /person/day
5) Average total travel time (min)/person/day
6) Average travel time & distance by purpose of travel
7) Average travel time & distance by mode of travel

4.3.4. How Mobile Positioning Data Could Be Used for Transport Statistics

As mentioned above, mobile positioning data can provide some useful indicators on travels taken by people within the country and across borders. However, it is rather difficult to identify the travel mode, except in some special cases, and also largely the purpose of the travel, both of which are important parts of the statistics. The ability to identify the place of

residence for domestic subscribers and the country of origin of foreign tourists is helpful. The trips of residents to and from their homes can describe the total number of movements, the duration of travel, and the separation between the movement and visit segments of the travel, which are represented in time and space (total mileage and km per day).

4.4. Population, Migration and Commuting Statistics

4.4.1. Definitions

Table 18. Concepts and definitions used in the field of population and migration statistics.

Variable	Definition
Place of residence	The place at which a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holiday, visits to friends and relatives, business, medical treatment or religious pilgrimage or, in default, the place of legal or registered residence.
Immigration	The action by which a person establishes their place of residence in the territory of a Member State for a period that is, or is expected to be, of at least twelve months, having previously been usually resident in another Member State or a third country.
Emigration	The action by which a person, having previously been usually resident in the territory of a Member State, ceases to have their place of residence in that Member State for a period that is, or is expected to be, of at least twelve months.
Citizenship	The particular legal bond between an individual and their state, whether acquired by birth or through naturalisation, whether by declaration, choice, marriage or other means according to national legislation.
Emigrant/out-migrant	A person who moves from one settlement unit to reside in another settlement unit /country and has registered the respective departure in the settlement unit of their previous residence. From the perspective of the previous residence, a person who has moved to reside in another settlement unit within a country is called an out-migrant.
Immigrant/in-migrant	A person who has moved to reside in another settlement unit/country and has registered their arrival from the settlement unit /country of their previous residence. From the perspective of the new residence, a person who has moved from another settlement unit within a country is called an in-migrant.
Internal migration	Changes of residence from one settlement unit to another within a country. A change of residence from one settlement unit to another within the same county is called intra-county migration, and a change of residence from a settlement unit of one county to a settlement unit of another county is called inter-county migration.
Migration	A cross-border change of the residence from one settlement unit to another
Migration event	A change of permanent residence across the border of a settlement unit.
Place of residence	Area or settlement in which the person resides according to their statement or according to the Population Register.
Time of migration	A date at which a person registered their arrival in or departure from their residence in a settlement unit.
Commuting	Everyday working or studying in a municipality which differs from the municipality in which their residence itself resides.

During the calculation of the usual environment (for domestic data), additional calculations are also made for the place of residence and other meaningful locations

(frequently visited places), and these can be of interest in population statistics. A number of people living (also holiday homes, secondary homes) and working (work-time anchor points can also be places for studying) in specific administrative units can be estimated based upon mobile data. For a longer period of time, domestic migration statistics can be presented based upon the data for changing the place of residence and other anchor points. Although mobile data does not provide the accuracy of a census, a comparison of mobile data with a census and the population register shows a high correlation (see Figure 25 and Figure 26) and is much more dynamic in the aspect of rapid changes. The question of whether such data can be used in official statistics is valid, as the methodologies and concepts are very different.

4.4.2. Data Sources

The main sources for population statistics are Population Census, administrative population register and household surveys for commuting.

4.4.3. Problems with Current Data Sources

A weakness of administrative population register may be low reliability of the data when it comes to the residence and where it involves a change of the residence in the case of the voluntary reporting of this data to the registry. The population census provides high quality data on migration and commuting, but the disadvantage is the rare interval between censuses; a census is conducted only every decade or more. Household surveys have all the disadvantages of sample surveys, the data is often of a low reliability and a low level of detail due to the small sample size and low response rate.

For estimating commuting and other regional statistics based upon mobile positioning data, the use of the administrative source when it comes to the actual location of the workplace might be necessary. In population statistics, mobile positioning data can be used in areas such as, for example, the assessment of the most frequent place of locality or the residence.

4.4.4. How Mobile Positioning Data Could Be Used for Population Statistics?

Mobile positioning data can benefit the population statistics in several aspects; however, with limitations similar to use of other statistics. Based upon the identification of the place of residence of the subscribers and along with the estimation, a number of residents living in different administrative levels can be presented (see Figure 32). The calculation of

the usual environment makes it possible to estimate the number of commuters between and within municipalities. These indicators can be compiled over a period of time representing the change in time of the population statistics. Long-term migration (changes of residence) can be detected more frequently and some statistics is more representative than 'low'. Although mobile positioning data cannot easily be linked to actual people, there is a technical possibility use mobile positioning data in connection with registry-based census if legislation allows for such a linkage. This would add an additional personalised data source that can be used in parallel with other data. The benefit of using mobile data is that it is more dynamic in its nature. The timeliness of the data (with its rapid data updates) provides the possibility of presenting the situation or changes in population much more quickly than the traditional 'slow' registries. Such opportunities can be appreciated by the users of population and other statistics in transportation planning, urban planning, regional development and other fields.
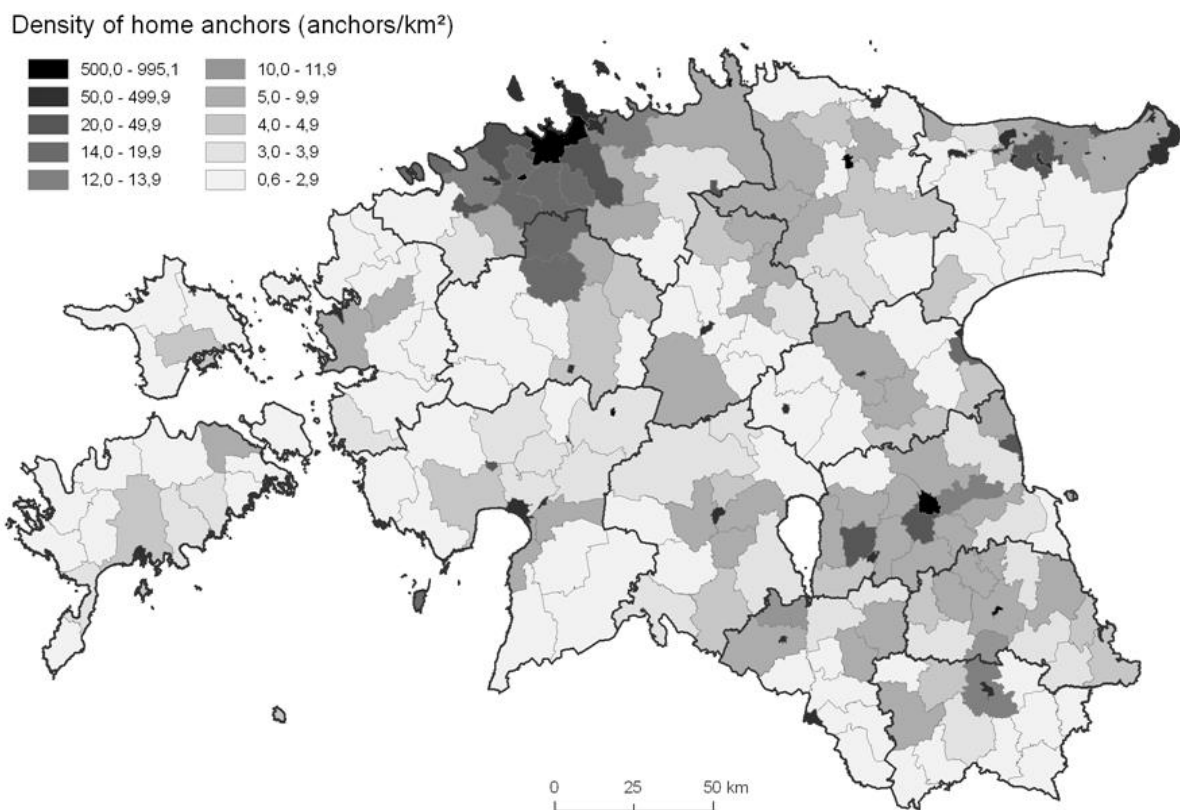

Figure 32. The density of home anchor points in municipalities modelled using the passive mobile positioning data (Positium 2012).

5. Conclusions

This report describes the methodology for producing tourism statistics based upon passive mobile positioning data. The methodology covers all the processes from frame formation and data processing to aggregation. This description of the methodology could be taken as a step-by-step guide for how to produce tourism statistics, it being rather detailed but at the same time general enough for broad use.

The methodology assumes that several variables relevant exist in the dataset for the forming of new variables and that the activities of anonymous subscribers can be followed over a longer period of time in order to establish their residency and/or usual environment. The latter assumption regarding 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 available data describes only a short period then processing errors become an issue 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.

It should be noted that the assessment of the quality of the final outcome relies heavily on existing external information e.g. accommodation statistics, transport statistics, information about the mobile operators' market share and number of subscribers, etc. Overall, the estimation part of the described methodology is very general allowing a frequentist or bayesian framework and various models to be applied as the data available at the estimation stage can be very different from country to country.

The quality of the described methodology is assessed by looking at the output validity, accuracy and comparability. The accuracy, especially coverage issues, is the most problematic quality aspect for this type of data. It is problematic because many components 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 based upon Estonian mobile positioning data to describe and illustrate the problem.

To assure comparability, quality assessments should be carried out when changes in methodology occur, just as it should be carried out for traditional surveys. In addition, it is important to be ready to update the methodology if changes in the telecommunication technology or in the data structure occur.

The possibility of using the same framework for producing tourism statistics and other official statistics is also explored in this report. Here, it is mainly the coherence of definitions used in different fields of official statistics that is discussed. The results show it is possible to use joint methodologies that will add value to this data source.

References

Legal documents

Regulation 692/2011 of the European Parliament and of the Council of 6 July 2011 concerning European statistics on tourism and repealing Council Directive 95/57/EC: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:192:0017:0032:EN:PDF

Publications

Ahas R. et al. (2010) Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1), pp 3-27

Ahas, R., Silm, S., Järv, O., Saluveer E., Tiru, M. 2010. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones , Journal of Urban Technology, 17(1): 3-27.

Eurostat (2013a) Methodological manual for tourism statistics, Version 2.1. Cat. No: KS-GQ-13-007-EN-N

Eurostat (2013b) Rolling Review Tourism Statistics. Executive Summary. March 2013.

GSMA (2011) European Mobile Industry Observatory report: http://www.gsma.com/publicpolicy/wp-content/uploads/2012/04/emofullwebfinal.pdf

Positium (2012) Mobiilsusuuringute metoodika rakendusuuring POSMETRAK. 2009-2011.

Special Eurobarometer 414 (2014) e-Communications Household Survey and Telecom Single Market Survey Roaming Results. TNS Opinion & Social at the request of European Commission: https://ec.europa.eu/digital-agenda/en/news/e-communications-household-survey-and-telecom-single-market-survey-roaming-results-special

Särndal, C.-E. and Lundström, S. (2005) *Estimation in Surveys with Nonresponse*. Chichester: Wiley & Sons

UN (2008) International Recommendations for Tourism Statistics, Studies in Methods, Series M, No. 83/Rev.1

Annex 1. Limitations with Regard to Variables of Regulation 692/2011

Table A1. Complete list of variables given in Regulation 692/2011 and the comment on the possibility of providing the required variables based upon the mobile positioning data.

ANNEX I INTERNAL TOURISM	
Section 1. CAPACITY OF TOURIST ACCOMMODATION ESTABLISHMENTS	
...	It is not possible to distinguish either the accommodation facility or any specific information concerning the establishment. Therefore, variables concerning the accommodation establishment will not be covered further in this table.
Section 2. OCCUPANCY OF TOURIST ACCOMMODATION ESTABLISHMENTS (DOMESTIC AND INBOUND)	
A. Variables and breakdowns to be transmitted for annual data	The difference between headings A and B in Section 2 is the same as between data compiled initially (rapid key indicators) and the final compilation of the data (recompiled later with more accurate results). For example, rapid key indicators might provide inaccurate information about the length of trips (tourists still travelling during the data update so that a final trip duration can only be established later); or the identification of usual environment (migration to a new residence which is considered as a tourism trip due to data cut-off).
(1) At regional NUTS level 2 and at national level	
NACE 55.1; 55.2; 55.3	It is not possible to distinguish the type of accommodation (NACE classification 55.1; 55.2; 55.3). Therefore this classification will be avoided in further nomenclature.

Number of nights spent by residents at tourist accommodation establishments	The number of nights spent by domestic tourists outside their usual environment. Not possible to distinguish between spending the night in the accommodation establishment or non-rented accommodation.
Number of nights spent by non-residents at tourist accommodation establishments	The number of nights spent by inbound tourists outside their usual environment. Not possible to distinguish between spending the night in the accommodation establishment or non-rented accommodation.
Arrivals of residents at tourist accommodation establishments	The number of domestic trips/visits by tourists outside their usual environment. Not possible to distinguish between spending the night at the accommodation establishment or non-rented accommodation. A trip can involve several visits in different locations. At the national level one trip is classed as being one visit to a country of reference.
Arrivals of non-residents at tourist accommodation establishments	The number of inbound trips/visits of the tourists outside their usual environment. Not possible to distinguish between spending the night in the accommodation establishment or non-rented accommodation. A trip can involve several visits in different locations. At a national level one trip is equivalent to one visit.
Net occupancy rates of bed places	Not identifiable
Net occupancy rate of bedrooms	Not identifiable
Breakdowns	See Section 3 of Annexe I
(2) At national level	All same as (1) of Section 2 of Annexe I
...	
B. Variables and breakdowns to be transmitted for monthly data at national level	All same as (1) of Section 2 of Annexe I. This can be considered as being the initial compilation of the data
...	
C. Limitation of the scope	Not identifiable
D. Rapid key indicators	See comment at Heading A of Section 2 of Annexe I
Section 3. CLASSIFICATIONS TO BE APPLIED FOR SECTION 1 AND SECTION 2	
A. Type of accommodation	Not identifiable
B. Type of locality (a) degree of urbanisation of the tourist accommodation establishments	

- densely populated area	Visits to specific geographical locations can be identified; the classification of the area is out of scope of mobile positioning data itself - it is the question of the classification of the geographical location itself.
- intermediate area	Same as above
- thinly populated area	Same as above
C. Type of locality (b) location close to the sea of the tourist accommodation establishments	
- coastal	Visits to specific geographical locations can be identified; the classification of the area is out of scope of mobile positioning data itself - it is the question of the classification of the geographical location itself.
- non-coastal	Same as above
D. Size class of the accommodation establishment	Not identifiable
E. Countries and geographical areas	Classification of countries is out of scope of the mobile positioning data. Mobile positioning data provides the specific country of the tourist. In case of domestic data, the (e.g. a municipality) can be identified.
Section 4. INTERNAL TOURISM IN NON-RENTED ACCOMMODATION	
A. Variables to be transmitted for annual data	
[optional] Number of tourism nights spent in non-rented accommodation during the reference year.	As it is not possible to identify the establishment where the tourists spend their nights, it is not possible to distinguish between visits to rented and non-rented accommodation from mobile positioning data itself, without comparing accommodation statistics.
B. Breakdown	
[optional] The variable listed under heading A shall be broken down by country of residence of the visitors as far as Union residents are concerned, while visitors residing outside the Union shall be grouped in a single additional category.	See above. Possible with comparison to accommodation statistics.
ANNEX II NATIONAL TOURISM	
Section 1. PARTICIPATION IN TOURISM FOR PERSONAL PURPOSES	
…	Not possible to determine whether trips are for personal or business purpose.
Section 2. TOURISM TRIPS AND VISITORS	

MAKING THE TRIPS	
A. Variables to be transmitted	
Variables and categories	
1. Month of departure	OK. Month (also day, week) of the moment when a person leaves one's usual environment.
2. Duration of the trip in number of nights	OK. Duration can be presented in number of total nights spent, days spent outside the usual environment as well as the number of nights and days spent in specific location during a stay.
3. [Only for outbound trips] Duration of the trip: number of nights spent on the domestic territory	OK. Possible to distinguish separate sections of the trip (e.g. domestic section leaving the usual environment, outbound section in foreign countries, domestic section before arriving to the usual environment).
4. Main country of destination	OK. Also possible to distinguish the transit countries and several destination countries (long trips with different longer stays in different countries).
5. Main purpose of the trip	Identification of the purpose of the visit is very limited (work/study trips can be derived through spatio-temporal algorithms with some limitations, also possible to determine visiting e.g. big concerts or when the visit can be linked with a specific time and location).
...	
6. [Only for trips for personal purposes] Type of destination, with multiple answer possibilities	Not possible to determine whether trips are for personal purposes unless the visit can be linked to a specific event in the destination (e.g. big concert).
(a) City	Only for domestic trips: visits to specific geographical locations can be identified; the classification of the area is out of scope of mobile positioning data itself - it is a question of the classification of the geographical location itself.
(b) Seaside	Same as above
(c) Countryside (including lakeside, river, etc.)	Same as above
(d) Cruise ship	Not possible to determine exclusively as the cruise ship is not associated rather with the sea passage and not with a specific location.
(e) Mountains (highlands, hills, etc.)	Same as above (a).
(f) Other	Same as above (a).

7. [Only for trips for personal purposes] Participation of children in the travel party	Not possible to determine if children were included in the trip.
...	
8. Main means of transport	Identification of the mode of the transportation is very difficult and might be possible only in specific situations.
…	
9. Main means of accommodation	Not possible to identify
...	
10-15 (booking the trip)	Not possible to identify
...	
16-19 (expenditure)	Not possible to identify
...	
20. Profile of the visitor: gender	Only if the socio-demographic data is provided by the MNOs.
...	
21. Profile of the visitor: age, in completed years	Only if the socio-demographic data is provided by the MNOs.
22. Profile of the visitor: country of residence	Fully identifiable.
23-25 (socio-demographics)	Not possible to identify unless such information is provided by MNOs; however, MNOs might not hold such information.
B. Limitation of the scope	
The scope of observation shall be all tourism trips with at least one overnight stay outside the usual environment by the resident population aged 15 and over and which ended during the reference period. The data on the population under 15 years of age can be transmitted separately on an optional basis.	Age can be differentiated only if the socio-demographic data is provided by the MNOs. Otherwise, scope can be defined as all tourism trips with at least one overnight stay outside the usual environment by the resident population of all ages and which ended during the reference period.
C. Periodicity	OK
Section 3. SAME-DAY VISITS	
A. Variables and breakdowns to be transmitted on an annual basis (outbound same-day visits)	
Variables:	
1. Number of outbound same-day visits for personal purposes	Not possible to distinguish between trips for personal or professional purposes. All trips/visits presented.
2. Number of outbound same-day visits for professional reasons	Not possible to distinguish between trips for personal or professional purposes. All trips/visits presented.
3. Expenditure on outbound same-day visits for personal purposes	Not possible to identify.

4. Expenditure on outbound same-day visits for professional reasons	Not possible to identify.
Breakdowns:	
(a) by country of destination	OK.
(b) by expenditure category: transport, shopping, restaurants/cafés, other	Not possible to identify.
Socio-demographic breakdowns:	See socio-demographic breakdowns in heading A of Section 1.
...	
B. Variables and breakdowns to be transmitted on a triennial basis (domestic same-day visits)	
Variables:	
1. Number of domestic same-day visits for personal purposes	Not possible to distinguish between trips for personal or professional purposes. All trips/visits presented.
2. Number of domestic same-day visits for professional reasons	Not possible to distinguish between trips for personal or professional purposes. All trips/visits presented.
3. Expenditure on domestic same-day visits for personal purposes	Not possible to identify.
4. Expenditure on domestic same-day visits for professional reasons	Not possible to identify.
Breakdowns:	
(a) by expenditure category: transport, shopping, restaurants/cafés, other	Not possible to identify.
Socio-demographic breakdowns:	See socio-demographic breakdowns in heading A of Section 1.
...	
C. Classifications to be applied for socio-demographic breakdowns	See socio-demographic breakdowns in heading A of this Section 1.
...	
D. Limitation of the scope	
The scope of observation shall be all same-day visits outside the usual environment by the resident population aged 15 and over. The data on the population under 15 years of age can be transmitted separately on an optional basis.	Age can be differentiated only if the socio-demographic data is provided by the MNOs. Otherwise, scope can be defined as all tourism trips with at least one overnight stay outside the usual environment by the resident population of all ages and which ended during the reference period.
E. Periodicity and first reference periods	OK.
...	