When mobile network operators and statistical offices meet - integrating mobile positioning data into the production process of tourism statistics

Paper submitted for Session 1 [Mobile phone data for tourism statistics] of the 14th Global Forum on Tourism Statistics (Venice, Italy, 23-25 November 2016)

Lead authors: Gerdy Seynaeve[1], Christophe Demunter[2],

Co-authors: Freddy De Meersman[1], Youri Baeyens[3], Marc Debusschere[3], Pieter Dewitte[3], Patrick Lusyne[3], Fernando Reis[2], Hannes I. Reuter[2], Albrecht Wirthmann[2].

[1] *Proximus (Belgium)*
[2] *Eurostat*
[3] *Statistics Belgium*

Keywords: big data, mobile phone, mobile positioning, signalling, tourism statistics

1. INTRODUCTION

Since the pioneering work of Ahas et al. [1] in exploring the use of mobile phone data for statistics (in particular tourism statistics), nearly ten years ago, the constellation has tremendously changed.

Up to now, experiments with using mobile phone data were largely limited to the use of call detail records (CDR). A comprehensive overview of this source, and the methodological issues, opportunities and weaknesses was reported in the Eurostat *Feasibility study on the use of mobile positioning data for tourism statistics* [2].

On the one hand, changed behaviour of mobile phone users is more and more affecting the relevance of call detail records (alternative non-SIM based messaging services, alternative voice or video call systems), which necessitates auxiliary data to assess the selectivity bias of this source, and to correct/calibrate for this bias[1]. On the other hand, mobile network operators are shifting to the use of other data sources available within their network infrastructure.

As one of the follow-up actions to the above mentioned feasibility study, Eurostat, Proximus and Statistics Belgium set up a partnership. This partnership has the ambition to develop methodologies for population statistics and tourism statistics using data held by Proximus, the leading mobile network operator in Belgium. The project, which runs from end 2015 till end 2016, focuses on signalling data. This data – as compared to call detail records (CDR) -

[1] See also the paper submitted by ISTAT submitted for Session 1 of the 14th Global Forum: *How many SIM cards in your luggage? A strategy to make mobile phone data usable in tourism statistics* [3]

increases the number of observations with a factor 10, allowing for much more precise information regarding time and place. Intermediate results are presented in this paper.

The outcomes of this project are expected to have relevance for other countries, in terms of methodology but also in terms of institutional or organisational set-up of the partnership.

2. METHODOLOGICAL APPROACH AND DATA SOURCES

The methodology in this research project has as a main innovation the shift to signalling data. To link the theoretical insights with practical data compilation, Section 3 of this paper compares mobile phone data with existing official statistics for Belgium in the area of tourism. In the context of this research, the scope focuses on outbound tourism with overnight stays made by residents of Belgium.

2.1. Mobile phone data

2.1.1. From call detail record to signalling information

The use of administrative records initially stored for billing purposes, i.e. call detail records (CDRs), has been intensively used for monitoring population, mobility and tourism [2][2]. However, a main shortcoming of this approach is that it is highly dependent on the behaviour of the subscriber: whether a phone is observed on the network is heavily influenced by the intensity of the mobile phone use. In the case of tourism, for instance, it has proven difficult to distinguish between same-day visits and trips with one or more overnight stays.

Network probing systems, on the other hand, offer a much better temporal granularity (and indirectly also a better geographical granularity since the increased number of observations will capture more changes in location at cell level). These capture all signalling events, billable and non-billable. The amount of useful signalling events is up to ten times higher as compared with CDRs [4]. The Proximus network detects the position of a device minimum every three hours (unless the device is switched off). For devices with data 'on', this drops to approximately 1 hour. In practice, through usage of the phone for calls, messages or data, devices are observed with a much higher frequency. During daytime hours, 7 out of 10 devices are observed after one hour during a given timeframe; 1 out of 3 devices are detected within 15 minutes. The mix depends on the actual usage and on the technology (e.g. 4G devices are typically giving more location points than 2G devices).

Signalling data opens a new perspective, in particular for monitoring day-to-day mobility. The cascade system used for this study determines firstly the usual place of residence of the subscriber (roughly approximated by the place where the subscriber is most often observed at 4a.m. in the morning over a given period of time) – for preliminary results regarding measurement of the present population, see [4]. Secondly, all movements away from the usual place of residence are observed. Thirdly, those movements that are outside the usual environment, thus relevant for tourism statistics are separated from the day-to-day activities. In a way, only the *noise* is relevant for tourism statistics.

[2] An overview of use cases – up to 2013 - can found in Report 1 (Stock-taking) of the feasibility study.

The above is relevant for all devices on a network (devices on their own network, and devices roaming on the network – so-called 'roaming in'). For 'roaming out' (devices outside their home network), the visited network has to fetch the profile of the user from the home network. These types of signalling events are useful in the case of outbound tourism. If the device is powered on during the change of country, the timestamp of this procedure is usually very close to the actual entry time (in practice the device will lock on the new network at the time it has lost connection to the previous network).

2.1.2. Focus on outbound trips

In the current experimental phase, but also linked to computing time and data protection, the focus is on tourism trips with a destination outside Belgium by subscribers to the Proximus network. Limiting the research to the trips abroad also simplifies the delineation of the usual environment as all trips within the country of residence are by default excluded from the scope.

Linked to the storage policy of the operator and to the resources needed for the calculations, the data used for this study refers to trips observed in the six months' period April 2016 to September 2016. The scope of observation covers trips that ended during this period (but may have started before April 2016), in line with the approach in official statistics[3].

2.1.3. From signalling events to trips and nights

From observations of international location updates, it is possible to create a concept of "trip" whereby a trip starts when leaving the home country (i.e. the first transaction abroad), and ends when returning to the home country (i.e. the first transaction back in the home country). If a trip contains multiple countries, the country with the longest stay can be called the destination (comparable with the "main destination" used in official statistics).

The concept of number of nights stayed can be implemented from the start and end timestamp of the trip. For this study, the number of nights was calculated as the number of hours divided by 24, or in case the duration of the trip is less than 24 hours, at least 10 hours with a return point after 04:00a.m. (i.e. assuming an overnight stay).

2.1.4. Evaluating the usual environment

To remove frequent trips to the same country from the observations - i.e. to detect and remove trips presumably taken within the usual environment – the algorithm calculates for each trip to a given destination the number of trips observed for the same mobile phone (SIM card) to that destination during a given reference period (for this study set at 253 days). The threshold was (arbitrarily) set at 5 trips. In other words, if the number of trips to a given destination for a given subscriber exceeds 5, all these trips will be considered as part of the usual environment and removed for the further analysis of outbound tourism.

A frequency analysis of how often a SIM is seen at a destination during a reference period is discussed in section 3.1 of this paper.

Future research and fine-tuning of the methodology relating to the usual environment could include applying dynamic thresholds (for instance treating short trips of 1 to 2 nights differently from longer trips of 7 nights or more, to the same destination by the same

[3] As laid down in Regulation 692/2011 concerning European statistics on tourism.

subscriber). This should allow to better separate 'usual' and 'non-usual / tourism' trips; indeed a subscriber can frequently travel to a neighbouring country within his/her usual environment but in addition spend a true holiday trip in that same country. In the current approach, all trips are excluded once the threshold is reached. It should be pointed out that the operator can only detect the country of destination without further detail on the regions visited. In this context, a trip just across the border cannot be separated from a trip 300 or 400km away within the same neighbouring destination country.

Another relevant parameter to be further examined is the reference period. Indeed, a mobile device needs to be observed during a certain timeframe in order to determine the user's usual environment. The length of this reference period (e.g. 3, 6 or 9 months) will not only have an impact on the quality of the data but also on the resources (computing time) and feasibility of access (data protection clearance). Different scenarios (parameter settings) could be tested to pave the way for a more regular data production in a later phase. However, the current reference frame of around 250 days is assumed to be sufficiently close to the generally accepted recommendations and definitions[4] for tourism statistics.

2.1.5. Data exchange and privacy

To overcome privacy issues, adequate methods need to be found to preserve privacy protection. In that sense, the analysis of the mobile phone data for this paper was done at the operator side, while the analysis of the official statistics was done by Eurostat in its secure environment for micro-data. The joint analysis of the data was done on the basis of aggregates, without exchange of raw data that could be traced back to an individual Proximus subscriber or respondent in the tourism survey.

Internally at Proximus, for privacy protection, data is calculated on anonymised datasets, all user specific information is deleted, and data is aggregated into groups of at least 50 before extrapolation.

Defining proper aggregation levels is a major challenge when using big data sources, trying to maximise the usefulness *and* minimise any privacy risks. It is a crucial element in making mobile phone data work in a regular data production environment where official statisticians need to assess and continuously monitor the quality of the sources used.

2.2. Official statistics

2.2.1. Tourism statistics collected and compiled by Statistics Belgium

Regulation (EU) 692/2011 established a common framework for European statistics on tourism. Member States transmit harmonised data to Eurostat on trips made by their residents. This data source will be the benchmarking reference for the results stemming from the mobile phone data.

In the context of this Regulation, Statistics Belgium conducts sample surveys on the population in which respondents reply to questions about tourism trips taken and the characteristics of those trips (main destination, duration, purpose, expenditure, means of transport, means of accommodation, combined with socio-demographic variables such as

[4] See for instance Eurostat's Methodological Manual for Tourism Statistics [5] or the International Recommendations for Tourism Statistics [6].

gender or age). This results in a sample of around 10 000 tourism trips for each reference year. For this research, only the variables main country of destination and duration were relevant.

To align the scope with that of the mobile phone data, only the subsample of outbound trips between April and September was used in the analysis (*n*=4919).

2.2.2. Inconsistencies between the two data sources

It is essential to point out that the official statistics refer to April-September 2015 while the mobile phone data refer to 2016. Official statistics for 2016 are *not yet* available[5], while mobile phone data for 2015 can *no longer* be retrieved (because of storage policy and regulation). The hypothesis is that the analyses undertaken in Section 3 of this paper discuss rather the structure of tourism trips than the absolute volume and should therefore be only slightly affected by the different reference year[6].

2.3. Equal treatment of sources

The research in this paper does not define the existing official statistics to be the so-called "ground truth" but challenges this data with insights gained from the mobile phone data. Many methodological issues inherent to new data sources exist in one way or another also in traditional statistical techniques and as such affect the quality of statistics produced pursuant to those techniques. In the end, both approaches lead to an estimate, not to the true value.

2.3.1. Recall bias or memory effect in traditional surveys

In tourism statistics, a particular problem is the recall bias or memory effect. Respondents reporting over a three months' reference period are likely to forget shorter trips. A Spanish study quantified this effect to cause a 15 to 20% underestimation of the number of trips made [7].

Machine-based data, e.g. mobile phone data, could contribute to overcoming this kind of measurement error.

2.3.2. Selectivity bias in mobile phone data: "the number you are trying to reach is not available at this moment, please try again later"

Once the barrier of getting access to mobile phone data is taken, other challenges arise. While traditional surveys suffer from non-response, mobile phone data can face comparable methodological weaknesses.

Firstly, mobile network operators have information on their market share (and the inverse of the market share would be a good first grossing-up factor to get to population estimates), but the market share can differ by region or by socio-economic group.

[5] Data for the reference year 2015 was timely transmitted to Eurostat by Statistics Belgium but unreleased at the time of writing.

[6] Based on an observed average annual growth rate of 1.4 % in the number of outbound trips by residents of Belgian over the period 2004-2015 *(source: Eurostat)*.

Secondly, penetration rates of mobile phone possession and use are not exactly 100%. This issue is similar to the issue of over-coverage or under-coverage of the sampling frame in traditional surveying.

Thirdly, subscribers may or may not make/take phone calls, send/receive message, connect to Wi-Fi networks depending on the time of the day or the place (e.g. while on holidays) or even switch off their device(s). This phenomenon, too, is comparable to the non-response or non-contacts that survey statisticians have to deal with. For the specific case of analysing outbound tourism through network signalling, bias could be introduced by devices being turned off before, or during tourism trips abroad, meaning country/network changes could go unnoticed.

The above sketched problems lead to a selectivity bias that needs to be taken into account when using mobile phone data. While it is generally expected that the use of big data can contribute to a reduction of respondent burden due to surveys, paradoxically the early phases of big data will see the necessity to collect auxiliary information via surveys to enable data scientists to correct for unevenly distributed market shares, for variable use patterns or for non-observation of devices.

Within the European Statistical System[7], initiatives are being set up to collect this kind of auxiliary information to support big data sources, not only for mobile phone data[8] but also for e.g. social media. Available data shows that the effects can be very significant. Recent data by ISTAT [3] evaluates the mobile phone use by Italian residents during tourism trips. Nearly 90% of respondents made calls during trips within Italy but the intensity of use dropped to just over 70% for trips abroad. On the other hand, Wi-Fi internet (not SIM) appears to be relatively higher during trips abroad, possibly avoiding perceived roaming charges.

3. RESULTS

3.1. How often is a SIM seen at the same foreign destination?

Figure 1 and Figure 2 analyse the frequency for continents and for EU Member States respectively. For each destination (continent/country), the distribution of the number of SIMs observed at that destination is given, in terms of the number of times a SIM is observed during the 250 days window.

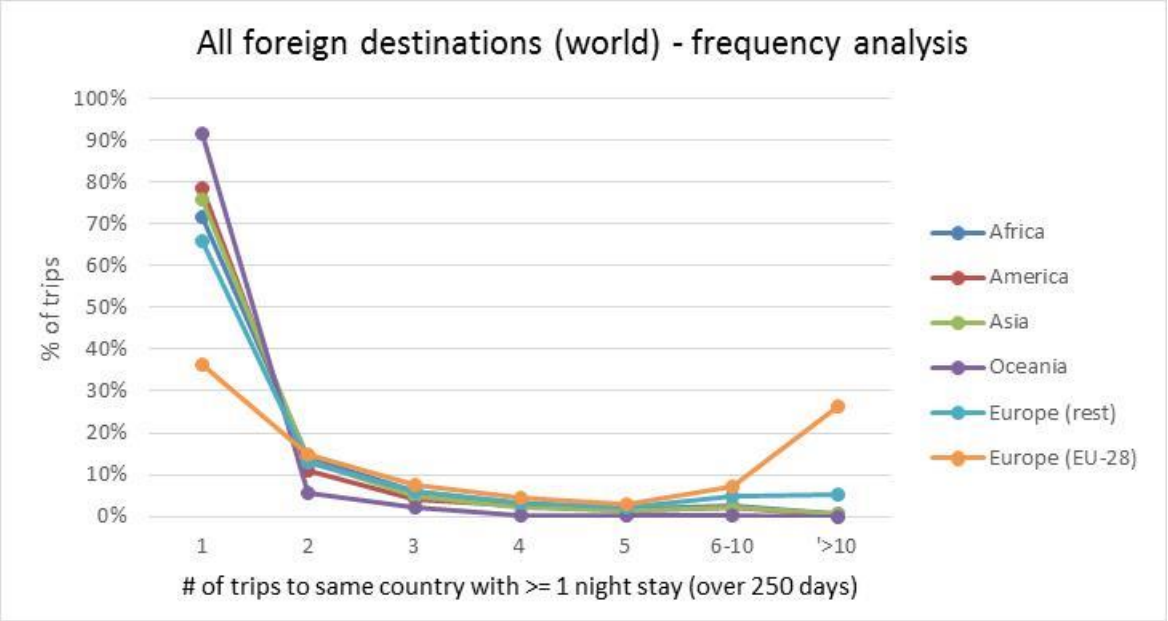
For instance, within the group of SIMs observed at destinations within the EU-28, 36 % are observed only once, 15 % are observed two times, 8 % are observed 3 times, 5 % are observed 4 times, 3 % are observed 5 times, 7 % are observed between 6 and 10 times, 26 % are observed more than 10 times in other EU countries during the 250 days period (see orange line in Figure 1).

[7] The European Statistical System (ESS) is the partnership between the Community statistical authority, which is the Commission (Eurostat), and the national statistical institutes (NSIs) and other national authorities responsible in each Member State for the development, production and dissemination of European statistics.

[8] A specific module on mobile phone use is being developed for the Community survey on ICT usage in households and by individuals (to be included by Member States on a voluntary basis).

For other continents, including European countries outside the EU-28, the majority of SIMs observed at these destinations are observed only once during the reference period. Less than 3% of SIMs observed in the remote continent of Oceania are observed on more than 2 trips to Oceania in the reference period (5.5 % twice, 91.8 % only once).

Figure 1: Distribution of SIMs, in terms of the number of times a SIM is observed during a 250 days period, by continent of destination

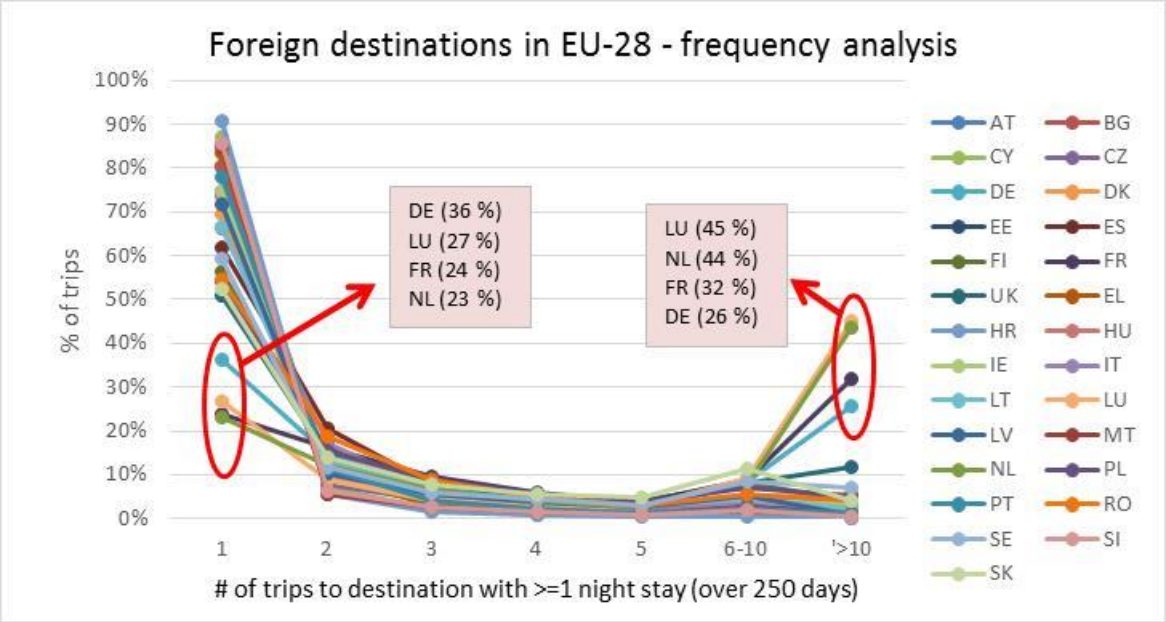


Source: Proximus

Proximity of the destination also plays a role when looking at non-Belgian destinations within the European Union (EU-28), as illustrated by Figure 2. Belgium's neighbours Germany, France, Luxembourg and the Netherlands show a distinctive pattern with relatively few SIMs observed in these countries being observed only once (i.e. few unique visits during the reference period) while relatively many SIMs observed in these countries are observed more than 10 times (i.e. many repeat visits during the reference period, making the destination country somehow part of a usual environment). For instance, 23 % of SIMs observed in the Netherlands during the 250 days, are observed only once but nearly half (44 %) are observed more than 10 times.

On the other hand, more remote and/or smaller destinations will most likely see a visiting Proximus SIM card only once during the reference period, for instance Croatia (91 %), Cyprus (87 %), Slovenia (86 %), Malta (85 %) or Greece (83 %).

Figure 2: Distribution of SIMs, in terms of the number of times a SIM is observed during a 250 days period, by EU Member State of destination



Source: Proximus

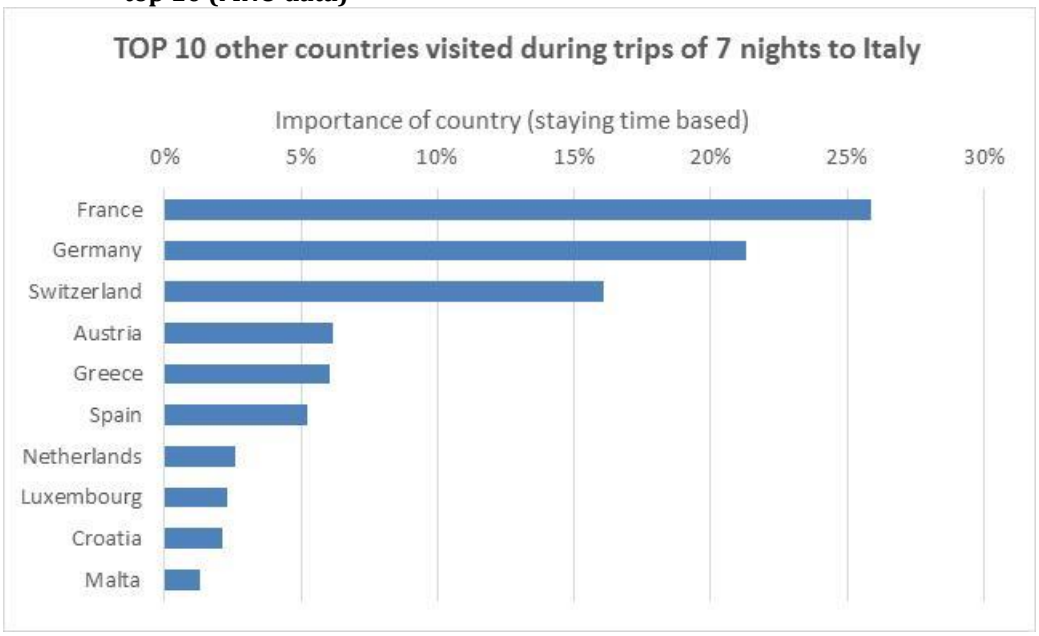
3.2. Main destination versus other destinations visited and transit countries

For reasons of respondent burden, tourism surveys in general only record the *main* destination of a trip and – for instance – all expenditure reported for the trip will be allocated to this single destination. Mobile phone data can more easily keep track of the countries where a SIM was observed during a trip away from the country of residence. On 43 % of all trips with Italy as a main destination, the SIM was only observed in Italy while in 57 % of the trips to Italy the SIM was also observed in other countries (transit or secondary destinations).

This data is also used to determine the main country of destination on a trip outside Belgium: for each foreign country were the SIM is detected, the time is calculated; the country with the highest value is the main destination.

This section takes as an example the case of trips to Italy with a total duration of 7 nights. Figure 3 list the 10 most visited 'other' countries during a trip with Italy as main destination (these 10 countries accounted for 92% of the total time spent outside Italy during trips of 7 nights with Italy as main destination). The countries are ranked according to their share in the total time spent in other countries than Italy during trips to Italy, for example 26% was spent in France. Two groups of countries can be distinguished: transit countries for trips by road (or rail), namely France, Germany, Switzerland, Austria, the Netherlands and Luxembourg, and secondary destinations such as Spain, Greece, Croatia, Malta (for example multi-country Mediterranean cruises).

Figure 3: Other countries visited during trips (of 7 nights) with Italy as main destination, top 10 (MNO data)



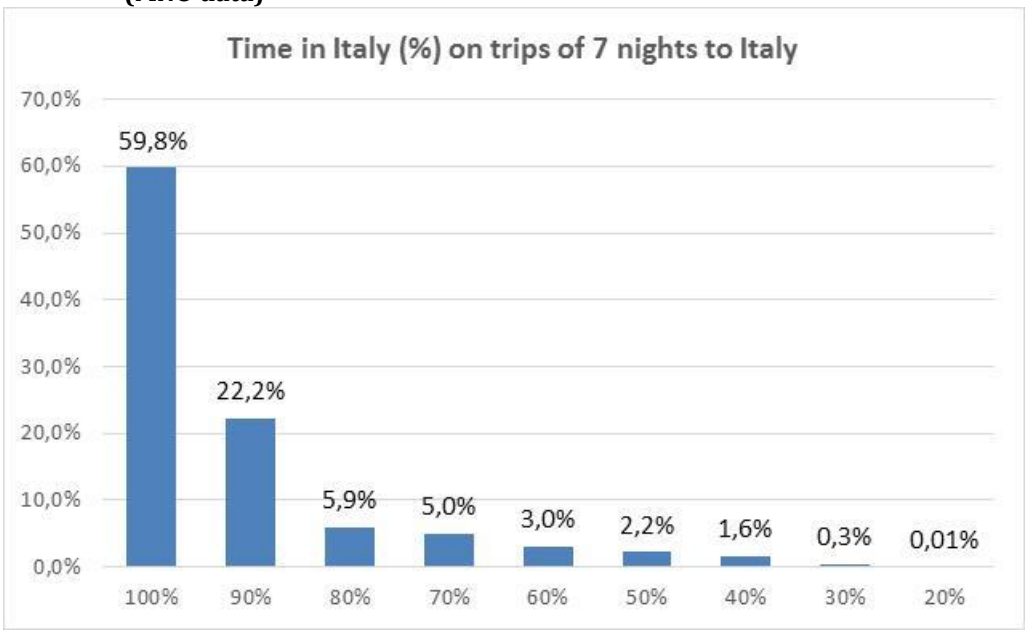
Source: Proximus

Another innovative analysis concerns the time actually spent in the main destination country (instead of allocating the entire duration to the main destination). Figure 4 shows, again for trips of 7 nights to Italy, the distribution of trips by decile of time actually spent in Italy. Nearly 60 % of trips of 7 nights were entirely[9] spent in Italy – i.e. the SIM was detected in no other country but Italy between leaving from and returning to the Proximus network. At the other end of the spectrum, the chart reveals that for a small part of the trips to Italy, less than half of the time was actually spent in Italy. While somehow counter-intuitive, it is possible that a SIM is detected during less than 25% at the main destination, namely in cases where the other destinations each accounted for even less[10].

[9] These are rounded groupings: "100%" means more than 95%, "90%" means between 85% and 95%, etc.

[10] Example from the dataset, trips spanning eight countries but Italy as main destination: IT(23%), HR(20%), SI(15%), HU(13%), AT(9%), CZ(9%), DE(9%), LU(0%) [note that the share don't add up to 100% due to rounding]

Figure 4: Distribution of trips (of 7 nights) to Italy, by decile of time actually spent in Italy (MNO data)



Source: Proximus

Mobile phone data can offer additional insights in the composition of trips in terms of countries visited. On the one hand, this can enrich the existing official tourism statistics by – for instance – better allocating expenditure to the different economies. On the other hand, this information can also enrich the mobile phone data – for instance as a proxy for estimating the means of transport used on the trips (by comparing the data with the obvious itineraries to reach the country of destination, or by assuming that trips to non-neighbouring countries without detection of the SIM in other countries were made by plane).

3.3. Distribution of outbound trips by destination (EU-28)

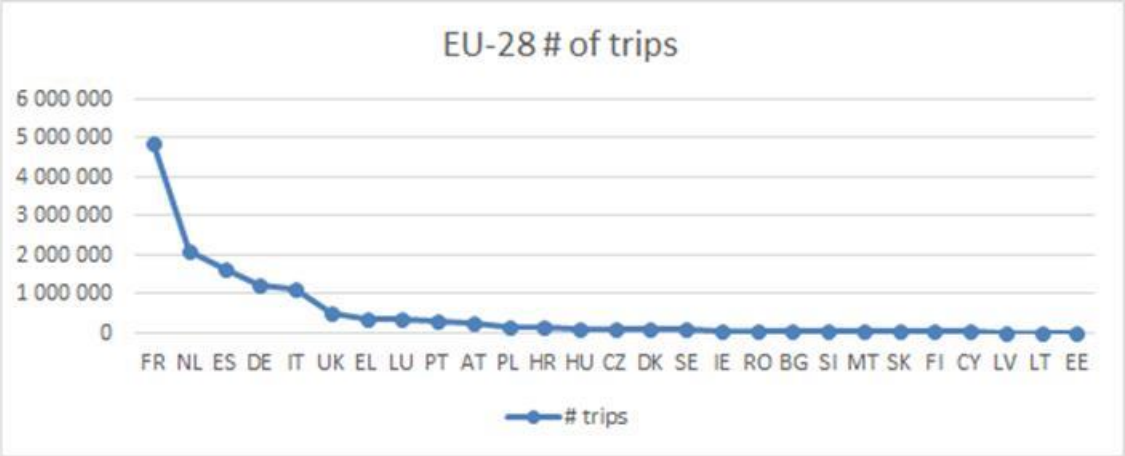
Exploring the mobile phone data, Figure 5 ranks the EU-28 countries in decreasing number of trips made by residents of Belgium. The top 5 is a mix of neighbouring countries (France, the Netherlands, Germany) and popular destinations (Spain, Italy). Smaller and/or more remote countries of destination appear at the bottom of the ranking.

In terms of nights spent (see Figure 6), Spain and Italy jump over Germany and the Netherlands in the ranking, not surprisingly since both countries (together with #1 France) are typical destinations for long(er) holidays.

The picture changes when focusing on the ratio of the nights and the trips, i.e. the average length of stay (in nights). More remote destinations such as Romania, Croatia, Greece or Bulgaria are now in the lead with an average duration of more than 10 nights (rounded) while nearby countries like the UK, Germany, the Netherlands and Luxembourg record durations of less than 5 nights on average.

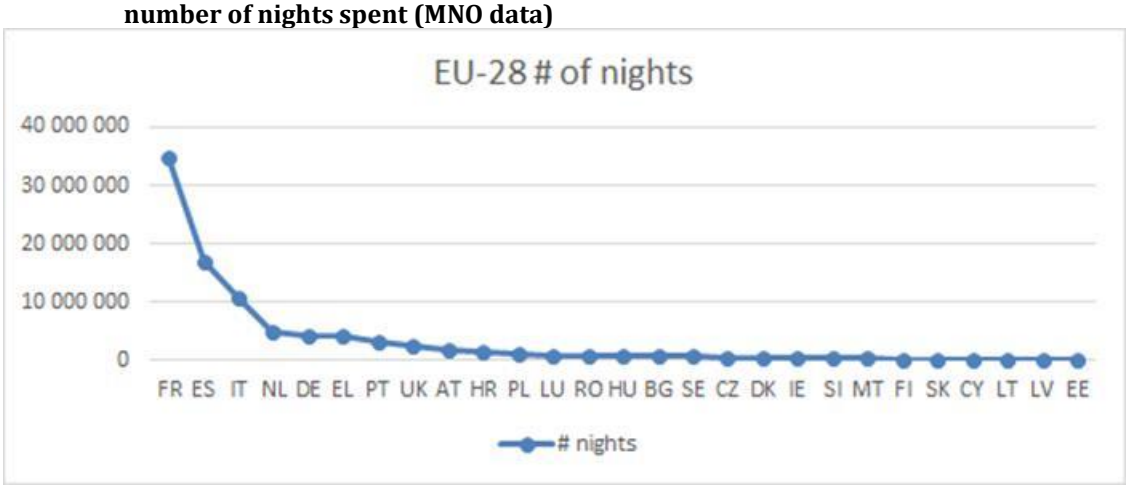
Similar data is available from traditional official tourism statistics. However, the data presented in Figures 5, 6 and 7 refer to the current year whereas official tourism statistics will not be available before spring 2017.

Figure 5: Ranking of EU-28 countries as destination for Belgian outbound trips, according to number of trips (MNO data)



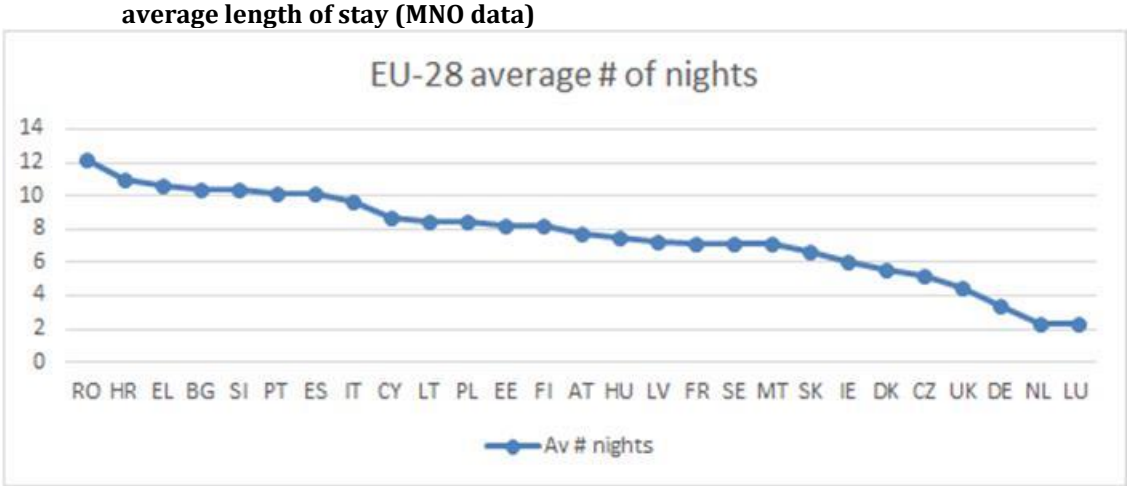
Source: Proximus

Figure 6: Ranking of EU-28 countries as destination for Belgian outbound trips, according to number of nights spent (MNO data)



Source: Proximus

Figure 7: Ranking of EU-28 countries as destination for Belgian outbound trips, according to average length of stay (MNO data)



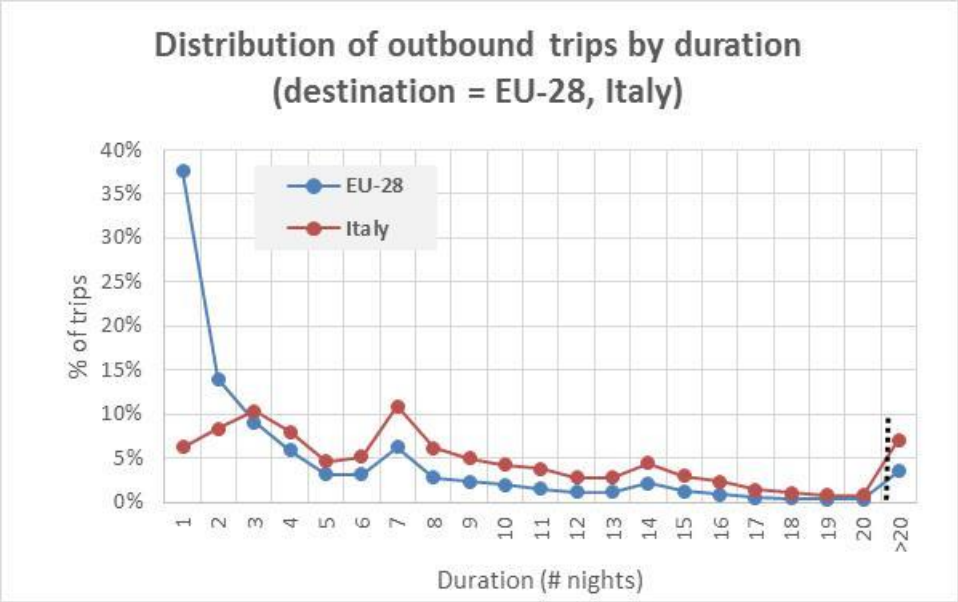
Source: Proximus

3.4. Distribution of outbound trips by duration of the trip

Figure 8 shows the distribution of trips made by Belgian residents by duration, respectively trips to any other EU-28 country and trips to Italy, as observed in the mobile phone data.

The typical/traditional holiday trip duration of 7 or 14 nights (i.e. 1 week or 2 weeks) can be clearly observed in both series shown in the graph, with relative peaks for these values. The main difference between trips to the EU-28 and trips to Italy more specifically, can be seen among the shorter trips, with far less trips of 1 or 2 days to Italy as compared with the total of all EU foreign destinations. The big share of trips with 1 overnight stay can partly be explained by the big weight of neighbouring countries.

Figure 8: Distribution of outbound trips, by duration of the trips (MNO data)



Source: Proximus

Figure 9 compares the distribution of trips by duration for the two data sources, mobile phone data and official tourism statistics. Two striking differences can be seen.

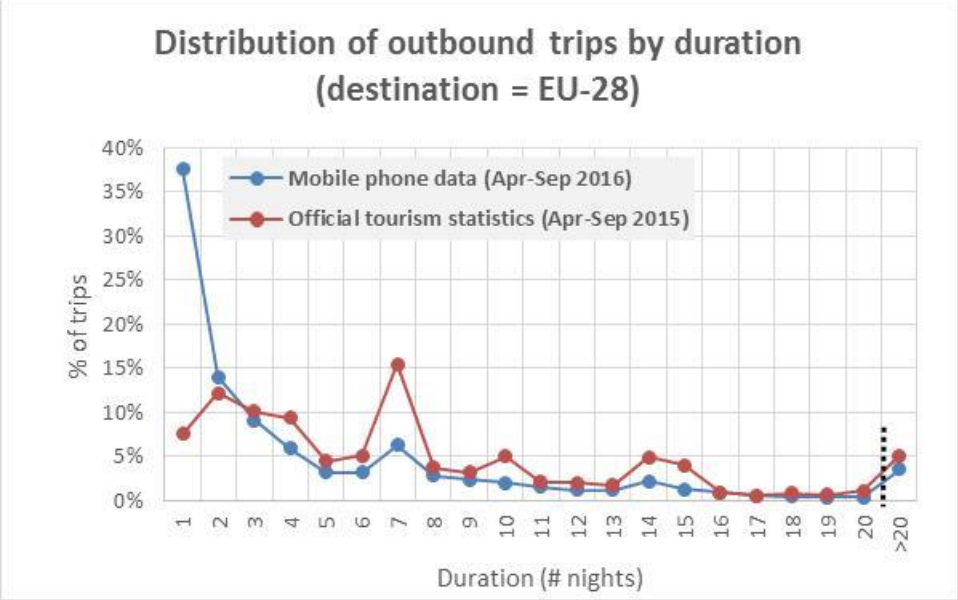
Firstly, the peak values for a trip duration of exactly 7 or 14 days are more pronounced for the survey based data. Here, the reason could be the rounding bias arising when respondents don't remember the exact duration (6? 7? 8? nights) and approximate ("a week" – recorded as 7 days).

Secondly, the weight of ultra-short trips of one overnight stay is much higher in the mobile phone. Two possible explanations could be the parameter settings for the mobile phone data (namely minimum duration of 10 hours and return after 4am) and the memory effect in the traditional tourism surveys (the shorter the duration, the more likely the respondent forgets to report the trip).

Actions for improvement (regarding the mobile phone data) could include a revision of the threshold (number of trips to the same destination in the reference window) or a more dynamic approach that is more strict/exclusive for short trips and more inclusive for longer trips that are more likely to be tourism trips taken outside the usual environment.

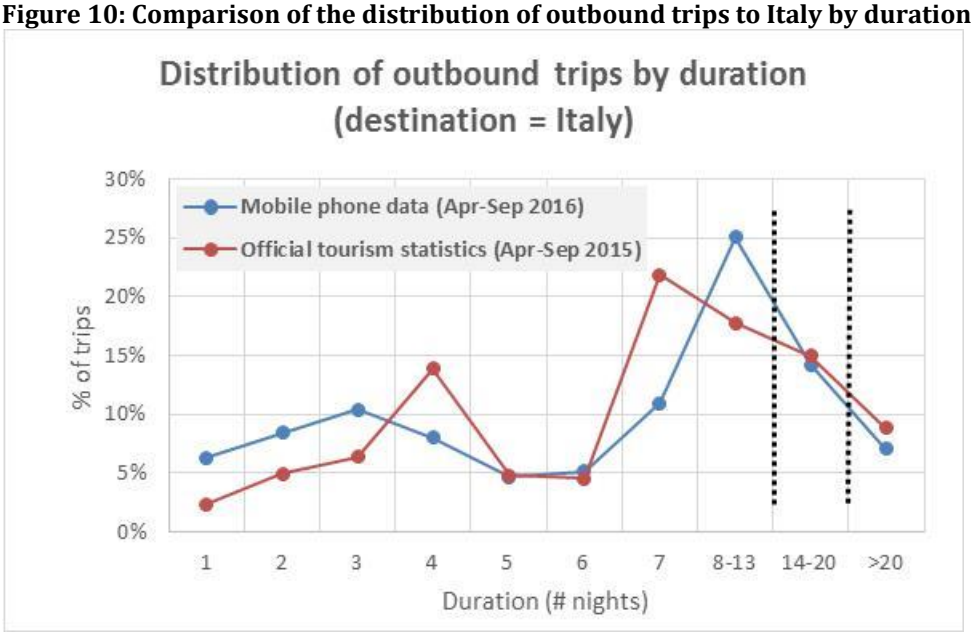
Figure 10 repeats the comparison, but focusing on trips to Italy only[11]. For a relatively far away destination like Italy, the series tend to be more coherent. The relation between the memory effect and the distance of the destination most likely plays an important role.

Figure 9: Comparison of the distribution of outbound trips to EU-28 by duration of the trips



Source: Proximus & Eurostat / Statistics Belgium

Figure 10: Comparison of the distribution of outbound trips to Italy by duration of the trips



Source: Proximus & Eurostat / Statistics Belgium

3.5. Comparison of volumes of trips and nights

The previous sections already highlighted areas for improvement or known methodological weaknesses for both data sources. These issues have an impact on the feasibility of

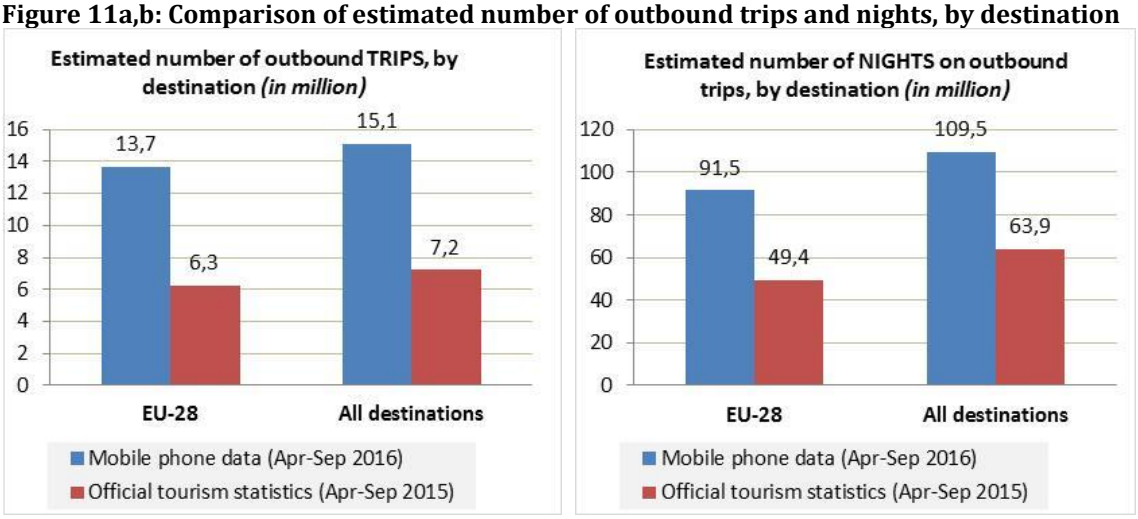
[11] Note that the different values of duration were regrouped due to small frequencies in the survey data (the necessary number of observations is minimum 50; durations were regrouped into classes that respect this rule).

comparing mobile phone data and official tourism statistics, with the risk of comparing apples and oranges. However, for the sake of completeness of the analysis in this paper, and to foster future research in this area, this closing section will take a look at the volumes, i.e. a comparison of the absolute numbers in terms of estimated trips and nights.

Figures 11a and 11b compare the estimated number of outbound trips made by residents of Belgium between April and September. The estimate for trips obtained from mobile phone data is (more than) twice as high as the official statistics, for nights the deviation is a bit less pronounced but still largely exceeding 50%.

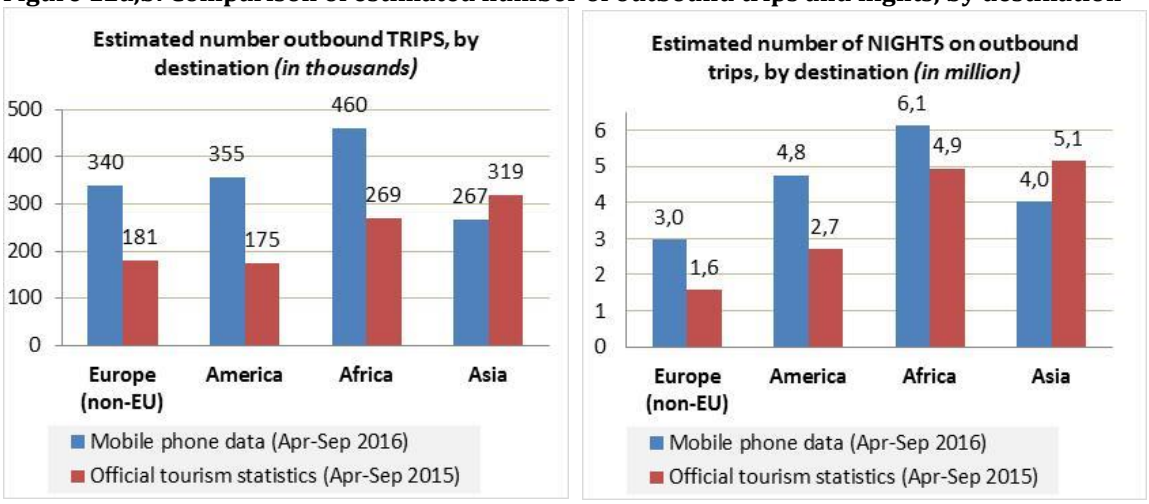
Figures 12a and 12b deepen the analysis to the level of the continents[12], but the conclusions remain similar (except for Asia).

Figure 11a,b: Comparison of estimated number of outbound trips and nights, by destination



Source: Proximus & Eurostat / Statistics Belgium

Figure 12a,b: Comparison of estimated number of outbound trips and nights, by destination



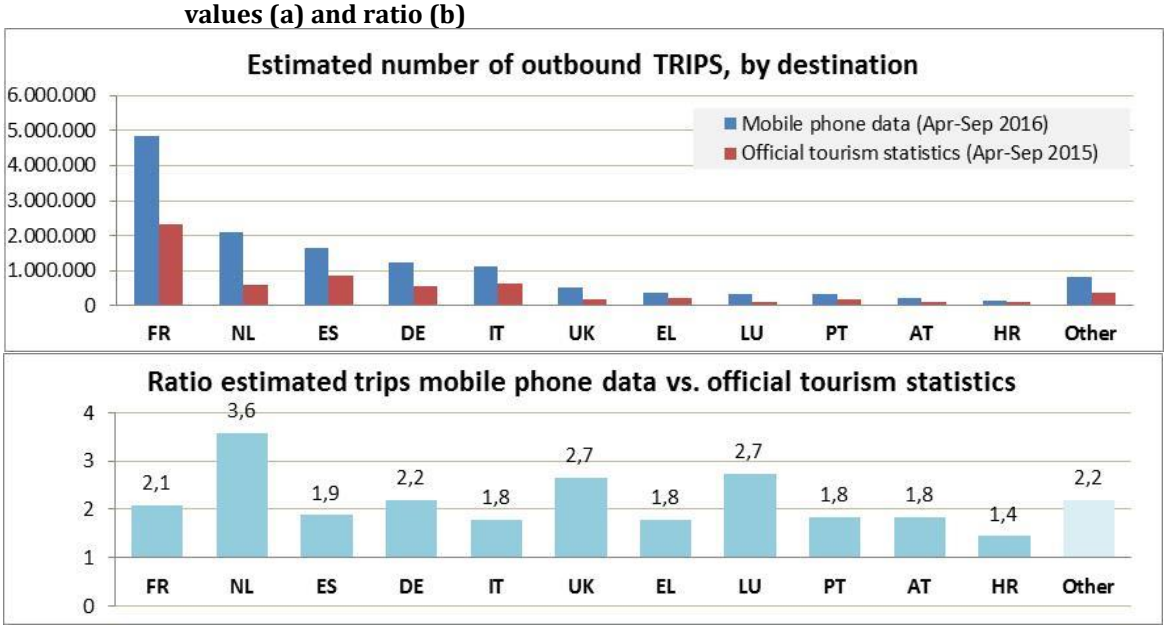
Source: Proximus & Eurostat / Statistics Belgium

[12] Note that the mobile phone data file Turkey and Russia under Asia. While this is not the case in official tourism statistics, the geographical regrouping of the latter has been aligned to the first, for the sake of meaningful comparison in this paper.

The detail for the countries of the European Union[13] is shown in Figure 13a (trips) and Figure 14a (nights). For both charts, the ratio between the mobile phone data estimate and the estimate of official statistics is presented in Figure 13b and Figure 14b respectively.

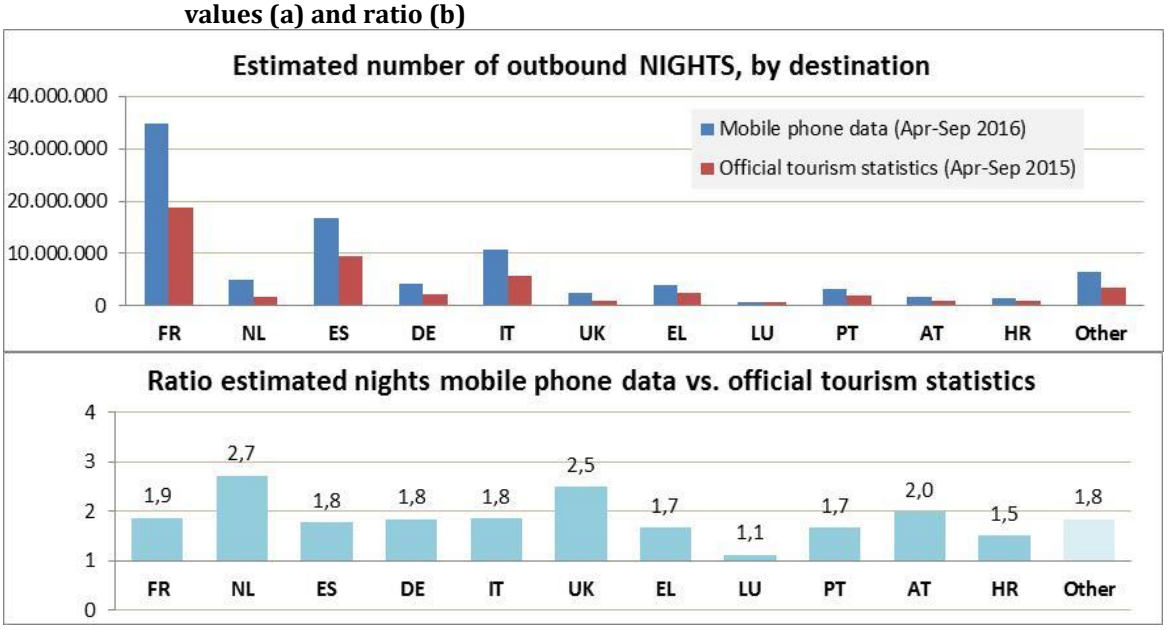
Within the EU, differences can be observed for countries located close to Belgium (relatively high ratios – the Netherlands, Luxembourg, the UK, Germany) and destinations further away (lower ratios – Croatia, Greece, Portugal, Italy, Austria).

Figure 13a,b: Comparison of estimated number of outbound trips, by destination: absolute values (a) and ratio (b)



Source: Proximus & Eurostat / Statistics Belgium

Figure 14a,b: Comparison of estimated number of outbound nights, by destination: absolute values (a) and ratio (b)



Source: Proximus & Eurostat / Statistics Belgium

[13] Due to sample size limitations, the survey data cannot be released for an important number of countries. The estimates for these countries were grouped in the category 'Other'; the mobile phone data was treated accordingly.

The differences between the estimates obtained from the two sources used are big, but the differences seem to be systematic (although the proximity to Belgium tends to play a role).

The systematic differences make it possible to use the mobile phone data for trend analysis, contrary to the analysis of the volume of tourism. Given the freshness of mobile phone data – the current analysis uses data which includes September 2016 – this opens opportunities for increasing the timeliness of tourism statistics by using mobile phone data as short-term intermediate source or auxiliary source.

In terms of volume, intensive work is needed to explain the differences between the series, assess the possible shortcomings of both series and adjust where possible and necessary.

A number of issues can be put forward as explanations. Firstly, the two series differ in scope in the sense that official tourism statistics only cover the population aged 15 or over. Age will certainly induce a selectivity bias in mobile phone data (e.g. infants do not have mobile phones) but the effect is expected to be smaller. Other socio-demographic characteristics of Proximus subscribers may also play a role. Proximus entered the market relatively early which could introduce a bias towards proportionally more business subscribers, older subscribers or 'wealthier' subscribers (depending on the extent to which there is still an effect in 2016 in a more mature, saturated market where subscribers can easily switch from one provider to another).

Secondly, the selectivity bias can have an impact. The mobile phone data from Proximus (observed phones/SIMs) is extrapolated to the known population of Belgium (who may or may not use a mobile phone). One could expect a link between not using a mobile phone and not travelling. For illustration, if data for all three mobile network operators would be available, no grossing up on the basis of the market share would be necessary. The sum of the three sources would give the population of observed SIMs, which is expected to be different from the resident population. Auxiliary information on mobile phone use can produce correction factors (see also Section 1).

Thirdly, the use of mobile phone data is still in an experimental phase and scenario testing is ongoing. This will allow to fine-tune the parameter setting in the algorithms and lead to more accurate estimates. As is the case in a survey setting, the delineation of the usual environment seems to be a major challenge when using mobile phone data. M2M (machine-to-machine) applications could also lead to a bias, these are filtered out but additional checks could be useful.

Fourthly, official tourism statistics could benefit from correcting for the recall bias in order to come closer to the true value. Research in other countries has shown that the memory effect leads to a significant underreporting of trips.

Fifthly, known threats to the quality of survey based statistics can be further verified for tourism statistics. The way that non-response is dealt with, can impact the overall results. Is non-response random or are persons who did not travel during the reference period more likely to drop out (because they don't feel concerned by the subject of the survey)?

4. CONCLUSION

This paper explored the possibility of using mobile phone data (in particular signalling data) for tourism statistics, with a focus on Belgium. The research is conducted in the framework of a partnership between a mobile network operator (Proximus) and statistical services (Eurostat, Statistics Belgium). Although a discussion of the partnership is out of the scope of this paper, it has proven to be a success factor in this project (which has as an aim the set-up of a regular data production system).

The first results are promising and demanding at the same time. Mobile phone data is able to capture existing tourism definitions and concepts and can serve as a relevant data source for tourism statistics – currently rather for trend analysis than for estimating volumes.

Observed differences between mobile phone data and official tourism statistics are very high, but to a certain extent systematic. Understanding of the differences and finding ways to correct for these shortcomings (in both sources!) is the condition sine qua non for integrating this big data source into the production system of official statistics. The paper highlighted a number of issues to be explored; these will the subject of further research in the months and years to come.

Furthermore, the current analysis focuses on outbound tourism. Domestic tourism needs to be added to the work programme, this will necessitate more complex algorithms to determine the usual environment.

REFERENCES

[1] R. Ahas, A. Aasa, A. Roose, U. Mark, S. Silm, *Evaluating passive mobile positioning data for tourism surveys: An Estonian case study*, Tourism Management 29 (2008) 469–486.

[2] European Commission (EUROSTAT), *Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics* (2014).

[3] B. Dattilo, R.Radini, M. Sabato, *How many SIM cards in your luggage? A strategy to make mobile phone data usable in tourism statistics,* paper for the 14th Global Forum on Tourism Statistics, forthcoming (2016).

[4] F. De Meersman, G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis F., H.I. Reuter, *Assessing the Quality of Mobile Phone Data as a Source of Statistics*, paper for the European Conference on Quality in Official Statistics (2016).

[5] European Commission (EUROSTAT), *Methodological manual for tourism statistics - Version 3.1* (2014)

[6] United Nations / UN World Tourism Organisation, *International recommendations for tourism statistics* (2008)

[7] Instituto de Estudios Turisticos, *Memory Effect in the Spanish Domestic and Outbound Tourism Survey (FAMILITUR)*, paper presented at the 9th International Forum on Tourism Statistics (2008)