Determining an optimal time window for roaming data for tourism statistics

Martijn Tennekes
Statistics Netherlands
Email: m.tennekes@cbs.nl

May Offermans
Statistics Netherlands
Email: mpw.offermans@cbs.nl

Nico Heerschap
Statistics Netherlands
Email: n.heerschap @cbs.nl

***Abstract*—Mobile phone data can be used for making Official Statistics on various subjects, such as safety, mobility and tourism. One of the major challenges is the protection of privacy when analyzing and processing these data for publication or fundamental research. Using sensitive information sparingly is one of the principles that is often put forward in privacy discussions. The time dimension is key for analyzing Call Detail Records (CDRs) but also for privacy. Therefore, in this paper an optimal time frame is determined for roaming visitors that provides sufficient information to create official tourism statistics of high quality while ensuring privacy. For this study we used anonymized aggregated CDR data from Vodafone in collaboration with Mezuro. The results show that a time window of 15 days is sufficient for official tourism statistics. However, a time window of one month is preferable in order to compare mobile phone based tourism estimations with the current tourism publication numbers.**

I. Introduction

One of the advantages of using Call Detail Records (CDR's) is the enormous detail of information in time and space. This advantage provides new opportunities for National Statistical Institutes like Statistics Netherlands for creating statistics on daytime population and tourism [1-4]. New and more precise information becomes available making it possible to create fast, almost real time statistics on a national level with high regional detail. One of the major challenges when making statistics with an extensive time horizon is the management of privacy and data-access privileges. Since 2009, Statistics Netherlands has collaborated with Mezuro and Vodafone in research on CDR's [5]. No data provided by Vodafone may lead to the identification of a person or company. Anonymization, aggregation of data, and remote systems that leave all personal sensitive data at the providers data center are important measures to ensure privacy. Different privacy rules are applied to ensure that even if information from space and time is combined, no identification is possible. One of these privacy rules in the data provided by Vodafone and Mezuro is that a device receives an anonymized IMSI (International Mobile Subscriber Identity (IMSI) for a period with a maximum of 31 days. After this period, a new anonymized IMSI is created which ensures that the device cannot be tracked for a substantive period of time.

The authors and Statistics Netherlands would like to thank VodafoneZiggo and Mezuro (www.mezuro.nl) for providing data and processing the CDR's. The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands.

This ensures privacy and provides enough information for statistical research purposes. However, devices that have roaming switched on in the Netherlands, the period was set at 24 hours. The reasons for this shorter, more strict period is the fact that there is no legal framework on collecting data from foreign devices that roam on the Dutch network. Each tourist receives an SMS when it starts roaming on the network. However, form a legal point of view it is impossible to obtain full informed consent or to provide an opt-out option. Therefore, the 24 hour IMSI-hash is introduced. The disadvantage of this 24-hour privacy rule is that the roaming data does not meet the requirements needed for tourism statistics. For example, we can roughly estimate the number of Japanese tourists in Amsterdam within a 24-hour timeframe, but it is not possible to see how long these tourists stay and where they consecutively go when visiting the Netherlands. These statistics contain important information on, for example, national and local tourism planning, public transport and (international) festivals. Determining an optimum period for this privacy rule is therefore important. On the one hand, one wants to produce useful statistics for the general public, government, and industry, and on the other hand, only use information that is necessary to produce these statistics.

II. Data

For this specific analysis, we used two aggregated and anonymized data tables that were extracted from CDRs at the Vodafone data center from all roaming customers on the Vodafone network in the Netherlands during the period from May 1st 2014 to May 31st 2014. The minimum cell count of these CDR tables is 15. Table cells with values lower than 15 are coded as missing to prevent identification. For this 31 day period, the anonymized IMSI numbers were not renewed. Instead, all geolocation data has been left out. The specific descriptions of the two tables that we have used are as follows:

1) The first table contains three columns: first date, last date, and number of foreign roaming devices. The number of foreign users refers to the number of users for which an CDR event is logged every day between (and including) the first and the last date. We refer to this date range as *consecutive stay*. Note that we do not know the length of the consecutive stay if the first date is May 1st or the last date May 31st, because foreign users could

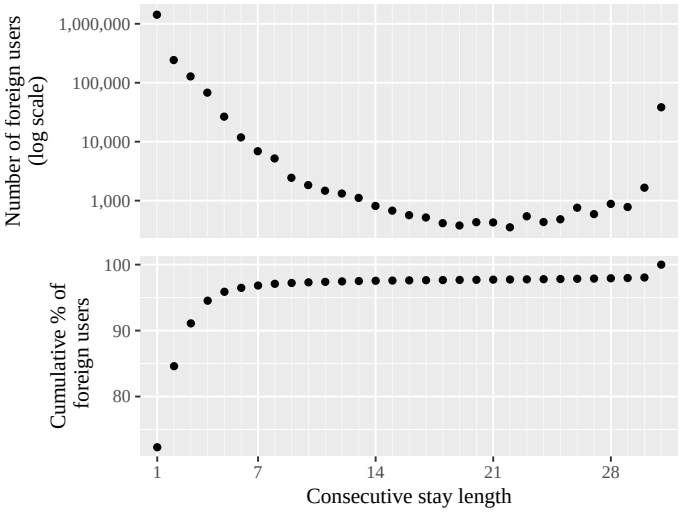

Fig. 1. Number of foreign users per consecutive stay length. Absolute numbers are shown above, cumulative below.

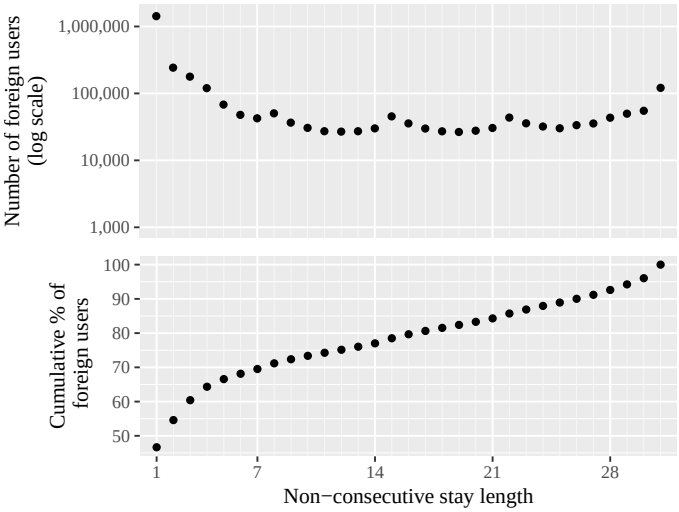

Fig. 3. Number of foreign users per consecutive stay length. Absolute numbers are shown above, cumulative below.

also have been logged before or after the observed time window.

2) The second table contains four columns: first date, last date, number of foreign users, and country of origin. In this table, the number of foreign users refers to the number of users for which the CDR events are logged between (and including) the first and last date, but not necessarily every day in between. Therefore, we refer to this date range as *non-consecutive stay*. To prevent low cell volume, the countries of origin were aggregated as follows: Belgium, Germany, France, United Kingdom, Eastern Europe, Southern Europe, Europe other, Middle-East, China-Japan-Australia-New Zealand, Asia other, North America, South America, and Inter Standard Roaming (ISR). Devices that roam with ISR use a different technology in their country of origin than GSM, such as CDMA. ISR is used in order to make roaming possible in Europe. ISR as a category contains users of different counties, mainly US, Canada, Japan and the Middle-East.

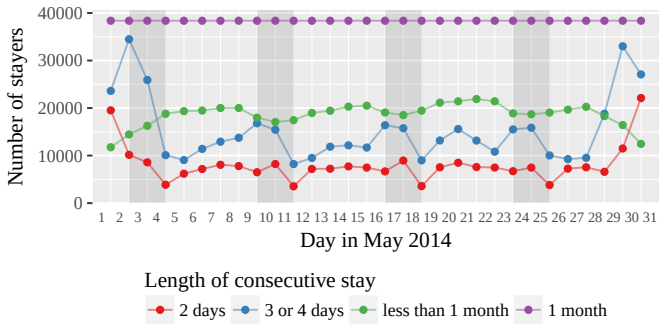

Fig. 2. Consecutive period of foreign users.

III. Results and Discussion

In this study, we focus on the length of stay of foreign roaming devices in the Netherlands. For tourism statistics, it is important to classify foreign users into standard categories, such as tourists, workers that cross the border on a regular basis, and normal border crossings (for example to do shopping). However, it was not possible to classify foreign users in this study, because it is unclear from the used data whether foreigners stay overnight in the Netherlands, or travel back and forth. It can generally be assumed that many people from Belgium and Germany, and, to a lesser extent France, will make day trips, whereas people from other countries will stay overnight in the Netherlands.

Figure 1 shows the number of users per consecutive stay length. For users who were logged on the 1st or the 31st of May, the consecutive stay can be longer, since they could have been in the Netherlands earlier or later than May.

The vast majority of users have only stayed for one day in the Netherlands. The number of users who have stayed between 15 and 30 consecutive days is relatively low. However, there are 38373 regular foreign users (almost 2 percent), probably people who live abroad and work or study in the Netherlands, or people who temporarily live in the Netherlands to work or study.

One of the aims for tourism statistics is to determine the total number of (overnight) stays and day trips of tourists in the Netherlands. In order to obtain those, the absolute numbers shown in Figure 1 will have to be multiplied by the active stay lengths. This will put more emphasis on foreign users who stay for longer periods.

Figure 2 shows the number of (overnight) stays during the observed period, grouped by the total consecutive stay length. For instance, the left-most green dot means that there are 23617 foreign users who have stayed between the 1st and the 2nd of May 2014 in the Netherlands (which can either

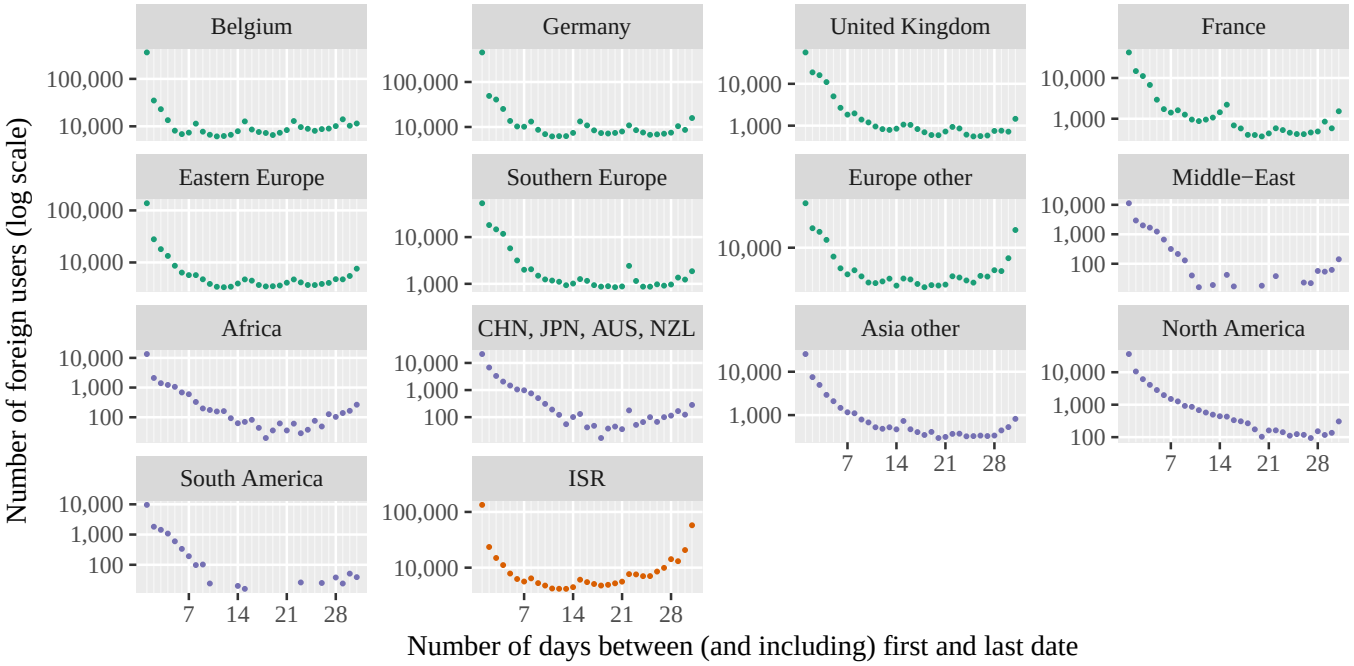

Fig. 4. Consecutive period of foreign users.

be an overnight stay or back and forth traveling) and 3 or 4 consecutive days in total.

Figure 2 also shows a typical tourism pattern. First, there is a holiday period from the 26th April 2014 until the 5th May 2014 of which the first days are visible in this dataset with high frequency of roaming data. These are mainly tourists that stay for 2 days or 3-4 days. Second, there is a weekend pattern (grey area) with again mainly tourists that stay for 2 days and 3-4 days. Many tourists stay from Thursday to Sunday. Third, there are the one-day events. The 29th of May is Ascension Thursday (a National day), and a lot of people also take holiday on the Friday after this Ascension Thursday, again 2 days or 3-4 days. The 5th of May is called Liberation Day (WWII) but is not visible, showing that tourists from other countries are not participating. Also for mothers day (for most countries on the 2nd Sunday of May), no differences are visible.

The number of users per non-consecutive stay length is shown in Figure 3. A stay length of 20 only means that the difference between the first and last date equals 20 days. It does not say anything about how many days in between the user actually stayed in the Netherlands. Observe that 30 percent of the foreign users have stayed non-consecutively for longer than 1 week in the Netherlands. An explanation for this could be the large amount of Belgium and German foreigners who work in the Netherlands or who regularly make day trips to the Netherlands.

Figure 4 shows the number of foreign users per non-consecutive stay length by (grouped) country of origin. The color of the dots correspond to the continent: Europe (green), other (purple) and Inter Standard Roaming Roaming (orange). Small local peaks are visible at 8, 15, and 22 days, which probably corresponds to standard holiday stays of one, two, and three weeks. There are differences visible per country of origin. Germany and Belgium show for example higher rates of visitors that stay short, usually one day , but we will not describe them in detail.

The total number of foreign users per (grouped) country are depicted in Figure 5 shows as expected that the number of German and Belgium users is relatively high. Also, the number of Inter Standard Roaming user is high, especially among regular users.

This dataset contains several limitations. First, in the Netherlands there are four network providers and Vodafone is one of them. In this study, no corrections or extrapolations were

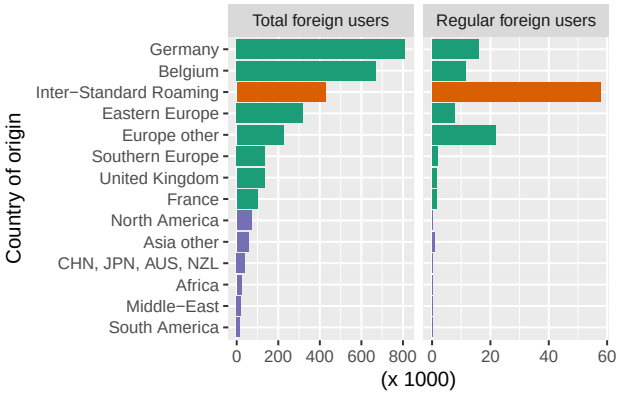

Fig. 5. Consecutive period of foreign users.

made. Second, a time window of 31 days was used. Therefore, the stay length of foreign users that were already in the Netherlands at May 1st or were there at May 31st cannot be determined exactly, since they could have been in the Netherlands earlier or later than May. Third, there was no geolocation provided in this dataset. Therefore, it was not possible to make a clear distinction between tourists on the one hand and border traffic and foreign workers in the Netherlands on the other hand.

IV. Conclusion

Based on the results presented in Figures 1, 2, and 3, it can be concluded that a time window of 15 days is sufficient for covering almost 98% of consecutive stays and 80% of non-consecutive stays. It is advisble to choose the time window larger than the desired stay length to be measured, in order to decrease the probability of overlap.

For tourism statistics, a time window of 1 month is preferable in order to compare the mobile phone based estimations of tourists with the current tourism statistics.

For future research, it is necessary to classify foreign users as tourists, day trip visitors, and cross border workers. For tourism statistics, also the geospatial information is important. Once roaming data is available for time window of 15 days or longer, geospatial information can be used to classify foreign users based on the places they visit.

References

[1] Raun, J. Ahas, R., 2013. Distinguishing tourism destinations with behavioural data. Brussel, New Techniques and Technologies for Statistics, NTTS.

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

[3] Altin, L., Tiru, M., Saluveer, E., Puura, A., 2015. Using Passive Mobile Positioning Data in Tourism and Population Statistics, Brussel, New Techniques and Technologies for Statistics, NTTS.

[4] Meersman, F. de, Seynaeve, G., Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., Wirthmann, A. Demunter, C., Reis, F., Reuter, H. I., 2016. Assessing the Quality of Mobile Phone Data as a Source of Statistics, Paper for the European Conference on Quality in Official Statistics (Q2016).

[5] Offermans, M, Priem, A, Tennekes, M., 2013. Rapportage project Impact ICT mobiele telefonie. Technical report (in Dutch). Statistics Netherlands.