Closer to the total? Long-distance travel of French Mobile Phone users

Maxim Janzen
Maarten Vanhoof
Zbigniew Smoreda
Kay W. Axhausen

Working paper 1169
Institute for Transport Planning and Systems

June 2016
Abstract

Analysis of long-distance travel demand has become more relevant in recent times. The reason is the growing share of traffic induced by journeys related to remote activities, which are not part of daily life. In today’s mobile world, these journeys are responsible for almost 50 percent of the overall traffic. Traditionally, surveys have been used to gather data needed for the analysis of travel demand. Due to the high response burden and memory issues, respondents are known to underreport the number of journeys. The question of the real number of long-distance journeys remains unanswered without additional data sources. This paper is the first to quantify the underreporting of long-distance tour frequencies in travel diaries. We take a sample of mobile phone billing data covering 5 months and compare the observed long distance travel with the results of a national travel survey covering the same period and the same country. The comparison shows that most of the error estimates calculated by researchers so far are too low. Our work suggests that the number of long distance journeys is twice as high as reported in surveys. It is shown that there two different reasons causing for the underreporting. On the one hand, soft-refusals travelled long-distances, but reported no long-distance tours. On the other hand, respondents underestimate their number of long-distance tours. Consequently, there is a need to use alternative data sources in order to gain better estimates of long-distance travel demand.
**Keywords**
long-distance travel demand; mobile phone data; travel surveys; soft-refusal

**Preferred citation style**
1 Introduction

Analysis of long-distance travel behavior has become more important in recent years since the contribution of long-distance journeys to the overall traffic is growing continuously. Therefore, its influence on plans of urban areas, highways, railroads etc. is becoming greater. Long-distance travel is usually defined by trips, which take place outside of a person’s environment. In order to develop tools, which are able to provide reliable predictions, one needs data sources describing the current state of long-distance travel demand. Tourism researchers have a special interest to this topic since predictions of tourism flows are essential for market size estimates.

Data collection methods in the field of travel demand research were investigated in the past (Axhausen et al.; 2002a; Armoogum and Madre; 2002; Bonnel et al.; 2009; Zmud et al.; 2013; Richardson et al.; 1995; Arentze et al.; 2000; Draijer et al.; 2000). The most frequently used data sources are surveys. In case of long-distance travel the number of these is limited (the main sources are national travel surveys). However, all long-distance travel surveys are facing similar problems. On the one hand, due to the high response burden these surveys have a low number of respondents. On the other hand, it is known that number of journeys reported in surveys is too low (Madre et al.; 2007; Armoogum and Madre; 2002). Both facts limit the explanatory power of studies and leave the question of the quality of the results unanswered (Kühnimhof and Last; 2009).

To overcome these limitations alternative data sources are needed. We propose in this paper to use mobile phone billing data in order to obtain better estimates of long-distance travel demand. The advantage is the large number of people that can be tracked and this without having to ask them to spend a lot of effort on a survey. We analyzed 5 months of mobile phone billing data covering one third of the total French population. The data was provided by the Orange™ Labs. After reconstructing long-distance journeys from the data we could quantify the error reported by the French National Travel Survey. More precisely, a lower bound for the average tour rates was calculated indicating that the number of tours is heavily underreported in the survey.

The paper is structured as follows. After a literature review we describe in detail the mobile phone data made available for our studies as well as the National Travel Survey used in this paper. In section four the tour reconstruction methodology is described. Afterwards, we present the outcomes and comparisons. We conclude this paper with a discussion and a conclusion.
2 Previous Work

Data collection has always been an important issue in the field of travel demand research. Different methods of data collection were investigated in the past (Axhausen et al., 2002a; Armoogum and Madre, 2002). The data sources mostly used are surveys, which have various forms (Dillman, 2000) to suit the diverse requirements of the researchers.

In case of long-distance travel the number of recent surveys is limited. Focusing on Europe, the *Mobidrive* studies are available (Zimmermann et al., 2001; Axhausen et al., 2002b; Chalasani and Axhausen, 2004). These studies focus on a six-week period, which is usually not sufficient for a deep analysis of long-distance travel behavior. Other sources are national travel surveys like the French (Armoogum et al., 2008), British (Department for Transport, 2016) or the Austrian (BMVI, 2012). An additional longitudinal perspective is provided by the INVERMO study from Germany (Chlond et al., 2006). Several European studies have been combined for an analysis of long-distance travel demand in Europe (Frick and Grimm, 2014). A similar approach led to a nationwide model for the United States (Outwater et al., 2015a,b; Bradley et al., 2015).

An overview of available studies of annual long-distance travel rates can be found in table 1. The studies included are: California Statewide Household Travel Survey (CSHTS) (Bierce and Kurth, 2014; Cambridge Systematics Inc., 2013), an ifmo study (Frick and Grimm, 2014; Kühnönhof et al., 2014), the INVERMO project (Zumkeller et al., 2005; Chlond et al., 2006), the Knowledge Base for Intermodal Passenger Travel in Europe (KITE) (Frei et al., 2010), the DATELINE study (Neumann, 2003), the French national travel survey (ENTD) (Armoogum et al., 2008), a Eurostat report (Weckström-Eino, 1999), Methods for European Surveys of Travel Behaviour (MEST) (Axhausen and Youssefzadeh, 1999), US National Transportation Statistics (US NTS) (Bureau of Transportation Statistics, 2016). All studies surveyed 8-12 weeks of long-distance travel and estimated annual tour rates. The ifmo study reports 15.9 one-way long-distance trips resulting in a tour-rate value that is smaller than 8.0. A correction factor is incorporated in all tour rates.

More long-distance travel studies were performed with a special interest in tourism. Guidelines for tourism studies (Harris et al., 1994) and preferred analysis methods (Crouch, 1994) were presented in the past. Many tourism studies were performed, e.g. the Travel Market Switzerland study (Bieger and Lässer, 2008) or the Net Traveler Survey (Schöntal and Williams, 1996). Almost all of them focus on tourism activities within a single country. A summary of international studies can be found in Lennon (2003) or the Eurostat database (Eurostat, 2016). Nevertheless, outcomes of tourism surveys are limited due to the known issue of unobserved tourism (De Cantis et al., 2015).
Table 1: Annual LD tour frequencies: Other studies

<table>
<thead>
<tr>
<th>Year</th>
<th>Area</th>
<th>Destination</th>
<th>Long-dist. Definition</th>
<th>Exclude single-day</th>
<th>Annual tours per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATELINE</td>
<td>2001-02</td>
<td>Europe</td>
<td>international</td>
<td>75km</td>
<td>No</td>
</tr>
<tr>
<td>DATELINE</td>
<td>2001-02</td>
<td>France</td>
<td>international</td>
<td>75km</td>
<td>No</td>
</tr>
<tr>
<td>ENTD</td>
<td>2007-08</td>
<td>France</td>
<td>France</td>
<td>80km</td>
<td>No</td>
</tr>
<tr>
<td>MEST</td>
<td>1997-98</td>
<td>France</td>
<td>international</td>
<td>100km</td>
<td>No</td>
</tr>
<tr>
<td>MEST</td>
<td>1997-98</td>
<td>Europe</td>
<td>domestic</td>
<td>100km</td>
<td>No</td>
</tr>
<tr>
<td>ifmo</td>
<td>2011</td>
<td>Germany</td>
<td>international</td>
<td>100km</td>
<td>No</td>
</tr>
<tr>
<td>KITE</td>
<td>2008-09</td>
<td>Switzerland</td>
<td>international</td>
<td>100km</td>
<td>Yes</td>
</tr>
<tr>
<td>KITE</td>
<td>2008-09</td>
<td>Portugal</td>
<td>international</td>
<td>100km</td>
<td>Yes</td>
</tr>
<tr>
<td>CSHTS</td>
<td>2012</td>
<td>California</td>
<td>state-wide</td>
<td>50 miles</td>
<td>No</td>
</tr>
<tr>
<td>Eurostat</td>
<td>1999</td>
<td>France</td>
<td>international</td>
<td>100km</td>
<td>No</td>
</tr>
<tr>
<td>INVERMO</td>
<td>2001-03</td>
<td>Germany</td>
<td>international</td>
<td>100km</td>
<td>No</td>
</tr>
<tr>
<td>MEST</td>
<td>1997-98</td>
<td>Europe</td>
<td>international</td>
<td>100km</td>
<td>No</td>
</tr>
<tr>
<td>KITE</td>
<td>2008-09</td>
<td>Czech Rep.</td>
<td>international</td>
<td>100km</td>
<td>Yes</td>
</tr>
<tr>
<td>US NTS</td>
<td>2001</td>
<td>USA</td>
<td>international</td>
<td>50 miles</td>
<td>No</td>
</tr>
<tr>
<td>MCS</td>
<td>2010</td>
<td>Switzerland</td>
<td>international</td>
<td>100km</td>
<td>No</td>
</tr>
</tbody>
</table>

Due to the high response burden that is usually associated with long-distance surveys (Axhausen et al., 2015; Axhausen and Weis, 2010), it can be expected that the number of long-distance trips is usually underreported. The reasons are not responding frequent travellers as well as travellers claiming not to travel, while answering other questions, so called soft-refusers (Madre et al., 2007). Furthermore, there is a memory effect. Respondents tend to forget tours, which happened some time before the survey (Smith and Wood, 1977; Bradburn et al., 1987; Tourangeau, 1999). Additionally, the vehicle miles travelled are usually heavily underestimated as shown by Wolf et al. (2003). Consequently, there is a need of survey weighting and expanding (Bar-Gera et al., 2009). Assumptions about underreporting long-distance tour rates in surveys led to an introduction of correction factors in several studies (Cambridge Systematics Inc., 2013; Armoogum et al., 2008). In case of tourist surveys, a weight correcting for the response bias is essential (Lee worthy et al., 2001). Nevertheless, a correction factor is the only method available now to account for the underreporting. Assumptions about inaccuracy of long-distance travel surveys are supported by the fact that two surveys with the same scope can suggest non-consistent travel behavior (Perdue and Botkin, 1988).

In order to estimate the level of underreporting in the surveys, one needs alternative data sources.
Nowadays, there are mainly two alternative sources available for the analysis of travel demand. Both use passive data collection. On the one hand, GPS data can be used to collect information about travel behavior (Montini et al., 2014). But the collection of GPS data is limited since the cooperation of the respondent is needed and smartphone GPS collection is battery consuming discouraging participation. On the other hand, mobile phone network operators produce mobile phone billing data that provides an enormous amount of data and has been already utilized in various fields (Blondel et al., 2015) including the field of transportation. One of the first applications was the analysis of travel demand induced by tourism (Ahas et al., 2008a, 2007). GSM data has been also used to estimate OD matrices (Friedrich et al., 2010; Pan et al., 2006; Cik et al., 2014). Altogether, GSM data is a very powerful tool for predictions of human mobility (Song et al., 2010). We will show in this paper that it is as well very useful for the analysis of long-distance travel demand.

Several studies have been carried out on data quality comparisons. These studies compared, for instance, GSM data with GPS trajectories (Iovan et al., 2013; Hoteit et al., 2014; Smoreda et al., 2013) as well as on the sociological aspects of using mobile phones like, for instance, in analysis of places relevant in transport science (Licoppe et al., 2008), concluding that CDR data forms a good proxy for overall tendencies of human mobility thanks amongst others to the large samples of persons and days involved.

Mobile phone data and National Travel Surveys have only sporadically been compared so far. Bekhor et al. (2013) did an evaluation in Israel. But the study area was comparably small in person-days involved, the focus was not longitudinal travel behavior, and the first data source is preceding the second one by 10 years (including a 25% population increase). Similar work has been done in USA (Huntsinger and Donnelly, 2014), but it was as well limited to a regional level (North Carolina). Neither of the two studies provides statements about the long-distance travel demand since they focus on other topics in travel behavior. We will close the gap in this paper.

3 Data Sources

3.1 Mobile Phone Billing Data

The study described in this paper is based on an anonymised mobile phone billing data set recorded by Orange™France. It consists of Call Detail Records (CDRs) covering mobile phone
usage of around 23 million users of the Orange™ network in France during a period of 154 consecutive days (May 13, 2007 to October 14, 2007). Given a population estimate of 63.945 million inhabitants in 2007 \(^1\), that is roughly 35.9% of the French population. The numbers concord with estimates made at Orange™ about mobile phone penetration in France anno 2007: 86% (ARE, 2016) and with the supposed market share of Orange™ in that year: 43.5%.

Each CDR contains information about the action (outgoing/terminating call or SMS) which took place in the network. The information needed for our purpose is the caller id, the time and duration of the action, and the antenna, which was the connection point of the mobile phone at the start of the action. In turn, the location of each antenna is known. Given location and time information for each action, a user can be traced and thus his or her movements can also be extracted. The accuracy of the movement reconstruction is dependent on the frequency of actions.

The data set has several limitations. Firstly, the action frequency is comparably low, because mobile data usage was not as intense in 2007 as it is today. Secondly, the data set does not cover a full year. Thus, any estimates for the missing time periods have to be supported with complementary data sets. In addition to the temporal inaccuracy due to the low call frequency, there is also spatial inaccuracy. In case of CDR data, the spatial information is limited to the position of the mobile network antennas. For less densely populated areas of the country, the antenna can be several kilometers away from the actual position of the mobile phone. Finally, no information about phone calls made abroad is available in this data set. Even though it is known that France has one of the highest ratios of domestic trips to trips abroad within Europe (OECD, 2012; Eurostat, 2016) this circumstance limits the range for which we can make valid estimates. We will account for this limitation with respect to the special situation of a large central European nation in the validation section below.

It has been shown that mobile phone billing data should be used with caution when analyzing mobility (Ranjan et al., 2012). Nevertheless, most limitations do not have a substantial impact when focusing on long-distance travel demand. The spatial and temporal inaccuracies described above become relatively small since we are working on large spatial and temporal scales. Still, mobile phone data can provide a lower bound to the real value. When comparing with survey data, we have to account for the missing roaming data and focus on national travel, though.

\(^1\)This is the average of the monthly estimates for the period between Mai and October 2007 as obtained from the French National Statistics Website (www.insee.fr)


3.2 Survey Data

The results of the CDR data analysis will be compared to a national travel survey. We used the *Enquête Nationale Transports et Déplacements* (ENTD), the French National Travel Survey. The ENTD is conducted every 10-15 years (1967, 1974, 1982, 1994, 2007-08). Various actors are involved in the ENTD, including the Ministry of Transport, the INSEE (French National Institute of Statistics and Economic Studies) and IFSTTAR (French institute of science and technology of transport, planning and infrastructures). The last ENTD was performed from April 2007 to April 2008 and most parts are publicly available (IFSTTAR, 2016). Since the survey includes the time period covered by the CDR data described above we use the ENTD 2008. One of the goals of this survey was the analysis of long-distance mobility. This fact is advantageous, because it ensures that we can compare the two data sources in terms of long-distance travel behaviour.

Nevertheless, the sample size of the ENTD 2008 is much smaller than the available CDR data. In total, 20'178 households and 44’958 individuals were surveyed. Just 18’632 (representative) persons were chosen for the long-distance travel module of the survey (Armoogum et al., 2008). The latter were asked to report their long-distance travel practices within the preceding 4 weeks as well as within the preceding 13 weeks. Detailed information on the tours of the 4-week period are publicly accessible, while for the 13-week period just the absolute number of tours is available. In the ENTD 2008 a long-distance journey is either a journey with the furthest destination being more than 80km away from home (crow-fly distance) or a journey, which includes at least one overnight stay (or both). We will account for the differences of the data sources in the comparisons (section 5). Unless not stated otherwise, we use the 4-week records to compare the ENTD to the CDR data since the 13-week records are not accessible to the public.

4 Methodology

The mobile phone billing data set described above was far too big to be analyzed completely in the framework of this study. Thus, we had to reduce the number of mobile phone users considered and their CDRs. We performed two selection steps. Firstly, a set of cities was chosen. Secondly, from each city a small subset of customers was selected in order to investigate their travel behaviour. We wanted to see, how large the impact of the population size on the long-distance travel of their residents is, as the German literature suggests that inhabitants trade-off their daily travel against more long-distance travel (Holz-Rau et al., 2014; Schlich, 2001). Both selection
processes are described in detail in the following. Subsequently, the algorithm used to extract the long-distance tours from the mobile phone data is presented.

In order to prevent misunderstanding take a note of the definitions, which are used in the remainder of the paper.

**Definitions:**
- **User Environment:** The area within a radius of 80km around the home location.
- **(Home Based) Tour:** A chain of activities and trips starting and ending at the home location (sometimes referred to as journey).
- **LD Tour:** A tour, which leaves the user environment.
- **LDF Tour:** A LD tour within France.

### 4.1 Municipality Selection

As described above, we want to limit the number of tracked persons in the CDR data. In the first step, we chose some cities and focus our analysis on the inhabitants of these cities. The cities were selected considering three attributes:

- Size of population (Inhabitants within the municipal borders).
- Economic strength (GDP per capita).
- Share of tourism sector (number of employees in touristic sector per 1000 inhabitants)

These municipality attributes were taken, because it is expected that they influence the long-distance travel behaviour of the inhabitants. All values are taken from the French Institute for Statistics - the *Institut national de la statistique et des études économiques* (INSEE, 2016). The values of all three attributes were assigned to three different categories (high, medium, low). We chose 32 cities covering all combinations of levels of the three attributes. Additionally, Paris was chosen, which is an outlier for all three attributes. Finally, some smaller communes were randomly chosen in order to cover also villages and rural areas.

In addition, two geographic attributes were considered during the municipality selection process. Firstly, the selected cities are uniformly distributed in France. Secondly, the next country border is not close. The latter was considered in order avoid possible frequent international travel in the data set, which can not be recognized in the underlying CDR data. The final selection of the mobile phone towers is shown in Figure [1]. In total, 23′438 towers in 3′631 distinct locations serve the chosen municipalities. These towers were used to identify the inhabitants of the municipalities. Bold alike crosses indicate that there are many towers close to each other.
This is the case in dense cities. The cities located closest to a border are Calais (sea side), Lille, Strasbourg and Mulhouse. Furthermore, all regional centers (identified by high population densities) are included in our selection.

### 4.2 Identification of Residents of the Selected Municipalities

In order to decide whether a customer is an inhabitant of one of the municipalities considered one needs to impute the place of residence of this customer. An analysis of home anchor points
(Ahas et al., 2008b; 2010) is undertaken for this purpose. Anchor points are the mobile network towers, which are most frequently used by a customer during a specific time of a day. Computing home anchors, we focus on the night (9 p.m. to 6 a.m.), because most people are expected to be at home for the majority of nights. An additional requirement is needed in order to avoid that a home anchor is wrongly set by call actions of a single night. Thus, we demand that a tower is a home anchor candidate only if the phone was in use at this location for at least 7 distinct days of a month. Following these rules, home anchor points were computed for each customer and each of the covered six months.

Around 18 million of the users had at least one month, where it was possible to identify a home anchor point. A customer is considered to be a resident of a municipality if he or she has at least three of his/her - maximal six - home anchor points within this municipality. This threshold was chosen, because two of the six monitored months had just 15 days of observations. Thus, there was a substantial share of customers who did not have anchor points in these months. Therefore, most of the persons had just four home anchors. Hence, we assume that people live at a place if they have three quarters of their home anchors at the same place. We chose all customers that are inhabitants of any of the selected municipalities. This subset contains more than 1.4 million customers, and therefore, captures over 17% of the population of the selected municipalities.

Afterwards, an algorithm was applied to identify machine-to-machine devices. These machines are sim card devices that are not used by humans but are automated. One can find these machines by looking for specific periodic behavior. All identified machines were removed from the previous subset. As a consequence the size of the customers subset was reduced to 1.39 million.

<table>
<thead>
<tr>
<th>Population [in 1000]</th>
<th>Tracked Persons</th>
<th>Number of Communes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paris</td>
<td>4953</td>
<td>1</td>
</tr>
<tr>
<td>200-900</td>
<td>19394</td>
<td>10</td>
</tr>
<tr>
<td>100-200</td>
<td>25294</td>
<td>13</td>
</tr>
<tr>
<td>50-100</td>
<td>9580</td>
<td>5</td>
</tr>
<tr>
<td>20-50</td>
<td>7461</td>
<td>4</td>
</tr>
<tr>
<td>10-20</td>
<td>7730</td>
<td>5</td>
</tr>
<tr>
<td>5-10</td>
<td>3190</td>
<td>5</td>
</tr>
<tr>
<td>1-5</td>
<td>1376</td>
<td>7</td>
</tr>
<tr>
<td>rural (&lt;1)</td>
<td>896</td>
<td>8</td>
</tr>
</tbody>
</table>
In order to identify the persons that actually did at least one LDF tour, an additional filter had to be implemented. We chose a single month (June 2007) and investigated whether the persons left their user environment during this month. More than 814,000 of our identified residents did so. Among these we randomly selected a subset of persons for a detailed analysis of their long-distance travel behavior. 5,000 residents of Paris as well as 2,000 persons from each of the other cities and all identified persons of the smaller communes were chosen. Initially the user environment radius was 50km, but had to be extended to 80km, because the ENTD uses a radius of 80km. After the extension 79,874 persons were left and studied regarding their long-distance travel behavior. Table 2 shows the number persons and communes by population size.

4.3 Reconstruction of Long-Distance Tours from CDR Data

Unlike surveys, mobile phone data does not provide information about tours undertaken directly. The information available is a series of time-space points. We have shown how the series can be used to impute home locations of mobile phone users. In the following, the extraction of long-distance tours is described in detail.

Scanning the CDRs of the users we suppose that a LD tour starts every time a CDR with a location outside the user environment occurs following a CDR located within the user environment. The tour is assumed to end with the first CDR back in the user environment. A sketch of a single construction process can be found in figure 2(a). The initial situation consists of the home anchor (H) and the user environment (green circle). Now, the locations of CDRs are identified by C1, C2,...C6, where their sequence is given by their numbers. The black dashed arrows show an potential path of the user, while the red arrows form the reconstructed tour.

In the sketch of figure 2(a) the constructed tour fits the initial real world tour quite well. This is not always the case. A problem is the boundary of the traced time period. The tours that are not finished before the end of the observed time period have to be truncated without any information of the further duration (figure 2(b)). Likewise, the tours started before the recording time have to be truncated (figure 2(c)).

Moreover, the character of CDR data causes further limitations. Firstly, there is no information about the mobile phone usage outside of France. This lack of information induces wrong final destinations during the tour construction (figure 3(a)). Without any mobile phone activity between the user environment and the border even an around-the-world tour would be missed. Likely, this is the case for most of the international tours. Secondly, low-frequency mobile phone users might go on two distinct tours without any mobile phone activity within the user environment between the two tours. In this case the tour construction algorithm merges the two
Figure 2: Visualization of the tour reconstruction algorithm

(a) Perfect tour reconstruction
(b) tour with unobserved end: C4 is after 16th of Oct.
(c) tour with unobserved start: C1 is before 15th May

Legend
H - Home anchor, C1...C6 - CDR positions, ○ - User environment, → - Reconstructed tour, ⟷ - Real world tour

Figure 3: Failed tour reconstructions

(a) Missed tour abroad
(b) Two tours merged
(c) Tour completely missed

Legend
H - Home anchor, C1...C6 - CDR positions, ○ - User environment, → - Reconstructed tour, ⟷ - Real world tour

tours due to a lack of a separation CDR (figure 3(b)). Thirdly, the worst case is a user without any CDR’s that relate to his long-distance travelling. Without the CDRs indicating an exit of the user environment no tour can be reconstructed (figure 3(c)). This is the most critical and probably the most frequent case of a failed tour reconstruction. In addition, it is also possible to miss just some parts of the tour or the final destination. Note that all limitations lead to a lower number of tours in comparison to the real world. Therefore, we can assume that the number of LDF tours identified by the algorithm is a lower bound of the total.
4.4 Identification of Commuting in CDR Data

The long-distance tours reconstructed from CDR data do not contain any information about the travel purpose. Therefore, the purposes have to be derived from the available information, like tour duration. Commuting is exceptional among all travel purposes, as a few travellers account for a substantial part of the overall long-distance tours. Additionally, commuters have specific travel patterns. These patterns can be used to impute commuting in the long-distance tours based on the CDR data set.

First of all, the home and work locations of all persons have to be determined. We use the home anchors as described in section 4.2 to determine the home locations. The work locations are determined with a similar approach. The considered time of the day relevant for a work anchor is 9 a.m. to 4 p.m., which are the regular working hours. For further analysis, all persons with work anchors outside of their user environment were selected. In the next step, regular weekday tours between home and work anchors were identified. More specifically, a person is considered to be a commuter, if he or she did a substantial amount of home-work tours from Monday to Friday. Since we focused on persons with work places outside of the user environment, these commuters are long-distance commuters. The approach described above has several limitations. Shift workers are difficult to identify, and so are weekly commuters with two residences. Additionally, commuting tours are shorter than an average long-distance tour. This fact reduces the probability to spot all commuting activities in a CDR data set. Nevertheless, a lower bound for the commuting tour rate can be determined.

4.5 Seasonal Tour Frequencies in Survey Data

The CDR data analysis have to be compared to the National Travel Survey. Therefore, the survey data has to be adjusted in order to make the two data sets comparable (e.g. international journeys have to be excluded). A major difference between the available CDR and survey data is the time period covered. While the survey covers the whole year, the CDR data is limited to five months (mid of May to mid of October). Consequently, the share of tours within these five months has to be computed in the survey.

A detailed analysis of the tour frequency distribution was performed. For each day of the year the number of tours as well as the number of persons reporting for this day were computed. Subsequently, the number of tours was summed up and scaled by the number of respondents. The computation shows that around 46.2% of all LDF tours took place within the five summer
months. The share is higher than 5/12 confirming the assumption that people tend to travel more during the summer. We will account for the higher share of journeys in summer in the next section.

5 Results

We obtain the information from the mobile phone data regarding long-distance travel behaviour, (e.g. tour distance, tour frequency) and compare it to the ENTD 2008. The main differences are pointed in the following. The main result is that the yearly long-distance travel demand is heavily underestimated, if one relies on the numbers of the ENTD 2008. All results presented in this section are limited to journeys within France with a destination more than 80km away from the home location.

5.1 Tour Distance Distribution

We investigate the distribution of LDF tour distances. The distance of a tour is defined as the crow-fly distance between home location and the known point of the tour, that is furthest away from home. The results are shown in figure 4. We compare the cumulative frequency of the tour
distances for both data sources. Additionally, both data sets are subdivided by municipality size of the travellers home location. The considered classes are ‘less than 20 000’, ‘20 000 to 900 000’ and ‘Paris’.

One can see that the CDR data reflects the survey almost perfectly for the residents of the bigger cities. In the range of 450km-800km the ENTD reports a higher share of LDF tours than the CDR data. An explanation is the underestimation of the travelled kilometers for very long tours, i.e. tours that are around 1000km (see Wolf et al. (2003)). In case of the residents of smaller communes the CDR data reports a higher share of tours in the range of 80km-200km. An explanation might be that these persons travel usually longer distances and thus underestimate the distances of mid-distance tours. This assumption is supported by the analysis of the tours in the range of 50km-80km. The CDR data shows that residents of smaller communes do 5.9 of these tours per mobile person, while the other persons do just 3.9 tours in that range.

### 5.2 Commuting

A substantial part of the long-distance journeys is covered by commuters. Therefore, it is important to identify long-distance commuting patterns. While commuting tours are marked in the national travel survey, the CDR data does not provide this information. Consequently, a commuting detection algorithm was implemented (see Section 4.4). The share of commuters as well as the percentage of commuting tours can be found in table 3.

<table>
<thead>
<tr>
<th></th>
<th>Share of commuters</th>
<th>Share of commuting LDF tours</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENTD</td>
<td>0.17%</td>
<td>4.8%</td>
</tr>
<tr>
<td>ENTD, corrected</td>
<td>0.29%</td>
<td>6.6%</td>
</tr>
<tr>
<td>CDR data</td>
<td>0.38%</td>
<td>1.9%</td>
</tr>
</tbody>
</table>

The shares of commuters in the two data sets differ substantially. The algorithm described in section 4.4 identified 437 long-distance commuters in the CDR data indicating that 0.38% of the total population are regular long-distance commuters. On the other hand, just 0.17% of the ENTD respondents were commuting for more than 80km. Grimal, who did a detailed analysis of the ENTD 2008 and used a correction factor, reports that 0.29% of the people are long-distance commuters (Grimal, 2010). In case of the share of commuting tours the differences are remarkable, but here the value is higher in the ENTD. The inconsistency of the two data sets is highlighted here and clarified later in the discussion.
5.3 Share of Long-Distance Travellers

The number of long-distance travellers is a major question in transport demand modelling and thus also in tourism demand analysis. Survey respondents are known to underreport their long-distance tours due to the high response burden of the corresponding items. Due to the enormous amount of data, we restricted this analysis to a single month (June 2007). The result as well as the corresponding values in the ENTD 2008 are shown in table 4.

Table 4: Number of persons performing LDF tours

<table>
<thead>
<tr>
<th>Reporting/Tracked Interval</th>
<th>CDR Data</th>
<th>ENTD 2008</th>
<th>ENTD 2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surveyed Persons</td>
<td>1'388'941</td>
<td>18'632</td>
<td>18'632</td>
</tr>
<tr>
<td>LDF Mobile Persons</td>
<td>814'381</td>
<td>4'796</td>
<td>25.7%</td>
</tr>
<tr>
<td>Selected for further analysis</td>
<td>79'874</td>
<td>4'796</td>
<td>8'743</td>
</tr>
</tbody>
</table>

One can see that the share of travellers within one month of CDR data is more than twice the share of travellers reported in four weeks of ENTD 2008. Considering the 13-week reports of the survey increases the share of travellers. Nevertheless, the value is lower than the 59%, which can be observed in the CDR data, and lower than the 61% estimated by Weckström-Eno (1999). The results support the assumption that a substantial number of survey respondents does not report their long-distance journeys. Consequently, long-distance survey practice should not only pay attention to the response rates, but they should also find a way to convince the respondents to report their journeys.

5.4 Tour Rates for Mobile Persons

We have shown that the number of persons reporting LDF tours is much lower in the ENTD. The question is whether the tour rates differ for those who report tours. Firstly, we compare the number of tours within three months (In case of the ENTD, 13 weeks is the reported interval). Figure 5 shows the histograms for the two data sources. One can see that most of the ENTD respondents did just 1 tour in this period and just a very small share of persons travelled more than three times. The CDR histogram suggests that many persons do 2,3 or 4 LDF tours and also a substantial amount of tracked persons travels more than 5 times within three months.

We compare the tour rates also month by month, in order to identify seasonal effects that might
have an influence. The monthly LDF tour rates for mobile persons are shown in figure 6. The tour rates are substantially higher in the CDR data. Consequently, the assumption of underreported tour frequencies in surveys is confirmed. There are two notes that have to be mentioned. On the one hand, the reference intervals differ slightly. While the CDR data is cut into monthly chunks, the ENTD responses refer to a four week period. On the other hand, May and October are not fully covered in the CDR data set. Thus, the shown tour rates are likely to be lower than the actual ones.
Finally, the total long-distance travel demand was analyzed. The number of LDF tours per capita was calculated based on the CDR data as well as based on the ENTD and is shown in table 5. The frequencies reported show the tour rates in the period from 16 May to 15 October. We refer to this period as *summer period* in the following. In case of the ENTD three different data sources were used. Firstly, the number of reported tours within 4 weeks was taken. Secondly, the number of reported tours within 13 weeks was used. Thirdly, the number of projected yearly tours were taken into account. For the latter we used the information that 46.2% of the yearly LDF tours were undertaken in the summer period. Two different LDF tour rates are reported, namely including commuting tours and excluding commuting. The detailed results for the 13-week survey are not publicly available. Thus, the share of commuting tours is not known, and therefore, the tour rate excluding commuting tours is not reported here.

Table 5: Average number of LDF tours per capita from 16 May to 15 October

<table>
<thead>
<tr>
<th>Reference Interval</th>
<th>CDR data 5 months</th>
<th>ENTD 4 weeks</th>
<th>ENTD 13 weeks</th>
<th>ENTD weighted 1 year</th>
</tr>
</thead>
<tbody>
<tr>
<td>incl. commuting</td>
<td>4.27</td>
<td>2.25 (52.7%)</td>
<td>1.96 (45.9%)</td>
<td>2.36 (55.3%)</td>
</tr>
<tr>
<td>excl. commuting</td>
<td>4.14</td>
<td>2.14 (51.7%)</td>
<td>–</td>
<td>2.20 (53.1%)</td>
</tr>
</tbody>
</table>
One can see that the frequencies eventually suggested by the ENTD are approximately half as high as observed in the CDR data. Additionally, exclusion of commuting or adding the weighting factor proposed by the ENTD analysts does not change the main finding here. The factor of underestimation is much higher than usually assumed (e.g. up to 1.3 in (Cambridge Systematics Inc., 2013)). The seasonal effect has been taken into account, while there is also a spatial effect. Therefore, we analyze the LDF tour rates by the size of the home city in order to capture this effect (Figure 7). Additionally, the 95%-confidence intervals are presented in Figure 8. Due to the smaller sample size the, the confidence intervals of the survey results are wider than the intervals based on the CDR analysis. Once again, one can see that the CDR data suggests a long-distance rate, which is twice as high as the survey outcome.

Figure 7: Average LDF tour rates per capita by municipality size for the summer period

The differences described in this section have enormous impact. This can be shown by pointing to the resulting absolute numbers. The analysis of the ENTD 2008 led to the assumption that the French population undertook 325 million long-distance tours per year. Limiting it to national tours and the summer period lead to an estimate of 130 million tours. On the other hand, the CDR data suggests that there are almost 240 million tours for this population and time frame. It can be assumed that an extension to the whole population, the full year and international tours would not change the survey underreporting rate. Consequently, the ENTD 2008 underestimates the annual long-distance travel demand by more than 260 million tours and all their economic and environmental impacts. This number is based on simple scaling assuming that the share of tours in the summer period and the share of national tours reported by the ENTD 2008 apply. Nevertheless, it gives an idea of the magnitude of the error. Additionally, we have to stress the
Closer to the total? Long-distance travel of French Mobile Phone users  

June 2016

Figure 8: Confidence Intervals of the per capita LDF tours per capita

---

fact that the tour frequency suggested by the CDR data is just a lower bound to the real value, which might be even much higher.

6 Discussion

The methods used and the results obtained are discussed in this section. Especially, limitations and their implications are highlighted.

6.1 Commuting

Firstly, the focus of the discussion is on the analysis of long-distance commuting behavior. It is shown in subsection 5.2 that the results of the ENTD 2008 and CDR data differ enormously. On the one hand, the share of commuting tours is much lower in the CDR data (1.9% vs. 4.8%). On
the other hand, the CDR data reports twice as many long-distance commuters (0.38% vs. 0.17%). Likely, neither of the two data sets tells the full truth in this case. One can see that the ENTD analysts anticipate underreporting in case of long-distance commuters, because the weights correct this number to a higher value. It is likely as well that many commuting tours could not be identified in the CDR data, because these tours are usually relatively short and people might not have placed a call during working hours. Therefore, the share of commuting tours is expected to be underreported in the mobile phone billing data analysis. Consequently, the share of commuting tours in the ENTD gives an upper, but probably more realistic upper bound, while the data produces a more reliable estimate for the number of long-distance commuters. An upper bound because the question invited the respondent to generalize their behaviour, ignoring common exceptions such as illness, business trips, work at home etc.. Finally, it has to be mentioned that statistical noise is very likely here, because the absolute numbers in the ENTD are low compared to the CDR data (there are only 31 long-distance commuters).

6.2 Limitations

As mentioned before, the CDR data used in this work has several limitations. Some of them are obvious and can be taken into account like the restriction to national tours and the 5-month observation period. The results presented in section 5 account for these restrictions. The question whether the person selection of the CDR analysis is biased can not be answered. The tracked persons are randomly selected among cell phone users and compared to the ENTD, which represents the population aged 6 and older. Therefore, bias is not expected to have an influence here.

Additionally, low CDR frequency in the data set is a concern. It results in long-distance tours, which can not be identified in the data. Especially short tours (in terms of duration) are affected as it was already highlighted in section 6.1. Nevertheless, we pointed out that this issue does not lead to overestimation. Rather, the resulting number of long-distance tours is in fact a lower bound to the actual number, which is expected to be substantially higher. In addition, we assumed in our analysis that LDF tours are undertaken just by those persons, who had at least one LDF tour in June 2007. Thus, adding the tours of the other persons will lead to an even higher number of the total amount of long-distance tours.
7 Conclusion

We have analyzed the long-distance travel behaviour of the French population. The data source used was CDR data covering five months of mobile phone usage within the French Orange™ network. We found that the number of long-distance tours reported by the National Travel Survey is underestimated. The actual long-distance tour frequency is almost twice as high. Considering the fact that CDR data is just a lower bound and is probably underreporting short tours substantially (e.g. single-day commuting), the long-distance tour frequency is likely even higher than shown in section 5. It has been shown that there are two reasons for the underestimated tour frequencies. On the one hand, the average tour rates of mobile persons differ by in the ENTD and the CDR data. Hence, survey respondents are underestimating their number of long-distance tours. On the other hand, the number of persons reporting any long-distance tour is much higher than suggested by the ENTD data. Therefore, soft-refusals are responsible for a substantial part of the underestimated long-distance tour numbers. Consequently, it is obvious that alternative data sources are indispensable for a reliable estimate of the long-distance tour frequency. Possible sources are either mobile phone data as presented in this paper or extended GPS studies. Either way, inclusion of the device carried by us almost all the time - the mobile phone - is deemed necessary for a better understanding of long-distance travel behavior. Finally, the underestimation of long-distance travel has consequences to transport policy (e.g. due to wrong estimates of CO2 emissions) and especially to the tourism sector (e.g. some markets are greater than assumed so far).

8 Acknowledgement

We would like to acknowledge the Swiss National Science Foundation (SNF) for providing funds to the authors under grant 205120_165900.

We are also grateful to Jimmy Armoguum from IFSTTAR who provided expertise in the ENTD 2008.

We are also thankful to Aare Puussaar for assisting us with his database knowledge and programming skills.
References


Closer to the total? Long-distance travel of French Mobile Phone users

June 2016


