Detected or undetected, which trips are seen in mobile phone OD data? A case study of the Lyon region (France)

E. Casassa[a*], E. Côme[a], L. Oukhellou[a]

[a] Université Gustave Eiffel, COSYS-GRETTIA, Marne La Vallée, France
eliane.casassa@univ-eiffel.fr, etienne.come@univ-eiffel.fr, latifa.oukhellou@univ-eiffel.fr
* Corresponding author

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)
September 02-03, 2024, Crete, Greece

April 26, 2024

Keywords: mobility demand, mobile phone data, household travel surveys, discrepancies

1 INTRODUCTION

Traditionally, mobility studies use household travel surveys (HTS) to characterize mobility demand in order to adapt infrastructure and services to the needs of citizens. HTSs provide a high level of detail (origin-destination, mode choice, trip purpose) and additional information about the individual (e.g. age, gender, socio-professional category). In addition, surveys try to be representative of the whole population. However, as they are very expensive, they are only carried out every 5 to 10 years, with small samples and on average only one day a week.

The last few decades have seen an explosion in the use of mobile phone networks, smart cards, and GPS devices. They produce huge amounts of data and cover large geographical areas with fine temporal granularity. However, these data are very often less complete than mobility surveys. Indeed, it is much more difficult to access the metadata of the trips made by users. Moreover, these data sources raise questions about representativeness and bias.

In the literature, several studies have been conducted to enrich mobile phone data and then validate it against HTS ground truth. Some have already attempted to infer people's places of residence and work from mobile phone data (Chrétien (2016)), while others have detected the transport mode (Huang *et al.* (2019)). Some authors have also processed raw cell phone trajectories to estimate origin-destination (OD) matrices and then compared them with origin-destination matrices from HTS (Fekih *et al.* (2021), Bonnetain *et al.* (2021)). Other recent work has focused on identifying biases between different data sources (Sfeir *et al.* (2024), for example, analyzing HTS biases using smart card data).

Following the same objective, our work focuses on highlighting the discrepancies between the two data sources namely, HTS and mobile phone OD matrices. Applying the definition of a *trip* according to a French telecom operator to HTS data, we want to answer the main question:

- What are the features of trips detected by the mobile phone data, in terms of transport mode, duration, distance, and purpose?

We also investigate two additional issues:

- Do the data provided by the telecom operator lead to volumes comparable to those of the surveys using the same definition of a trip?
- Are the OD matrices for home-work purpose provided by mobile phone data close to those estimated by the public statistics institute?

2 CASE STUDY: DATA SOURCES AND STUDY AREA

For this study, we used two different types of data:

- A HTS realized in 2015 by CEREMA (expert study centre on risks, environment, mobility, and urban planning) in Lyon Metropolis in France. 28,230 individuals, living in the study area, from 16,361 different households were surveyed on their trips over the day before the interview. Here a trip is defined as any movement performed by the respondent by any mode of transport, of any duration, and of any origin or destination. In this survey, we know for each trip the departure and arrival times and places (spatially divided into fine zones), the distance, the duration, and the purpose.

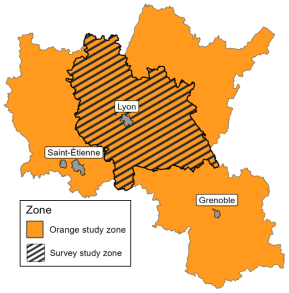

Figure 1 – *Map of the Study area*

- Mobile phone data provided by the main French telecom operator Orange. It consists of data collected in September and October 2022 on Lyon metropolis and its proximity (figure 1), and pre-processed by the telecom operator. The study zone is divided into 802 areas of 27 km^2 on average.

A trip is defined by this telecom operator as follows:

Definition 1 *Let U be an user. Let S_1, ... S_n be the different stationarity periods of U during a whole day detected by the telecom operator. Each S_i is related to a zone Z_i in which U is located and T_i is the duration of each S_i. A trip is detected between S_i and S_j by the telecom operator if it satisfies:* $\boxed{T_i \geq T_w,\ T_j \geq T_w \text{ and } Z_i \neq Z_j}$*, with T_w a chosen threshold (set to 1 or 3 hours).*

Anonymization being mandatory in France (by the CNIL, the French Data Protection Authority), mobile phone data are aggregated. The minimum volume of people making the same Origin-Destination trip must be greater than 20, otherwise, this flow is not taken into account.

3 COMPARATIVE ANALYSIS

3.1 Application of the definition of trip on survey data

The definition of *a trip* is not always the same in the literature. So we wanted to know what the definition of this telecom operator implies (definition 1). We started by applying this definition to HTS data for stationarity periods T_w of 1 and 3 hours.

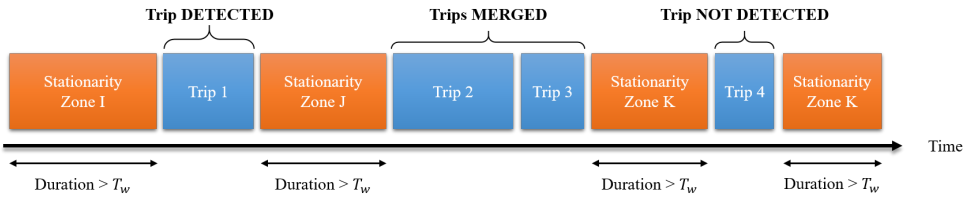

Figure 2 – *Illustration of possible trips detected, not detected, or merged with this definition.*

Each trip is labeled "detected" or "not detected" if it is taken or not into account by the definition. Others are labeled "merged" because they are part of a detected trip but multiple trips have been merged into one (figure 2).

With the definition of trips used by the telecom operator, we find that only 47% of survey trips (for $T_w = 1$ hour) would have been detected (in which 10% would have been considered merged). For 3 hours of stationarity, we found 36% (with 14% of merged trips).

As one would expect, short time and distance trips are mostly undetected (90% of the trips under 1km). In addition, long trips are over-represented (80% of trips above 10km are detected). This seems consistent as short trips have been grouped under the status "merged". For the same reasons of time and distance, we find that some modes of transport are less detected than others. For example, 85% of walking trips are not detected against 15% of public transport trips (for 1 hour of stationarity).

Finally, concerning travel purposes, we have plotted the 9 most common in Figure 3. We found that in the survey filtered data, about 80% of the Home-Work and Work-Home trips are retained, the same for Home-Studies with about 50%. But for many of the other patterns, at least 70% or more of the trips are lost.

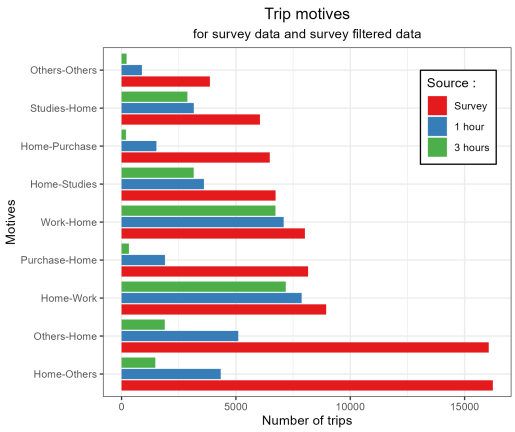

Figure 3 – *Trip motives and their impact on detection.*

3.2 Average number of trips per person and per day

To see if there are similarities between the mobile phone data and the survey filtered data, we first looked at whether the average number of trips per person per day during weekdays was close. For the telecom operator data, since the data are aggregated, we cannot retrieve the trips of a user for a whole day. Therefore, during pre-processing we assumed that these data were well scaled to the population level of the study area, and since the population considered in the study area is about 4.5 million inhabitants in 2022, we calculated the average number of trips per inhabitant by dividing the total number of observed trips by this value. Furthermore, the telecom operator data come from all mobile phones present in the area, while the HTS respondents are only residents. To compare responses from residents only, we worked with a version of the OD data with an additional split based on inferred residences, workplaces and on nationality (French or foreign). Note that this additional split implies a non-negligible loss of volume, as more cells fall below the anonymization threshold.

HTS 2015	Survey filtered data		Telecom (Residents trips only)		Telecom (All trips)	
	1 hour	3 hours	1 hour	3 hours	1 hour	3 hours
3.55	1.46	0.98	1.3	0.42	1.9	0.7

Table 1 – *Average number of trips per person and per day.*

With these considerations, we find the results of the table 1, where we set the lower (only resident trips) and upper limits (all trips) for the average number of trips recorded by the telecom operator. Compared to the data filtered by the survey, we can see that the averages are close, so the volume seems to be coherent.

3.3 Home-Work matrices

We explained earlier that the telecom operator's definition of a trip seemed to retain a majority of home-work trips. We therefore wanted to construct home-work matrices (from the telecom operator's preprocessing) and compare them with those obtained by INSEE (National Institute for Statistics and Economic Studies) during the population census. For the analysis, we placed these two types of data on the same spatial division. However, the INSEE data are provided by the municipality, while the mobile phone data have a specific spatial division chosen by the telecom operator.

For this reason, we decided to disaggregate the data from the telecom operator uniformly to a spatial structure smaller than both, and then re-aggregate by municipality. The result of the scatter plot between the home-work matrices is shown in Figure 4. Here we notice a lack of volume in the telecom operator data. But we found a decent coefficient of determination R^2 equal to 0.65. And even equal to 0.72 if we consider only flows above 200[a]. Moreover, 91.6% of INSEE flows above 100 are captured by the telecom operator, 97.6% above 200 and all flows above 500. Thus, the structure of the two matrices seems to be equivalent, although we observe fewer trips in the mobile data.

[a]As indicated by INSEE in its documentation, weak flows <200 should be considered as orders of magnitude.

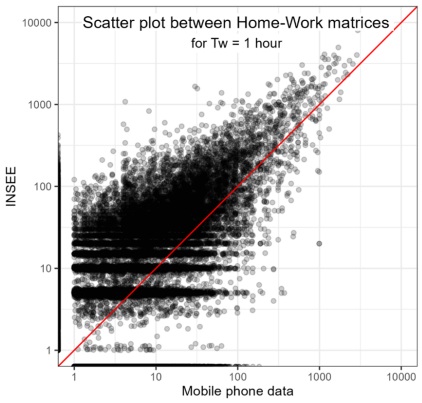

Figure 4 – *Scatter plot between Home-Work matrices (in logarithm).*

4 CONCLUSION

In conclusion, our work shows that the definition of *trip* in mobile phone data implies different features from those provided by HTS. Short trips in time and distance are lost, while long trips are over-represented. Modes of transport are not captured in the same way. This definition mostly retains home-work and home-study trips. However, when this definition is applied to the HTS data, the volume of data appears to be consistent between the survey filtered data and the mobile phone data. Similarly, the structure of the home-work matrix seems to be consistent with the INSEE census data. Due to lack of space, we have not been able to develop here our results concerning the analysis of metadata inferences (places of residence, places of work, nationalities, means of transport, etc.), the impact of the threshold T_w or the departure time of the trips on their detection. Finally, we must remember that our work has limitations. Firstly, the spatial study area and the period are not the same for both studies. Differences in urban and peri-urban trips compared to rural areas should be investigated. Secondly, there is a time lag between the HTS and the telecom operator data, COVID-19, and possible behavioral changes in the population.

Aknowledgment

This research is supported by the French ANR research project MOBITIC (grant number ANR-19-CE22-0010).

References

Bonnetain, Loïc, Furno, Angelo, El Faouzi, Nour-Eddin, Fiore, Marco, Stanica, Razvan, Smoreda, Zbigniew, & Ziemlicki, Cezary. 2021. TRANSIT: Fine-grained human mobility trajectory inference at scale with mobile network signaling data. *Transportation Research Part C: Emerging Technologies*, **130**, 103257.

Chrétien, Julie. 2016. Algorithm-induced biases in data representativeness: the case of activity inference from mobile phone traces. *In: American Association of Geographers Annual Meeting.*

Fekih, Mariem, Bellemans, Tom, Smoreda, Zbigniew, Bonnel, Patrick, Furno, Angelo, & Galland, Stéphane. 2021 (01). Suitability of Cellular Network Signaling Data for Origin-Destination Matrix Construction: a Case Study of Lyon Region (france). *In: TRB 2019, 98th Annual Meeting Transportation Research Board.*

Huang, Haosheng, Cheng, Yi, & Weibel, Robert. 2019. Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*, **101**, 297–312.

Sfeir, Georges, Rodrigues, Filipe, Zeid, Maya Abou, & Pereira, Francisco Camara. 2024. Analyzing the Reporting Error of Public Transport Trips in the Danish National Travel Survey Using Smart Card Data. *In: TRB 2024, 103th Annual Meeting Transportation Research Board.*