Human Mobility from GSM Data - A Valid Alternative to GPS?

Daniel Schulz
Fraunhofer IAIS
Schloss Birlinghoven
Sankt Augustin, Germany
daniel.schulz
@iais.fraunhofer.de

Sebastian Bothe
Fraunhofer IAIS
Schloss Birlinghoven
Sankt Augustin, Germany
sebastian.bothe
@iais.fraunhofer.de

Christine Körner
Fraunhofer IAIS
Schloss Birlinghoven
Sankt Augustin, Germany
christine.koerner
@iais.fraunhofer.de

ABSTRACT

Characteristics of human mobility are a valuable source of information in many applications. In this paper we evaluate the usability of call detail records for the extraction of mobility quantities. We derive several quantities from the simultaneously collected GPS and GSM mobility data of the Nokia Mobile Data Challenge. Our analyses show that GSM activity data underestimates average daily travel distance and radius of gyration when derived straightforward from the data. In addition, they indicate that the correlation between mobile phone usage and movement quantities is biased when using GSM activity data. Finally, our analyses confirm that long-term GSM activity data is well suited to detect frequent stop locations.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Spatial Databases and GIS

Keywords

GSM, GPS, comparative study, mobility quantities

1. INTRODUCTION

Mobile phone data is an interesting data source for mobility data analysis because it has a wide coverage within the population and of the geographic area. However, as the data is primarily collected for billing purposes and network optimization, it is not tailored to the needs of mobility analysis and modeling. On the one hand, call detail records (CDR) collected for billing purposes capture only snapshots of users in time because they are restricted to call activities. On the other hand, mobile phone data has a coarse spatial resolution which can vary between a few hundred meters and a few kilometers depending on network structure. In this paper we evaluate the usability of mobile phone data for the extraction of mobility quantities. We are especially interested in quantities that can be applied to enrich mobility models, such as average daily travel distance and radius of gyration. In addition, we analyze the correlation between mobile phone usage and mobile behavior as well as the locations of mobile phone usage. Our analysis is based on the Mobile Data Challenge (MDC) data set of the Open Track, which contains a parallel GPS and GSM survey of 38 participants in Lausanne, Switzerland, for over one year.

2. DATA PREPROCESSING

The MDC data set contains regularly recorded GPS positions and GSM cells as well as call activities. During preprocessing we use the GPS and GSM data to approximate the cell geometry of the mobile phone network and to obtain CDR like data. In addition, we perform stop detection on the GPS records.

We approximate the cell geometries by constructing Voronoi polygons based on the median of GPS positions within up to 15 minutes of recorded GSM cells (we increasingly searched for points in a 30 seconds, 2, 5, 10 and 15 minutes time window). A picture of the obtained tessellation in the area of Lausanne is depicted in Figure 1.

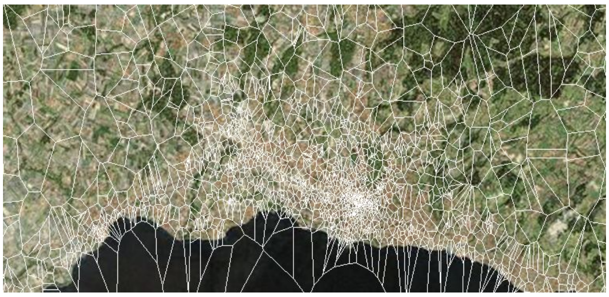

Figure 1: Approximated GSM cells

In order to produce a data set similar to CDRs, we reduced the continuously collected GSM cell data to records during call activities. To achieve this we selected the closest GSM record within a time window of 2 hours around a call activity. Not all call activities could be matched to a GSM record, resulting in a data set with about 77.6% of all call activities. The reason for this difference is not clear. However, it means that in our analyses we potentially underestimate mobility quantities from call activities. We will call the resulting set the *GSM activity* data set in the following.

This material was prepared for the Mobile Data Challenge 2012 (by Nokia) Workshop; June 18-19, 2012; Newcastle, UK. The copyright belongs to the authors of this paper.

Finally, we performed stop detection on the GPS data in order to identify spatial locations where the user has no or little movement. We consider a location in which the user remains within an radius of 300 meters for at least 1800 seconds as a stop. In a second step we clustered the obtained stops using the DBSCAN algorithm in order to identify frequently visited locations using a distance threshold of $\epsilon = 300$ meters and a minimum number of $minPts = 3$ neighbors.

3. COMPARISON

We began our comparison with two basic quantities describing mobile behavior, namely travel distance and radius of gyration. For the evaluation of both quantities we reduced the data set to measurement days where GPS, GSM as well as call activities were available. Thus the quantities relate to the same underlying mobility when compared across different measurement technologies.

3.1 Travel Distance

We calculated average daily travel distances for all users from the GPS and GSM activity data set. From the GPS data we hereby calculated two different travel distances. On the one hand, we summarized the distance between any two consecutive GPS points excluding those points inside of a stop. Second, we calculated the travel distance considering only the centroid coordinates of stop locations. For the GSM activity data we calculated the average daily travel distance between consecutive GSM activities using the estimated cell centroids. The results are depicted in Table 1.

Table 1: Comparison of average daily travel distances (in km)

GPS Sequence (excl. stops)	Between GPS Stops	Between GSM Activity
39.10	19.22	18.56

The daily average travel distance calculated from GPS fits well with statistics by the Swiss Bundesamt für Statistik (BFS), which states an average travel distance of 38.2 kilometers for 2005 [1]. Also more recent travel surveys from other countries affirm the result. For example, the German travel statistic "Mobilität in Deutschland" states an average daily travel distance of 41 kilometers per person for 2008 [3]. When calculating the distance only between stops detected within the GPS data, the travel distance divides in half (ratio of 0.47). This means that a substantial part of mobility is lost when relying only on stop locations. Certainly, the distance is underestimated because it represents the air-line distance between stop locations. However, the difference seems too large to be explained by this fact alone. It remains a task of future work to repeat the calculations using distance based on the street network. Interestingly, the average daily air-line distance between consecutive GSM activity locations is very close to the travel distance calculated from GPS stops. This is a first hint that GSM activity data may be adequate to detect frequent stop locations. We will explore this topic further in Section 3.4. In summary, we conclude that a direct estimation of travel distance from GSM activity data underestimates the true average daily travel distance.

3.2 Radius of Gyration

The radius of gyration (ROG) is a quantity to measure the spatial extent of a person's mobility and is defined as

$$r_g = \sqrt{\frac{\sum_{i=1}^{N} (p_i - \overline{p})^2}{N}} \quad \text{with} \quad \overline{p} = \frac{\sum_{i=1}^{N} p_i}{N}.$$

Hereby p_i denotes a single position in the trajectory of a user. The radius of gyration has been analyzed in several previous studies using mobile phone data ([2, 5, 6]). However, an evaluation of the quality of the radius of gyration when derived from mobile phone data has not been made yet. In this paper we derive the radius of gyration (a) per user and (b) per average day of a user. The former quantity allows insights about variation in long-term mobility of a user. However, in a long-term observation period it is likely that a user makes long-distance trips (e.g. for vacation), which distort results as the radius of gyration relies on squared distances. Therefore, we also formed average daily radii of gyration per user. Table 2 shows the results for both quantities when averaged over all users. Again we calculated the ROG from the GPS data excluding stop locations, from the GPS stop locations, from the GSM activity data and additionally from the full set of given GSM data.

Table 2: Comparison of radius of gyration (in km)

	GPS Seq.	GPS Stops	GSM Activity	GSM
avg. per user	25.60	20.54	17.14	16.77
avg. per user day	5.10	4.92	4.08	7.13

Considering the average ROG of users over the whole observation period, we see again that the GSM activity data captures only a part of the mobility captured via GPS. The ratio is 0.67, which is higher than the ratio for average daily travel distance (see Section 3.1). However, as the radius is used to describe an area of activity, the ratio becomes worse when comparing the two circular areas described by the radii (ratio of 0.45). Similar to the previous section we also calculated the ROG from the GPS stops. This ROG reaches about 0.80 of the GPS radius and lies thus considerable above the GSM activity radius. It means that although average daily travel distances between GPS stops and GSM activities have been similar, the GSM activity data does not cover the complete information of stop locations. In addition to these three radii, we also calculated the ROG for the complete set of GSM cells. Surprisingly, its value is close (even a little below) to the ROG of the GSM activity data. This indicates that the primary loss of information of ROG when derived from GSM does not come from the temporal selectivity of call activities. This result is reasonable considering that the evaluation relies on long-term observations and that human mobility is repetitive over time [4, 2]. One possible explanation for the effect could be the broader spatial granularity of observation in comparison to GPS.

In the second analysis we calculated the radius of gyration as daily average for each user. Here, the ROG of GPS and of stop positions is nearly identical. This means that on a daily basis stop positions provide a quite accurate picture of travel behavior. Further, the ROG of GSM activity data reaches 0.80 of the ROG of GPS. On first sight this number improves when comparing it to the ratio of the previous analysis. Our expectation would have been the opposite, i.e. to obtain a smaller ratio because it is less likely that a user will make calls or write messages from all his visited location within a single day. The ROG derived from the complete GSM traces provides a possible explanation for this contradiction. The average daily ROG of GSM is considerably higher than the ROG of GPS. This difference is likely to result from the coarser spatial resolution of the GSM data. Clearly, a coarse spatial resolution has a much higher impact on a small geographic region (area of daily movement) than on a large geographic region (area of long-term movement). In the case of GSM this seems to lead to an overestimation of ROG. In consequence it is likely that also the ROG of GSM activity is overestimated, resulting in a higher ratio than anticipated. However, this effect has to be studied in more detail in future work.

In summary, our analysis indicates that GSM activity data underestimates the radius of gyration. For short-term analyses covering a limited geographic area, an opposite effect due to the coarse spatial granularity of GSM cells seems to improve results. As the granularity of cells differs between geographic regions (e.g. cities, rural areas) the effect may differ in strength and has to be analyzed in further studies.

3.3 Correlation of Mobile Phone Usage and Travel Behavior

When mobile phone data shall be used for the estimation of mobility characteristics, a sufficiently high sampling rate (i.e. number of calls) should be available per user. For example, assume that we have one user with a high mobility and another user with a low mobility. If the call frequency of both users is low, reflecting only one typical location, the GSM activity data will not be able to distinguish the movement behavior of the two users. If the user with a low mobility calls more often than the user with a high mobility, the estimation from GSM activity data may even be reversed. Recently, [6] analyzed a large collection of GSM activity data and found a high correlation between the number of calls and the radius of gyration. However, their analysis contains only mobility characteristics from GSM activities. In our next analysis we therefore correlated mobile phone usage with both movement quantities derived from GSM activity data and quantities derived from the GPS data. Table 3 shows the correlation between the average number of daily GSM activities (e.g. incoming/outgoing call or message) and the average daily travel distance as determined in Section 3.1 as well as the radius of gyration per user and average user day as determined in Section 3.2. Figure 2 shows the detailed results for comparing GSM activity with average daily travel distance (left) and ROG per average user day (right).

Table 3: Correlation between average daily GSM activities and mobility quantities derived from GPS data (excluding stops) and GSM activity data respectively

	Average Daily Travel Distance	ROG per Average User Day	ROG per User
GPS Sequence	0.219	0.274	-0.126
GSM Activity	0.546	0.503	0.011

The correlation between mobile phone usage and GPS movement statistics is only weak. In case of ROG per user even negative. However, when performing the same analysis with movement quantities obtained from the GSM activity data, the correlation increased to a medium level. This indicates that correlation analyses based on GSM activity data are biased towards an overestimation. Surprisingly, the ROG per user did not correlate with mobile phone usage in either case. Several possibilities exist why our results differ from the results in [6]. First, [6] group the data before correlation analysis in order to reduce variation. Second, the data in [6] is available for only nine days. Lastly, the mobile phone usage between different nationalities may differ. However, our data set is very small, containing only 38 users. We therefore see it as necessary future work to repeat the analysis with the full set of users in the MDC data set and to explore the differences further.

3.4 Analysis of Call Locations

In our last analysis we examine call locations. As mentioned in Section 3.1 and analyzed by [5] GSM activity data seems well suited to identify frequent stop locations. In this analysis we first determined the proportion of GSM activities that take place within typical stop locations identified from the GPS trajectories. Typical stop locations are hereby defined as a cluster of at least three stops of a user that are close in space (see also Section 2). Figure 3 shows the proportion of calls that take place within the first most often visited stop cluster, the first two most often visited stop clusters etc. About 69,4% of all calls take place within the nine most important stop locations. This means that GSM activity data is a good source to identify and analyze stop locations. However, it also means that GSM activities mostly tell us about where people stay, not where they move. GSM activity data will therefore be less adequate to measure the amount of traffic on the street network at a given moment in time, which is also an important question in mobility analysis.

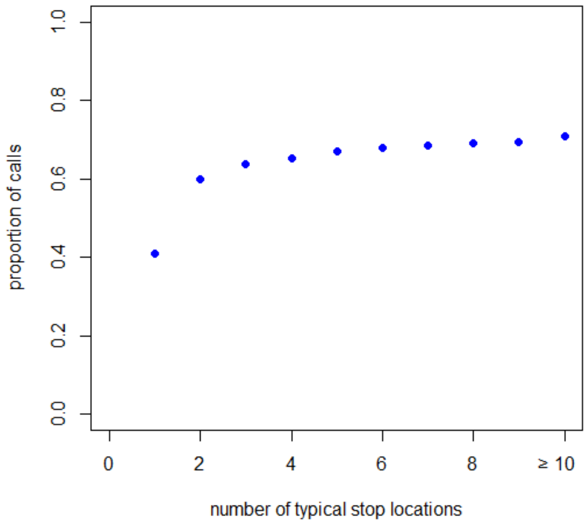

Figure 3: Proportion of calls conducted in the most often visited stop locations

So far we considered the proportion of calls that take place at stop locations. Of course it is also important to

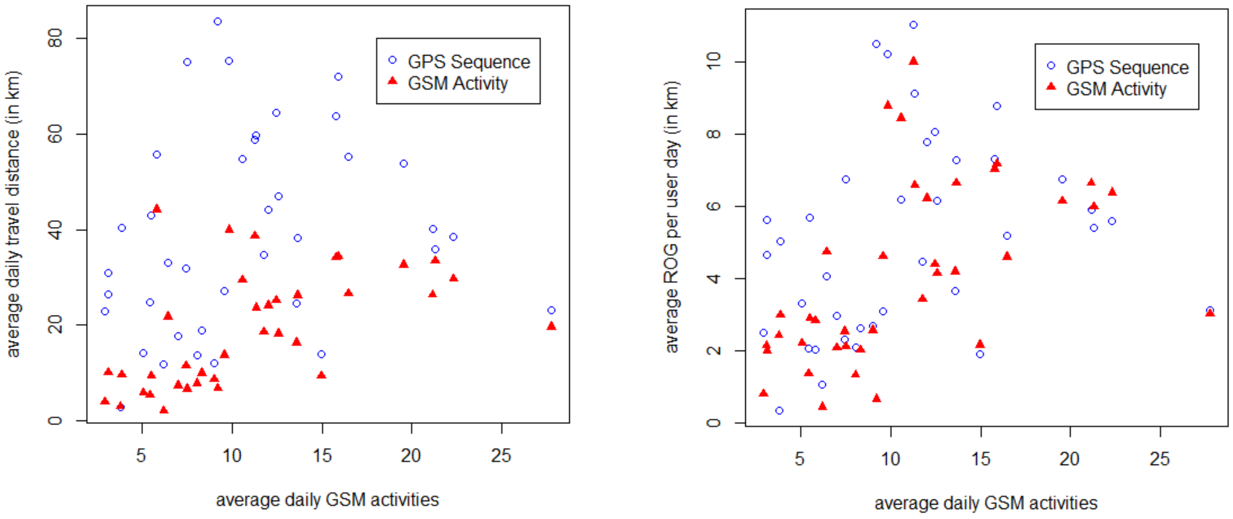

Figure 2: Correlation of GSM activity with average daily travel distance (left) and ROG per average user day (right)

know how many stop locations can potentially be detected through GSM activity data. When considering the clustered stop locations along with the remaining single stops, a total of 61,3% of the locations could be covered by GSM activities. When separating this number into frequently visitied locations and the remaining single locations, the covered percentages are 89.9% and 56.9% respecetively. This means that GSM activity data has a very high probability to identify typical stop locations, however, only about half of the not frequently visited locations are covered. This result also explains why GSM activity data has a lower radius of gyration than obtained from the GPS stops.

4. CONCLUSION

In this paper we evaluated the quality of mobility quantities when derived from GSM activity data. Our analyses are based on the Open Track data set of the Nokia Mobile Data Challenge, containing a long-term parallel GPS and GSM activity survey. We performed four types of analyses, comparing (1) average daily travel distance and (2) radius of gyration for long- and short-term observation. Further, we (3) calculated the correlation between mobile phone usage and the first two quantities and finally (4) analyzed typical call locations.

While our analyses confirm that long-term GSM activity data is well suited to identify typical stop locations, they also show that a straightforward derivation of average daily travel distance and radius of gyration from GSM activity data underestimates the respective quantity. In addition, our analyses indicate that the correlation between mobile phone usage and movement quantities is biased when using GSM activity data. However, these results have to be confirmed in further work as our data set contained only 38 persons. In addition, we were able to assign only 77.6% of all call activities to GSM cells, which will partially have caused the underestimation. Nevertheless, our analyses underline the necessity to evaluate GSM activity data with other mobility data sources and to assess advantages and shortcomings of this data source. Especially further analyses of simultaneously collected mobility data will contribute to such an evaluation and are required in order to obtain objective results. We are confident that such studies will allow to develop methods for the reliable estimation of movement quantities from GSM activity data.

5. ACKNOWLEDGMENT

This work has been carried out under the European FP7 Open-FET project Data Science for Simulating the Era of Electric Vehicles (DATASIM) (project number FP7-ICT-270833).

6. REFERENCES

[1] R. Evéquoz. Mobilität und Personenverkehr. In *BFS Aktuell, 11 Mobilität und Verkehr*. Bundesamt für Statistik (BFS), 2011.

[2] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[3] Institut für angewandte Sozialwissenschaft GmbH (infas), Deutsches Zentrum für Luft- und Raumfahrt e.V. - Institut für Verkehrsforschung (DLR). *Mobilität in Deutschland 2008, Abschlussbericht*. Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2010.

[4] R. Schlich and K. W. Axhausen. Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation*, 30:13–36, 2003.

[5] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human motion. *Science*, 327(5968):1018–1021, 2010.

[6] Y. Yuan, M. Raubal, and Y. Liu. Correlating mobile phone usage and travel behavior - a case study of harbin, china. *Computers, Environment and Urban Systems*, 36(2):118–130, 2012.