Kernel-based Approach to Reconstruct Travel Diaries from GSM Records

Zahra Eftekhar*, Adam Pel, Hans van Lint
Delft University of Technology, The Netherlands

November 2020

keywords

travel demand; GSM; Bayesian; Kernel; temporal pattern.

1 Introduction

Accurate estimation of travel demand allows decision-makers to improve the planning and operation of the transportation network (Chan et al. [2007]). This can traditionally be based on using only travel diaries which provide a high level of detail in activity and movement behavior. However, they have minimal sampling ratios; therefore, they are liable to sample bias and reporting errors [Hajek, 1977, Kuwahara and Sullivan, 1987, Groves, 2006].

Call Detail Records (CDR), on the other hand, with a reasonable size of low-cost GSM data on people's movements, contains (much) less detail than travel diaries. It contains discretized traces of users without the precise time and location of the underlying activities or activity types. Therefore, only after analysis, CDR data is useful to estimate the travel diaries. In fact, using a data fusion technique to combine travel diaries in such CDR analyses could lead to the best of both worlds; that is, the high sampling ratio of CDR combined with the high level of detail (spatial and activity patterns) in travel diaries.

CDR reported locations are initially **unlabeled**. In other words, whether the reported CDR location is *pass-by* (i.e., person was traveling) or *stay* (i.e., person was attending an activity) cannot be directly concluded [Zilske and Nagel, 2014]; Therefore, one must interpret the GSM data to **reconstruct** the underlying travel diaries. To achieve this, a number of previous research has specified a certain duration (or speed) to distinguish *stay* from *pass-by* locations.For instance, Iqbal et al., Alexander et al., and Demissie et al. only theoretically assumed that a trip is recorded if, in the CDR, each user's subsequent entries

*presenter

indicate location change with a time difference of more than 10 minutes but less than 1 hour. However, these studies raise a question of how to select and validate the threshold; thus, we aim to empirically derive the optimum one that accounts for the study context. To achieve this, we develop and validate a new method to interpret the unlabeled reported records. Our method separates *pass-by*s from *stay*s using the temporal patterns extracted from the associated travel diaries. For full experimental control and avoiding privacy issues regarding real GSM data, our method, **Kernel-based approach (KA)**, gets evaluated by applying it to synthetic travel diaries to see how successfully it can distinguish the type of reported locations given 1% of the associated population's travel diaries. Regarding this, the synthetic travel diaries were generated based on real-life travel surveys in the calibration process of an agent-based activity-based travel demand model (see Timmermans and Arentze [2011]).

Our evaluation indicated that in 94.4% of the time, the location type (*stay* or *pass-by*) was distinguished correctly; therefore, it seems that one can reconstruct travel diaries (associated with the GSM data) even from a small sample of travel diaries that contain as little as one percent of the population's movements. Since the method makes no assumptions on the temporal distributions of activities and trips, KA is potentially generalizable to empirical (i.e., GSM) data and other networks.

2 Material and Method Overview

The adopted synthetic dataset for evaluating KA contained travel diaries of 22 thousand car drivers conducted —albeit indirectly— during a representative working day within the Amsterdam network. The selected training set contains travel diaries of 220 drivers (1% of the population). Our research approach is threefold:

- pre-processing and selecting the training set from the entire travel diaries (i.e., travel diaries of 1% of the population)
- Developing and evaluating KA to reconstruct the travel diaries from the GSM data (i.e., unlabeled travel diaries)
- assessing how successfully KA performs by comparing the travel diaries with the reconstructed ones using the associated confusion matrices

In the next part, we explain our proposed method in more detail.

2.1 Kernel-based approach of interpreting the unlabeled travel diaries

To identify location type (*stay* or *pass-by*), we used a **Bayesian classifier** trained by a random one percent of the travel diaries (see [Yair and Gersho, 1990]). The classifier uses the distribution of duration and start time of the training records in each event (*stay* or *pass-by*) to identify the category of each

	observed *stay*	observed *pass-by*	total predicted
predicted *stay*	54231 (91.8%)	1799 (3%)	56030 (47.4%)
predicted *pass-by*	4822 (8.2%)	57254 (97%)	62076 (52.6%)
total observed	59053 (100%)	59053 (100%)	118106 (100%)

Table 1: Confusion matrix of applying KA on the entire travel diaries for location type recognition.

record in GSM data. The same way, the temporal distributions of *stay*s (in the training set) were used to detect the activity category (*home* or *work* or *other*) since detecting the location of *home* helps to distinguish the miss-identified *stay*s. The major reason of selecting duration and start time of events as explanatory variables is that location and activity categories have particular temporal patterns separating them from each other.

The Bayes Rule requires an **a-priori** knowledge about the probability density functions of the priors (i.e., duration and start time of events). Assuming that the observed data points in the training set are a sample from an unknown probability density function, **density estimation** is the construction of an estimate of the density function from the training set. Regarding this, we used **kernel density estimation** (KDE), currently the most popular **non-parametric** approach for probability density estimation [Scott, 2012, Simonoff, 2012]. Accordingly, our proposed method is named **Kernel-based approach (KA)**.

3 Method Result

KA performance was tested by applying it for location (and activity) category recognition on the entire travel diaries, using an example training set (i.e., travel diaries of 1% of the population). The generated confusion matrix is shown in Table 1.

Based on Table 1, *stay* detection had more false-negative errors than *pass-by* detection; therefore, underestimation in *stay* detection was more probable.

Based on our initial analysis, due to closeness of *stay* and *pass-by* starts (because travel time is generally shorter than activity duration), duration of events has a more significant role in differentiating them. Regarding this, observing the results showed that the minimum duration of correctly recognized *stay*s was about 44 minutes. Therefore, it seems that KA empirically specifies a duration threshold to separate *stay*s from *pass-by*s. Moreover, our analysis shows that about 88% of activities have a duration of more than 44 minutes, whereas 96% of trips endure less than 44 minutes. Thus, activities are less likely to be recognized. This justifies the underestimation in activity detection in Table 1 (KA accuracy was 92% for *stay*s and 97% for *pass-by*s).

Table 2 shows KA performance for activity category detection (*home*, *work*

	observed *home*	observed *work*	observed *other*	total predicted
predicted *home*	22460 (82%)	50 (0.4%0	144 (0.8%)	22654 (38.4%)
predicted *work*	307 (1.1%)	13008 (95%)	236 (1.3%)	13551 (22.9%)
predicted *other*	4631 (16.9%)	630 (4.6%)	17587 (97.9%)	22848 (38.7%)
total observed	27398 (100%)	13688 (100%)	17967 (100%)	59053 (100%)

Table 2: Confusion matrix of applying KA on the entire *stay*s for activity recognition.

or *other*); overall, 87% of activities were distinguished with the correct activity category.

4 Conclusion

The KA validation results on the synthetic data show that in 94% of the time, the location type (*stay* or *pass-by*) was distinguished correctly. Moreover, with our method, it is possible to reconstruct travel diaries even from a sample of travel diaries that contain as little as one percent of the population's movements.

Since the method makes no assumptions on the temporal distributions of activities and trips, KA is potentially applicable to actual data and other networks. To substantiate that claim, more elaborated analyses and experiments can and must be performed with KA to comprehensively understand the GSM data characteristics, particularly related to temporal and spatial discretization and user-group biases, which we have planned to investigate in separate research. Finally, KA is a promising approach that allows us to take advantage of a significant source of low-cost GSM data and reduces the need for detailed travel diaries. Furthermore, with its data-driven interpretation of records, no theoretical assumption is needed. Moreover, the insignificant requirement of travel diaries reduces the bias of the estimated travel demand.

References

L. Alexander, S. Jiang, M. Murga, and M. C. González. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58:240 – 250, 2015. ISSN 0968-090X. Big Data in Transportation and Traffic Engineering.

J. Chan et al. *Rail transit OD matrix estimation and journey time reliability metrics using automated fare data.* PhD thesis, Massachusetts Institute of Technology, 2007.

M. G. Demissie, S. Phithakkitnukoon, and L. Kattan. Trip distribution modeling using mobile phone data: Emphasis on intra-zonal trips. *IEEE Transactions on Intelligent Transportation Systems*, 20(7):2605–2617, 2019.

R. M. Groves. Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5):646–675, 2006.

J. J. Hajek. Optimal sample size of roadside-interview origin-destination surveys. *TRB Transportation Research Board*, 1977.

M. S. Iqbal, C. F. Choudhury, P. Wang, and M. C. González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63 – 74, 2014. doi: https://doi.org/10.1016/j.trc.2014.01.002.

M. Kuwahara and E. C. Sullivan. Estimating origin-destination matrices from roadside survey data. *Transportation Research Part B: Methodological*, 21(3): 233–248, jun 1987. doi: 10.1016/0191-2615(87)90006-3.

D. W. Scott. *Multivariate Density Estimation and Visualization*, pages 549–569. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-21551-3.

J. S. Simonoff. *Smoothing methods in statistics*. Springer Science & Business Media, 2012.

H. Timmermans and T. A. Arentze. Transport models and urban planning practice: Experiences with albatross. *Transport Reviews*, 31(2):199–207, 2011. doi: 10.1080/01441647.2010.518292. URL https://doi.org/10.1080/01441647.2010.518292.

E. Yair and A. Gersho. Maximum a posteriori decision and evaluation of class probabilities by boltzmann perceptron classifiers. *Proceedings of the IEEE*, 78(10):1620–1628, 1990.

M. Zilske and K. Nagel. Studying the accuracy of demand generation from mobile phone trajectories with synthetic data. *Procedia Computer Science*, 32: 802 – 807, 2014. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2014.05.494. The 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014), the 4th International Conference on Sustainable Energy Information Technology (SEIT-2014).