Dynamic Estimation of Urban Zonal Speed from Mobile Sensing Data and Macroscopic Paths

Manon Seppecher[1,2], Ludovic Leclercq[2], Angelo Furno[2] & Delphine Lejri[2]

[1]Citepa, Paris, France
[2]Univ. Gustave Eiffel, Univ. Lyon, ENTPE, LICIT, F-69518, Lyon, France
{manon.seppecher, ludovic.leclercq, angelo.furno, delphine.lejri}@univ-eiffel.fr

Extended abstract submitted for presentation at the hEART 2020 9th Symposium
Sept. 1–3, 2020, Lyon, France

Word count: 2685 words (excluding the references)
February 14, 2020

Keywords: GPS data, CDR data, traffic speed, travel time, regions, macro-paths

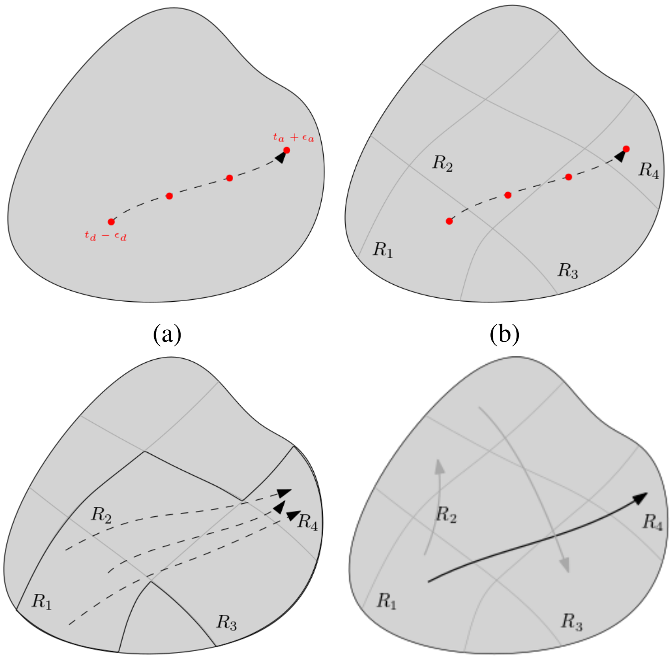
Extended abstract

In the past decades, call detail records (CDR), a specific type of mobile phone data, has been proven to be an accessible and rich source of information about mobility. This data, generated by mobile phone users while communicating, is collected and stored by the communication data providers. It registers for each communication event (calls, messages or data browsing) the user's unique identification key, the timestamp of the event, and the location of the base station antenna that processed it. This basic structure confers to this data a strong potential for mobility analysis.

However, when compared to GPS data classically used in mobility and traffic analysis, CDR data can present significant drawbacks. First, as the positional information is obtained at the base station scale, the spatial precision depends on the density of the base station network. Second, the data generation depends on the users' communication behaviours, making the data acquisition uneven and sparse in time; in particular, users with little communication activity will generate less location data and their mobility will be especially difficult to estimate. At the same time, CDR data have important advantages: they are usually accessible, massive, and of better spatial coverage because they are not restricted to a single category of users, unlike many GPS datasets.

Given those advantages, we wonder if CDR data could be used as a reliable alternative to GPS data in order to estimate the mean traffic speed dynamics at a zonal scale, in an urban area. Aside from the challenge that speed estimation from sparse data represents, or from the insight it can give on the traffic dynamics at a large urban scale and for a large sampled population, this speed estimation presents a high potential in the field of traffic emission estimations. While using GPS data for this purpose requires important complementary data and costly scaling up processes (10), an estimation using CDR data could be an efficient and light alternative.

An extensive literature exists on mobile phone data usage for mobility analysis, from the development of mobility choice models (7), to the construction of CDR-based origin destination matrices (1, 5, 8). However, the studies related to the traffic characterisation are more limited. In Toole et al. (11), a CDR-based origin-destination matrix is estimated in a first step, then assigned onto the road network in a second step, to estimate the traffic load. This step-wise approach is completed by studies exploiting two other types of mobile phone data, handovers and location area updates data, to estimate traffic speeds on specific highway segments (2), travel time (9) or macroscopic fundamental diagrams



(a) (b)

(c) (d)

FIG. 1 – *Visual recap of the methodology*

(6). Still, the potential of CDR data specifically for traffic characterisation remains weakly assessed and, in particular, the question of speed estimation needs to be more deeply investigated.

We propose an innovative methodology to estimate the temporal dynamics of the average traffic speed at zonal scale, in an urban context, using a limited and selected set of individual trips characterised by a low level of information. Our method relies on partitioning the studied urban network into a set of regions, and on uniformly discretising time into periods (15 minutes here). The individual trips are characterised at this spatio-temporal resolution, with the following trip features: *trip Id, vehicle Id, macroscopic path, arrival period, total travel time*, where the *macroscopic path* (later called *macro-path*) corresponds to the succession of regions traveled. This scaled representation of the data allows to identify clusters of individual trips sharing two characteristics (identical *macro-path* and *arrival period*) and to estimate for each of these clusters a robust travel time. Providing that a good estimation of the trip lengths at the city and zonal scale is available, and assuming that for a given period the speed is homogeneous and constant in each region, we can conduct, through the construction of a linear system, a large numerical combined analysis of those travel times at a given time period, to deduce an estimation of the traffic speed in each region at this same period. Essentially, this means combining and analyzing jointly the typical travel time information of macro-paths that overlap (or share common regions) to deduce the underlying the traffic speed in the intersecting reservoirs (figure 1(d)). The process can be iterated throughout the day to get the full speed profile in each region. In practice, we show that we can build at each time period the following linear equation system :

$$S_t = \left\{ \bar{T}_{P,t} = \textstyle\sum_{r \in P} \bar{L}_{r,P} Y_{r,t} \qquad \forall P, \text{ with } Y_{r,t} = \frac{1}{V_{r,t}} \right. \tag{1}$$

where $\bar{T}_{P,t}$ is the average travel time along the *macro-path* P at period t, estimated from the macro-trip information, $\bar{L}_{r,P}$ the average distance traveled in region r along *macro-path* P, assumed to be exogenously known, and $V_{r,t}$ the average speed in region r at period t and unknown to be determined. S_t has as many unknowns as regions studied, and as many equations as *macro-paths* observed during period t. This system is usually overdetermined as the number of possible *macro-paths* is higher than the number of regions, some of them might as well be conflicting. This means that the system can very likely only be solved approximately using a least square optimisation method. The inverse of the optimal solution $Y_{r,t}$ corresponds to the traffic speed $V_{r,t}$ for each region r.

Theoretically, GPS data are precise enough to provide accurate estimations of $\bar{T}_{P,t}$. However, when it comes to using CDR data, the direct estimation of the travel time might not be as reliable. For a given mobile phone user, the good characterisation of their mobility is highly dependant on their communication activity. The more active they are, the more location data are collected, and statics phases (stays) can be separated from the in-between mobility phases. More precisely, if the user is very active, the departure and arrival time will be estimated with a limited time imprecision. On the contrary, for a barely active user, estimating their exact departure and arrival time will be more difficult and the observed travel time will be an overestimation of the real one. This means that the estimated traveled time of a trip observed from CDR data is actually higher than the real one. Consequently, equation (1) becomes:

$$S'_t = \left\{\bar{T}_{P,obs} - \bar{\epsilon}_t \simeq \textstyle\sum_{r \in P} \bar{L}_{r,P} Y_{r,t} \qquad \forall P, \text{ with } Y_{r,t} = \frac{1}{V_{r,t}} \right. \tag{2}$$

where $\bar{T}_{P,obs}$ is the average of the travel time observed along P at period t from sparse data, and $\bar{\epsilon}_t$ is the average bias existing at period t between the observed travel time and the real one. Thus, if we are able to estimate the average temporal bias induced by the human dependent sampling rate of mobile data, then it becomes possible to de-skew the observed the travel time along each *macro-path* P, and therefore to correct the system.

In order to reduce the system size and to limit the conflicts between equations, we propose to filter out of the system the equations that are the least reliable. While some *macro-paths* are observed for many macro-trips, some others might be representative of only one individual. Also, for a same *macro-path*, the variability of the individual travel times can be high. To take into account those two aspects, we implement a confidence interval threshold above which the corresponding *macro-paths* will be filtered out of the system.

The method we just exposed presents the advantage of considerably reducing the complexity of the problem. First, individual trips are gathered for each time period per *macro-paths* which provides a robust estimation of $\bar{T}_{P,t}$. Second, we can select among all *macro-paths* the most representative ones before looking for the system solution, which permits to reduce the system size. Third and foremost, we only need a very limited time information about trips, basically the arrival period and the travel time. This makes the method perfectly suitable for a CDR data input. The drawback is that it requires a robust and exogenous estimation of mean traveled distance within each region $\bar{L}_{r,P}$. The spatial definition of the regions is also crucial as it may drive very different patterns for different *macro-paths* inside the same region. This topic is still under research.

The methodology is intentionally applied on a large GPS dataset, progressively simplified and downsampled to reproduce the typical CDR data temporal characteristics and the temporal biases of mobile phone data. The downsampling process aims to simulate the temporal imprecisions of mobile phone data compared to GPS data. To do so, we adapt the method developed in Chen et al. (4) for spatial bias analysis to our data and problematic. We introduce in the data temporal gaps that reproduce the characteristic inter-event time of CDR data. We specifically focus on the impact of our downsampling on the detected beginning and end of the trip. This downsampling introduces a bias both on the trip start time and on the trip arrival time, and thus implies an increase of the global traveled time. In reality, using CDR data does not only introduces temporal bias but also distance imprecisions due to the sampling done at the antenna location and the lack of intermediary trajectory points. For the moment, we neglect this limit and consider that the *macro-paths* remain observable.

At this stage, monitoring the data downsampling process instead of using real CDR data directly has several strategic advantages. First, it allows to limit discrepancies and understand how the jamming and progressive information loss impact the quality of the results. Second, the original raw GPS

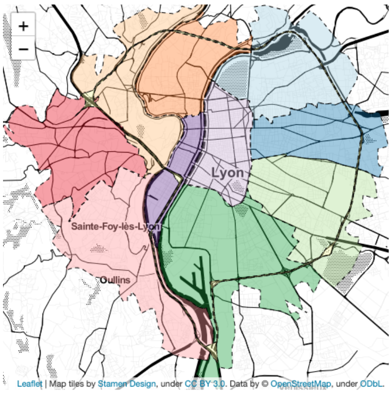

+
−
Lyon
Sainte-Foy-lès-Lyon
Oullins
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under ODbL.

FIG. 2 – *The 10 inner regions of Lyon*

data provides the average speed profile reference, considered as ground truth, to compare our method to the traditional ones, as well as the real travel time to which the observed CDR-like travel time can be compared to estimate the time bias. Last but not least, this full GPS data also gives access to a precise estimation of the trip length. Otherwise, without the GPS data, this estimation could be done with automatic network analysis (3). Eventually, the objective of this experiment in a very controlled environment is to make this method robust enough to be used later on real CDR data.

The city of Lyon, France, is chosen as case study. The study area includes both Lyon and the next-by municipality of Villeurbanne, which is located inside Lyon's ring road. The city is parted into a total number of 13 regions. 10 of them divide the inner city, while the 3 others split the ring road, as displayed in figure 2. For the inner regions, their geometry are constructed as an aggregation of "IRIS" units, the smallest census unit defined by the French National Institute of Statistics and Economic Studies, which fills both demographic and geographic criteria. This aggregation of regions is made so that it is consistent with Lyon's road network and its traffic characteristics.The GPS dataset used in the study was supplied by a European navigation system provider, and the traces come from vehicles equipped with their navigation system technology. This means that the dataset presents little bias compared to the GPS taxi datasets. Moreover, as each trace corresponds to a vehicle, there is no need to filter out pedestrian or cyclist travellers as there would be with non simulated CDR data. Though this will make the speed estimation process easier, it also means that the data we simulate do not exactly reproduce real CDR data that gather users making no distinction between their means of transport. The dataset consist of one year of GPS traces over the Great Lyon area. For this study however, we select the data of one typical weekday, 2018 February 5. As the data contains only a few trips during the night, the study time span is restrained to the day hours in-between 5 am and 8 pm. After preprocessing, the trip dataset is made of 3 4736 macro-trips, with a total of 1 125 distincts *macro-paths* observed.

We begin by applying our method on the low-resolution dataset (but without downsampling and therefore time bias). The speed profiles obtained are displayed in figure 3. The first ten regions of each figures correspond to the urban inner regions, while the last one characterises the speed dynamics on Lyon's ring road. On each subplot, the reference speed is represented in blue. The orange lines characterise the results obtained from the system resolution, after applying a physical filter to remove the few highly diverging speed values that occur when the resolution does not converge. An additional moving average filter is applied next to smooth the speed profiles, in green on the plots. Those results are very satisfactory for the inner regions. Generally speaking when it comes to the ring road regions, the obtained speed profiles fit less the reference profile. However, the results are very good for most of the speed drops at the peak hours, which is extremely encouraging. This demonstrates the ability of the method to properly capture the speed trends with only limited trip information.

Next, the method is applied to the biased dataset. The very negative effect of the temporal bias introduced by the downsampling process on the speed profiles evidences that the imprecision on

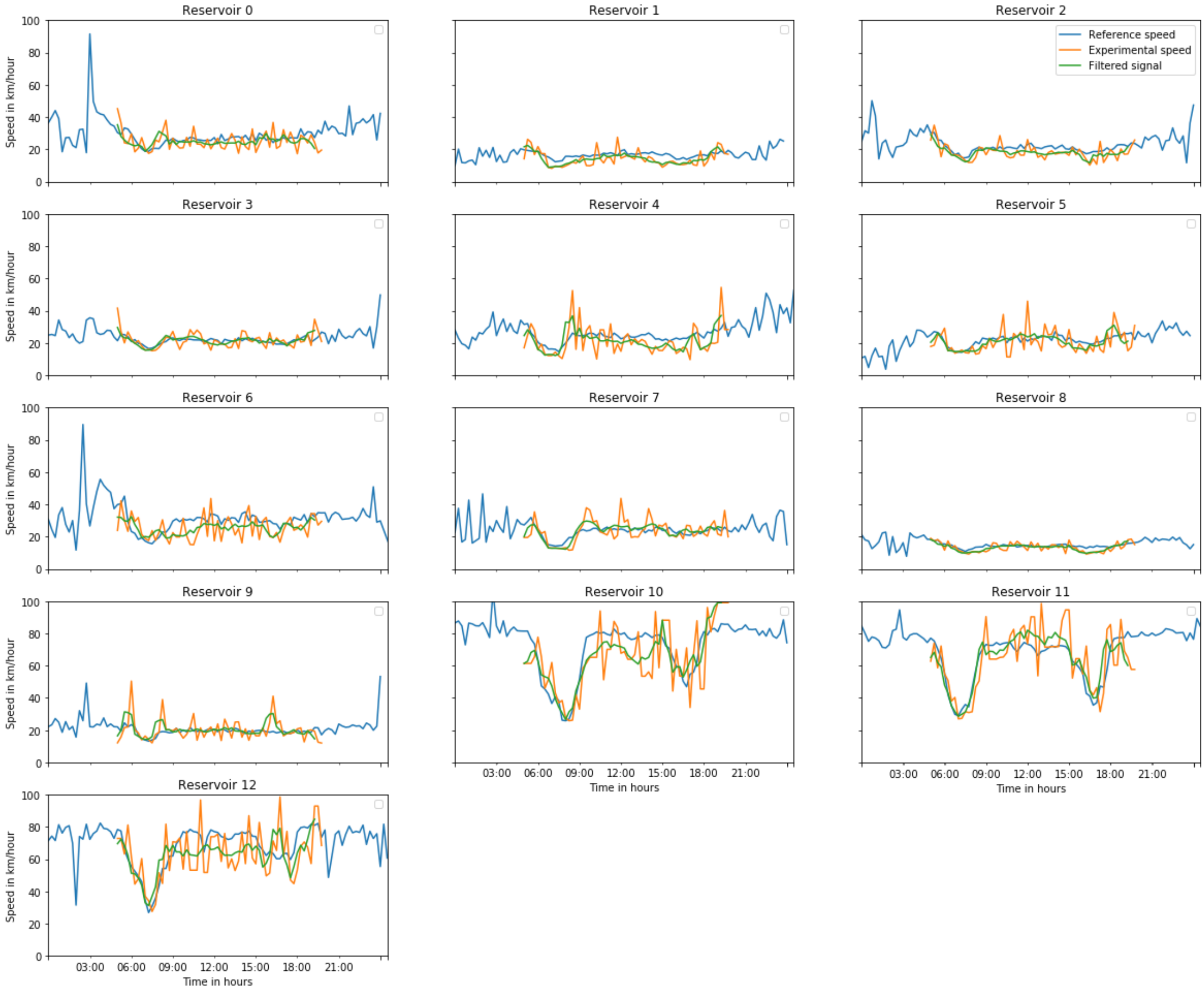


FIG. 3 – *Speed profile for low quality trip data without temporal downsampling*

travel times are too important to be neglected and confirm the necessity of a proper bias estimation. This bias estimation can be done with the comparison of the travel times between the low resolution dataset and the biased resolution dataset, to derive ϵ_t at each time period throughout the day. This can be injected in the equation system as in equation (2). From our first observations after bias estimation and removal, it appears that the result remain largely unsatisfactory. This can so far be explained by the low amount of trips for each *macro-path*, making the global average bias $\bar{\epsilon}_t$ sometimes highly different from the local average biases $\bar{\epsilon}_{P,t}$. We believe that this problem would be solved when dealing with a larger data amount and we are working on this aspect to demonstrate it.

To conclude, we propose a new methodology to estimate the dynamics of regional traffic speeds from mobile sensing data. Our method is based on the partitioning of the urban area in regions, and on the identification of groups of sampled trips sharing common *macro-paths* and arrival time period. This clustering of macro-trips provides an estimation of the travel times along each path. The construction at each time period of linear systems allows a combined analysis of the travel times and returns an estimation of the traffic speed in each region, providing that exogenous travel distance data are available. The structure of this method is particularly fitted to a CDR data input, as it requires very little temporal or itinerary information at the individual level and to takes into account the inherent temporal bias that characterises those data.

Through the application of our method to the set of GPS trips reduced to minimal temporal and path information, we could validate the global methodology. We then downsampled the trips temporal dimension in order to simulate the human uneven communication rhythms and to reproduce the temporal limits of CDR data. The direct application of the method the downsampled data showed that correcting the estimation of the travel time was a necessary condition to properly estimate the speed. Comparing the original and biased trip data specifically enabled us to estimate the temporal bias that CDR data can present compared to the ground truth GPS data. While our first observations show that removing the average bias does not permit to fully correct the travel time estimation and to obtain satisfactory speed results, we believe that this will be made possible by using a larger dataset. We are working on verifying it.

In future works, we plan to explore the sensibility of our method to the different parameters such as the size of the regions, the time period duration, or the significance threshold of the system equation. We also plan to explore the impact of an additional spatial downsampling, and develop methods for *macro-path* completion when the data is too sparse for the *macro-path* to be detected. Last but not least, our objective is eventually to apply the developed method on real CDR data.

Acknowledgements

This project has received funding from the Center on Atmospheric Pollution and Greenhouse Gases (CITEPA), Paris, France, under the Green City Big Data research and development Project.

References

[1] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta 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. doi: https://doi.org/10.1016/j.trc.2015.02.018. URL http://www.sciencedirect.com/science/article/pii/S0968090X1500073X. Big Data in Transportation and Traffic Engineering.

[2] Hillel Bar-Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel. *Transportation Research Part C: Emerging Technologies*, 15(6): 380–391, Dec 2007. ISSN 0968-090X. doi: 10.1016/j.trc.2007.06.003. URL http://dx.doi.org/10.1016/j.trc.2007.06.003.

[3] S.F.A. Batista, Ludovic Leclercq, and Nikolas Geroliminis. Estimation of regional trip length distributions for the calibration of the aggregated network traffic models. *Transportation Research Part B: Method-*

ological, 122:192 – 217, 2019. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2019.02.009. URL http://www.sciencedirect.com/science/article/pii/S0191261518311603.

[4] Guangshuo Chen, Sahar Hoteit, Aline Carneiro Viana, Marco Fiore, and Carlos Sarraute. Enriching sparse mobility information in call detail records. *Computer Communications*, 2018.

[5] Serdar Çolak, Lauren P Alexander, Bernardo G Alvim, Shomik R Mehndiratta, and Marta C. González. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transportation research record: Journal of the transportation research board*, 2526(1):126–135, 2015.

[6] Thierry Derrmann, Raphaël Frank, Francesco Viti, and T. Engel. Estimating urban road traffic states using mobile network signaling data. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7, Oct 2017. doi: 10.1109/ITSC.2017.8317718.

[7] Marta C. Gonzalez, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453:779 EP –, June 2008. URL https://doi.org/10.1038/nature06958.

[8] Md. Shahadat Iqbal, Charisma F. Choudhury, Pu Wang, and Marta C. González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63 – 74, 2014. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2014.01.002. URL http://www.sciencedirect.com/science/article/pii/S0968090X14000059.

[9] Andreas Janecek, Danilo Valerio, Karin Anna Hummel, Fabio Ricciato, and Helmut Hlavacs. The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, 2015.

[10] Jingbo Shang, Yu Zheng, Wenzhu Tong, Eric Chang, and Yong Yu. Inferring gas consumption and pollution emissions of vehicles throughout a city. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2014.

[11] Jameson L. Toole, Yves-Alexandre de Montjoye, Marta C. González, and Alex Pentland. Modeling and understanding intrinsic characteristics of human mobility. *Social Phenomena*, pages 15–35, 2015. doi: 10.1007/978-3-319-14011-7_2. URL http://dx.doi.org/10.1007/978-3-319-14011-7_2.