Transportation Research Record

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

--Manuscript Draft--

Full Title:	Dynamic Estimation of Urban Zonal Speed from Mobile Sensing Data and Macroscopic Paths
Abstract:	In the past decade, call detail records (CDR), a specific type of mobile phone data, has been proven to be an innovative source of human and urban mobility data. When compared to GPS probe data, CDR data often suffer from temporal sparsity and spatial imprecisions. However CDR data are also usually more massive and offer a larger population coverage. Given this advantage, we wonder whether CDR data could be used as a reliable source to estimate the mean traffic speed dynamics. In this article, we propose an innovative method, based on the partitioning of the urban area in reservoirs and on the analysis of basic trip features. The mean traffic speed in each reservoir is derived from the solution of a linear system that clusters individual trips per macro-path to make more robust travel time estimations. This method is fast and simple to implement in real-time but requires a prior analysis of the main trip patterns at the city and reservoirs scales. We apply this methodology on a large GPS dataset, that we reduce and downsample in order to reproduce the typical CDR data temporal characteristics. As we want to compare the method results on both GPS and CDR, monitoring the data downsampling process allows us to limit discrepancies. Moreover, the original raw GPS data also provides the ground truth reference. This experiment in a very controlled environment is a first step towards studies dealing with real CDR data.
Manuscript Classifications:	Data and Information Technology; Traffic Flow Theory and Characteristics AHB45; Planning and Forecasting; Traffic Flow Theory and Characteristics AHB45
Manuscript Number:	
Article Type:	Presentation
Order of Authors:	Manon Seppecher
	Ludovic Leclercq
	Angelo Furno
	Delphine Lejri
	Amélie-May Lupinski

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

Manon Seppecher*
Univ. Lyon, ENTPE, IFSTTAR, LICIT
UMR_T 9401, F-69518, LYON, France
CITEPA, PARIS, France
Tel: +33 4 72 04 72 95
Email: manon.seppecher@entpe.fr

Ludovic Leclercq
Univ. Lyon, ENTPE, IFSTTAR, LICIT
UMR_T 9401, F-69518, LYON, France
Tel: +33 4 72 04 77 16
Email: ludovic.leclercq@entpe.fr

Angelo Furno
Univ. Lyon, ENTPE, IFSTTAR, LICIT
UMR_T 9401, F-69518, LYON, France
Tel: +33 4 72 04 77 16
Email: angelo.furno@ifsttar.fr

Delphine Lejri
Univ. Lyon, ENTPE, IFSTTAR, LICIT
UMR_T 9401, F-69518, LYON, France
Tel: +33 4 72 04 77 16
Email: delphine.lejri@entpe.fr

Amelie-May Lupinski
Univ. Lyon, ENTPE, IFSTTAR, LICIT
UMR_T 9401, F-69518, LYON, France
Tel: +33 4 72 04 77 16
Email: amelie-may.lupinksi@entpe.fr

* Corresponding author

Paper submitted for presentation at the 99th Annual Meeting Transportation Research Board, Washington D.C., January 2020

Special call by AHB45 committee on “Improved flow modeling and advanced control strategies for large-scale urban traffic networks”

Word count: 4995 words + 1 table(s) × 250 = 5245 words

August 1, 2019

ABSTRACT

In the past decade, call detail records (CDR), a specific type of mobile phone data, has been proven to be an innovative source of human and urban mobility data. When compared to GPS probe data, CDR data often suffer from temporal sparsity and spatial imprecisions. However CDR data are also usually more massive and offer a larger population coverage. Given this advantage, we wonder whether CDR data could be used as a reliable source to estimate the mean traffic speed dynamics.

In this article, we propose an innovative method, based on the partitioning of the urban area in reservoirs and on the analysis of basic trip features. The mean traffic speed in each reservoir is derived from the solution of a linear system that clusters individual trips per macro-path to make more robust travel time estimations. This method is fast and simple to implement in real-time but requires a prior analysis of the main trip patterns at the city and reservoirs scales.

We apply this methodology on a large GPS dataset, that we reduce and downsample in order to reproduce the typical CDR data temporal characteristics. As we want to compare the method results on both GPS and CDR, monitoring the data downsampling process allows us to limit discrepancies. Moreover, the original raw GPS data also provides the ground truth reference. This experiment in a very controlled environment is a first step towards studies dealing with real CDR data.

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

1. INTRODUCTION

In the past few years, mobile phone data has been proven to be an innovative, accessible and very rich source of information about human mobility. Those data, passively generated by mobile phone users while communicating, moving around the network, or activating and deactivating their devices, are collected by the mobile phone data provider either for billing or network management purposes. The resulting massive databases form an incredibly large source of information about communication activities, but also mobility behaviours of the urban populations. In this paper, we wonder if a specific type of mobile phone data, call detail records, could be used in order to estimate the mean traffic speed dynamics at a zonal scale in an urban area.

Call detail records (CDR) register for each communication event (calls, messages or data browsing) the following information: the user's unique identification key, a timestamp characterising the start of the event, and the location of the base station antenna that processed the event. This basic structure, relating together user id, time and location, confers to those data a strong potential for mobility analysis. Compared to GPS data classically used in mobility and traffic analysis, CDR data present several advantages. The mobile phone penetration rates are generally high. For traffic studies, phone data usually provides a better spatial coverage as they are not restricted to a single users category, unlike taxis GPS probe data. However, CDR data also have drawbacks. First, as the positional information is obtained at the antennas level, the spatial precision depends on the base station network. Second, the data generation depends on the users' communication activities and behaviours, making the temporal acquisition of data uneven and sparse it time; in particular users with little communication activity will generate fewer location data and their mobility will be especially difficult to estimate.

Using CDR data, Gonzalez et al. (1) analysed the behavioural rules intrinsic to human mobility in order to construct realistic individual mobility models. At a more aggregated scale, CDR data were also used to estimate origin destination matrices (2, 3, 4), proving that they are an interesting alternative to the traditional and costly transportation surveys and census data at an urban scale. In Toole et al. (5) the origin-destination matrices estimated thanks to the aforementioned methods were later assigned onto the road network in order to estimate the traffic load and to identify the specific origin-destination flows that caused the highest share of it. Two other types of mobile phone data, handovers and location area updates data, have been specifically exploited for traffic analysis studies. Handovers data record the network-centred events that happen when the communication connection of a user engaged in a call session is transferred from one antenna to another due the user's movement on the network. Location area updates data record for every device (including idle ones) the transfer of the communication connections from one identified group of antennas to another one when the devices owners move over the network. Those data are less dependant to communication activities than CDR data and have a good spatial coverage, however they are not always accessible in practice, mainly for privacy reasons. Derrmann et al. (6) explored the potential of handovers data for the estimation of Macroscopic Fundamental Diagrams. On highway segments, Bar-Gera (7) used handovers data to estimate the traffic speed while Janecek et al. (8) analysed the travel time using both handovers and location area updates.

But, despite this extensive literature in the field of mobility and traffic, the specific question of speed estimation from CDR data appears to have been barely studied. This is probably partly due to the imprecisions of CDR data, which can seem *a priori* insufficient for traffic speed estimation compared to GPS or handovers and location area updates data. Aside from the challenge that speed estimation from CDR data represents, or from the insight it can give on the traffic dynamics over time 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. CDR data are accessible, massive, and representative of large population groups unlike most GPS probe datasets derived from taxi trajectories. While using GPS data for those estimation requires important complementary data and costly scaling up processes (9), an estimation using CDR data can be an efficient and light alternative. In that direction, Li et al. (10) introduced a method to estimate traffic related pollutant emissions from CDR data at a regular zonal scale. However the method used for estimating the emission factors (essential for the emission calculation and which depends on the traffic speed) was not made explicit. Consequently, and to the best of our knowledge, the estimation of the traffic speed dynamics from CDR data still needs to be investigated. This is what we propose to do in this article.

We aim to determine if CDR data can be used in order to estimate dynamically the traffic speed at a zonal scale in an urban area. As a first step into that direction, we propose an innovative method based on the partitioning of the urban area in reservoirs and on the identification of clusters of individual trips sharing similar simple features such as macro-path and arrival time period. Providing that a good estimation of the trip lengths at the city and reservoirs scale is known, the clusters allow to build at each studied time period a simple linear system, that, once solved, returns an estimation of the traffic speed. We chose the city of Lyon, France, as our case study and experiment our method on a GPS dataset that we downsample in order to get the same statistical characteristics than CDR data, following a method initially developed by Chen et al. (11) to estimate the spatial biases introduced by CDR data in mobility studies. As the GPS data use does not come from taxi compagnies, it does not present the specific bias mentioned above. This approach first allows to validate the general method with unbiased data, and then to evaluate the impact of the bias introduced by the downsampling and attempt to limit this impact. Eventually, the objective is to make this estimation method robust enough to be used on real CDR data rather than simulated ones.

This article is organised as follows : Section 2 exposes the principle of our approach and describes the proposed methodology. Section 3 presents our case study, as well as the exploited data. Section 4 focuses on the results we reached. Finally, Section 5 concludes with the achieved results, the limitations of this work and the on-going perspectives.

2. METHODOLOGY

Work scope and definitions

We propose a method to estimate the temporal dynamics of the mean traffic speed at zonal scale, using a limited and selected set of individual trips characterised by a low level of information. Our method relies on partitioning an urban network into a set of regions, called reservoirs, and on dividing the time dimension into a set of time intervals of equal duration. Our objective is to estimate the mean traffic speed in each reservoir for every time period.

The notions on which the methodology relies are defined below.

Reservoirs: We call reservoirs the result of the partitioning of the studied urban area into smaller regions of equivalent sizes and homogeneous characteristics in term of city fabric, demography, road network or traffic dynamics.

Micro-path: We call micro-path the ordered sequence of road segments traveled by a sampled vehicle along a trip.

Macro-path We call macro-path the ordered sequence of reservoirs crossed by a sampled vehicle along a trip. A macro-path is a scaled representation of a micro-path at the reservoirs level.

Micro-trip: We call a micro-trip the representation of a sampled individual vehicle trip according to the following features : ***(trip Id, vehicle Id, micro-path, exact arrival time, total travel time)***

Macro-trip: We call macro-trip the representation of a sampled individual vehicle trip according to the following features : ***(trip Id, vehicle Id, macro-path, arrival interval, total travel time)***. A macro-trip is a simplified representation of a micro-trip at the reservoirs scale and with a coarser time resolution.

The simplification process that reduces a micro-trip to its macroscopic representation allows to group similar (but not identical) micro-trips into clusters based on both the macro-path and the arrival interval. The methodology we propose in the next part relies on those macro-trip categories.

Linear System Construction

Generic System

Let us consider that the studied area is divided in a given number R of reservoirs. Let us also consider that the time span of the experiment is divided into a set of equal time intervals. Let I_t (of size n_t) be the set of macro-trips reaching their destinations during period t, and let $I_{t,P}$ (of size $n_{P,t} \leq n_t$) be the subset of I_t of macro-trips that match macro-paths P.

For a macro-trip i of $I_{t,P}$, its traveled time along P can be written as the sum of the traveled times $T^i_{r,P}$ of i in each reservoirs r of P:

$$T^i_P = \sum_{r\in P} T^i_{r,P} \tag{1}$$

We assume that the traffic speed in reservoir r during interval t $V_{r,t}$ is constant and homogeneous. This gives:

$$T^i_P = \sum_{r\in P} \frac{L^i_{r,P}}{V_{r,t}} \tag{2}$$

with $L^i_{r,P}$ the total distance traveled in reservoir r by instance i along the macro-path P.

Summing on the $n_{P,t}$ equations 2 characterising to the $n_{P,t}$ macro-trips of $I_{t,P}$, we get:

$$\begin{aligned}
\sum_{i=1}^{n_{P,t}} T^i_P &= \sum_{i=1}^{n_{P,t}} \sum_{r\in P} \frac{L^i_{r,P}}{V_{r,t}} && (3)\\
&= \sum_{r\in P} \sum_{i=1}^{n_{P,t}} \frac{L^i_{r,P}}{V_{r,t}} && (4)\\
n_{P,t}\bar{T}_{P,t} &= \sum_{r\in P} n_{P,t} \frac{\bar{L}_{r,P,t}}{V_{r,t}} && (5)\\
\bar{T}_{P,t} &= \sum_{r\in P} \frac{\bar{L}_{r,P,t}}{V_{r,t}} && (6)\\
& && (7)
\end{aligned}$$

With $Y_{r,t} = \frac{1}{V_{r,t}}$, and assuming that $\bar{L}_{r,P,t}$ is independent of t , we get:

$$\bar{T}_{P,t} = \sum_{r \in P} \bar{L}_{r,P} Y_{r,t} \tag{8}$$

For the time interval t, this equation characterises one given macro-path. Applying it to the whole set of macro-paths observed at interval t, we can construct the following system S_t:

$$S_t = \left\{ \bar{T}_{P,t} = \textstyle\sum_{r \in P} \bar{L}_{r,P} Y_{r,t} \qquad \forall P \right. \tag{9}$$

S_t is linear with R variables made of as many equations as macro-paths observed during interval t. This system is usually overdetermined as the number of possible macro-paths is higher than the number of reservoirs. Assuming that $\bar{L}_{r,P}$ can be estimated exogenously, and as $\bar{T}_{P,t}$ can be derived from the macro-trip information, the system can be solved using for exemple a least square optimisation method. The inverse of the solution $Y_{r,t}$ will correspond to the traffic speed $V_{r,t}$ for each reservoir r.

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 departure and the arrival 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 reservoir $\bar{L}_{r,P}$. The spatial definition of the reservoirs will also be crucial as it may drive completely different patterns for different macro-paths inside the same reservoir. This topic is still under research.

Parameter reliability and temporal bias

Theoretically, GPS data are precise enough to provide accurate estimations of $\bar{T}_{P,t}$ (corresponding to the mean traffic time along P when reaching destination during t) and $\bar{L}_{r,P}$ (the total traveled distance in reservoirs r along path P). However, when it comes to using CDR data (or simply sparser GPS data) the direct estimation of those parameters might not be as reliable. Indeed, CDR data include both temporal and spatial biases, due to users potentially large inactivity times and weak spatial resolution of base stations network. Using such data to estimate the system parameters implies reproducing those biases on the parameters and therefore twisting the system and its results.

Throughout the study, we make the assumption that $\bar{L}_{r,P}$ can be estimated exogenously thanks to another reliable method (probe vehicles, surveys, automatic network analysis, etc) and therefore do not consider the potential impact of distance bias on the data. We rather focus on the impact of the temporal bias introduced by the low acquisition rate of mobile phone communication events. 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, detecting their movements and therefore estimating the correct travel time will be much more difficult. This means that the estimated traveled time of a trip observed from CDR data is actually higher than the real one.

Consequently, for a macro-trip i of $I_{P,t}$ obtained with CDR data, we have:

$$T^i_{P,obs} = T^i_P + \varepsilon^i \tag{10}$$

where T^i_P is the real travel time of trip i, $T^i_{P,obs}$ is the observed traveled due to data imprecisions and ε^i is the temporal bias that exists in-between. Summing over $I_{P,t}$, this gives:

$$\bar{T}_{P,obs} = \bar{T}_P + \bar{\varepsilon}_t \tag{11}$$

Reinjecting $\bar{T}_P$ in (8), we get :

$$S'_t = \left\{ \bar{T}_{P,obs} - \bar{\varepsilon}_t = \sum_{r \in P} \bar{L}_{r,P} Y_{r,t} \qquad \forall P \right. \tag{12}$$

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 estimation the travel time along each macro-path P, and therefore to correct the system.

Macro-paths and equations selection

In order to reduce the system size and to limit the incompatibility risks between equations, we propose to filter out of the system the equations that are the least reliable, that is to say that have the least significance among the macro-trips. This significance can indeed be very uneven from one macro-path to another. While some macro-paths are observed for many macro-trips, some others might be representative of only one individual. In this later case, the spatial and temporal mean values will represent a single trip and so will be weakly reliable. Thus, we implement a significance threshold corresponding to the minimum number of times that a macro-path should be observed during period t in order for the corresponding equation to be selected in the system S_t. This threshold is set to 5 trips in this study.

3. CASE STUDY AND DATA DESCRIPTION

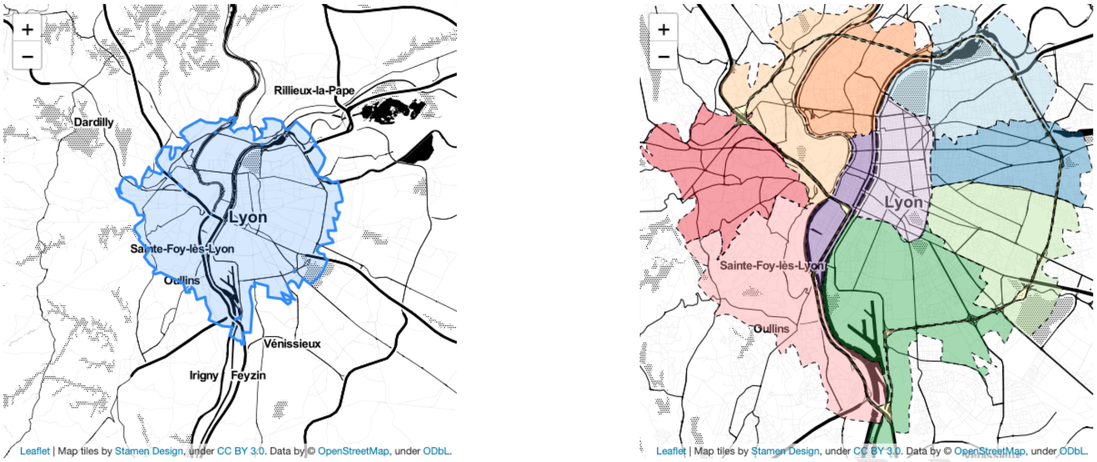
Case study

We select the city of Lyon, France, as our case study. The study area includes both Lyon and the next-by municipality of Villeurbanne, which is located inside Lyon's ring road. It is displayed in figure 1a. We have parted this territory into a number of 11 distinct reservoirs. 10 of them divide the inner city, while the last one corresponds to the ring road. Those reservoirs are displayed in figure 1b. For the inner reservoirs, 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 reservoirs is made so that it is consistent with Lyon's road network and its traffic characteristics.

Working data

Input data

In this study, we intentionally work with a GPS dataset, which we progressively simplified and downsampled in order to recreate the temporal biases of mobile phone data. This approach presents two advantages. First, it allows us to use the raw GPS data as ground truth for the estimation of the traffic speed. Second, it enables us to progressively reduce the data quality, and thus to understand how the jamming process impacts the quality of our results.


Rillieux-la-Pape
Dardilly
Lyon
Sainte-Foy-lès-Lyon
Oullins
Vénissieux
Irigny
Feyzin
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under ODbL.

(a) Considered area in Lyon, France

(b) Spatial partitioning of the studied area in 11 reservoirs, with the ring road as one of them

FIGURE 1 : Visualisation of the studied area and reservoirs

The data used 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. The data was provided by a European navigation system provider, and the traces come from vehicles equipped with their navigation system technology. This means that the dataset does not present the bias observed with GPS traces from taxis. Moreover, as each trace corresponds to a vehicle, there is no need to filter out pedestrian or cyclist travellers as usually in CDR-based studies. Though this will make the speed estimation process easier, it also means that the data we simulate do not exactly reproduce CDR data that gather users making no distinction between their means of transport.

The data structure is slightly different from typical GPS datasets, in that the GPS sampled locations of the monitored vehicles have already been map-matched by the data provider onto the road network. Thus, each GPS trace is made of an ordered sequence of traveled network links rather than a succession of geolocated points. The original data have the following structure : each monitored *vehicle* is associated to the different *trips* that it performed, and each *trip* corresponds to a sequence of *observations*. An *observation* relates each link of the trip to the considered vehicle, the timestamp of the entrance of the vehicle on this link, the speed of the vehicle on this link as well as the distance coverage that the vehicle does of this link (the first and last link of the trajectory might not be fully traveled, for exemple because of parking spots). This information will especially allow us to reconstruct both the travel distance and the travel time on each link.

As the data coverage is actually larger than the study area defined above, we filter out the data tracks that do not enter the inner Lyon area, and split into smaller trips the ones that go beyond those limits. Additionally, we further preprocess the data by filtering out any aberrant trace or static vehicle in order to obtain a clean and reliable data set.

Ground truth data

First, the raw data are used for computing the ground truth reservoir speed values throughout time. For each time interval t, and each macro-path P, the ground truth speed $V_{r,t}$ is calculated as :

$$V_{r,t} = \frac{TTD_{r,t}}{TTT_{r,t}} \tag{13}$$

where $TTD_{r,t}$ and $TTT_{r,t}$ are respectively the total traveled distance and time by the sampled users in reservoir r during interval t. This information can easily be derived from the *observations* dataset.

Distance parameters
To estimate the travel distance $\bar{L}_{r,P}$ we resort here to the full resolution of GPS data. In practice, this estimation is done thanks to a few days of data and supposed known thereafter. It will be used as a parameter in the systems constructions. On the other hand, the speed estimation is applied over another selection of days where only CDR-like data are supposed to be available.

Spatiotemporal aggregation : from trips to macro-trips
Reshaping the original preprocessed data, we build a T_0 dataset of micro-trips as defined in section 2, i.e. using the five following features: *trip id*, *vehicle id*, *micro-path*, *exact arrival time*, and *total traveled time*. From this initial dataset T_0, we are going to derive the two degraded datasets used for our experiments.

The first of those datasets corresponds to the macro-trips data set. First, we apply the spatial aggregation which relies on the city partitioning proposed above. This partition of the studied area allows us to relate each of the road network links to the reservoir it is included in. Subsequently, each vehicle *micro-path*, i.e. the succession of network links making up the trip itinerary, can be summarised as the ordered sequence of reservoirs that the vehicle crossed along its way.We call this ordered sequence of crossed reservoirs *macro-path*, in opposition to the micro-paths made of ordered sequence of links. Then, for each trip, we reduce the exact arrival time to the time interval it belongs it. In this study, we select time intervals of 15 minutes. Thus, an arrival time at *11:22:36* will be transformed into *11:15:00*.

This aggregating process allows us to transform trips into macro-trips of the shape *(trip Id, vehicle Id, macro-path, arrival interval, total travel time)*. We will call this new dataset T_1. For each vehicle trip, the exact arrival time is replaced by the closest anterior quarter of hour. In that way we part the day in a predefined grid of 96 time intervals. We can now identify similar vehicle trips, i.e vehicle trips that travel along the same macro-path and end in the same time interval. So far, we have erased the knowledge about the exact travelled path and exact arrival time. However, it is important to stress here that this data degradation has not impacted yet the information about total traveled time. It simply allows to categorise the different trajectories and to relate them together, in anticipation of the construction of the linear systems.

Temporal downsampling
In order to evaluate the potentials of our speed estimation method with mobile phone data, we apply on the dataset T_1 a downsampling process which 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. (11) 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 downsample

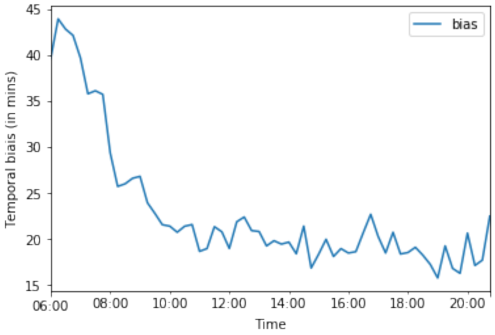


FIGURE 2 : Temporal bias evolution throughout the day

introduce a bias both on the trip start time and on the trip arrival time, and thus implies an increase of the global traveled time. The detected arrival interval can also change depending on how big was the resulting time gap at the end of the trip. This new resulting dataset will be named T_2.

In reality, using CDR data 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.

4. RESULTS AND DISCUSSION

In this section, we present the results of the application of our method to a specific day, 2018 February 5.

Bias characterisation

Using both T_1 and T_2, we can compare for each macro-trip the observed travel time (biased) to the real travel time. The difference characterises the temporal bias at an individual level. Measuring this bias for every macro-trip and averaging their values for every time period, we can estimate the mean bias ε_t throughout the day. The result of this estimation with our data is displayed in figure 2. In the early morning the mean temporal bias is quite high, up to 45 minutes of difference between the real travel time and the observed one. This average bias strongly drops between 6 am and 10 am before stabilising around 20 minutes. This difference is due to the changes around that time in the communication activity rates of people. This is understandable as this time bias depends itself on the human communication rhythms, much sparser at night. Thanks to this bias characterisation, it becomes possible to understand the impact of the CDR data temporal characteristics on the estimation of the travel time.

Method application to macro-trip data

As the macro-trip data contains only a few trips during the night, the study time span is restrained to the day hours in-between 6 am and 9 pm. Taking this restriction into account, our macro-trip dataset T_1 is made of 69 520 individual macro-trips, with a total of 1 862 distincts macro-paths observed. Out of those macro-paths, only 82 of them are observed more than 5 times (our significance threshold) during at least one time period, and therefore will be represented by system equations.

The results of the method application on dataset T_1 are displayed in figure 3. The first ten reservoirs of each figures correspond to the urban inner reservoirs, while the last one characterises

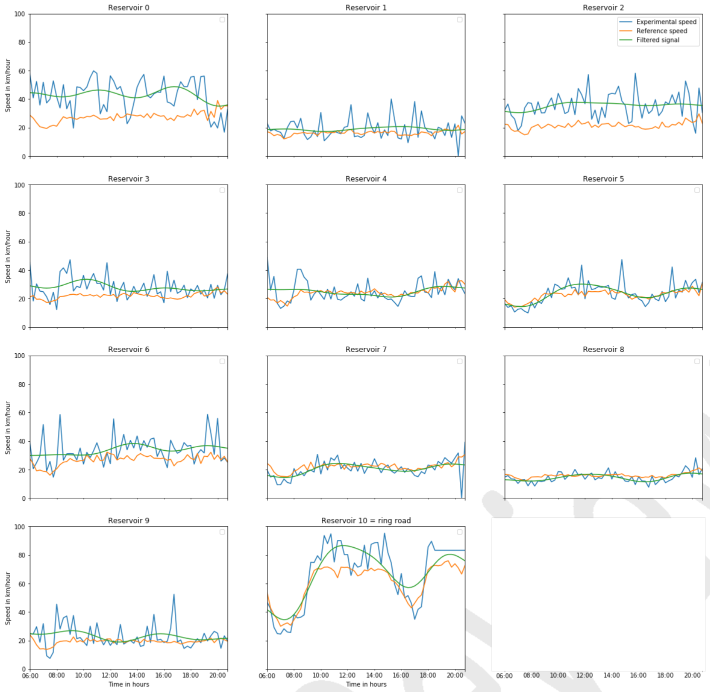


FIGURE 3 : Results of methodology applied to dataset T_1 of macro-trips

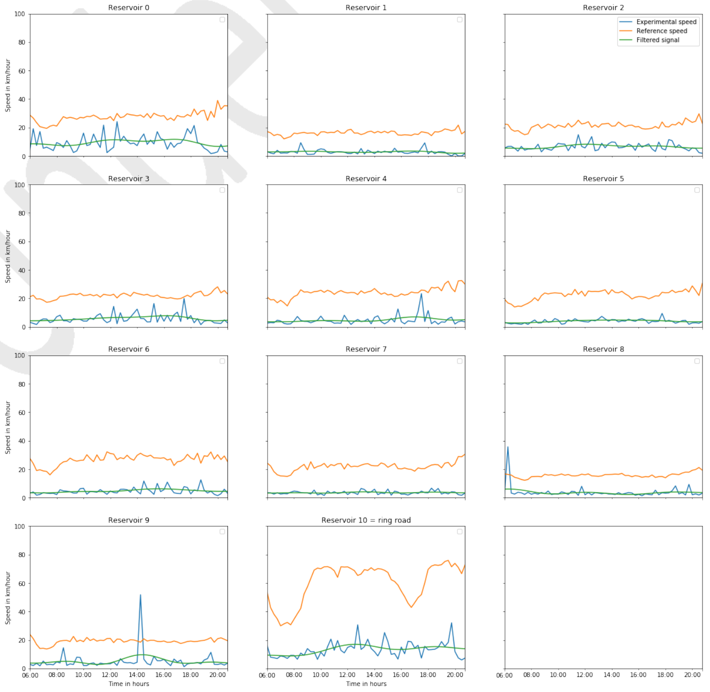


FIGURE 4 : Results of methodology applied to dataset T_2 of macro-trips without bias correction

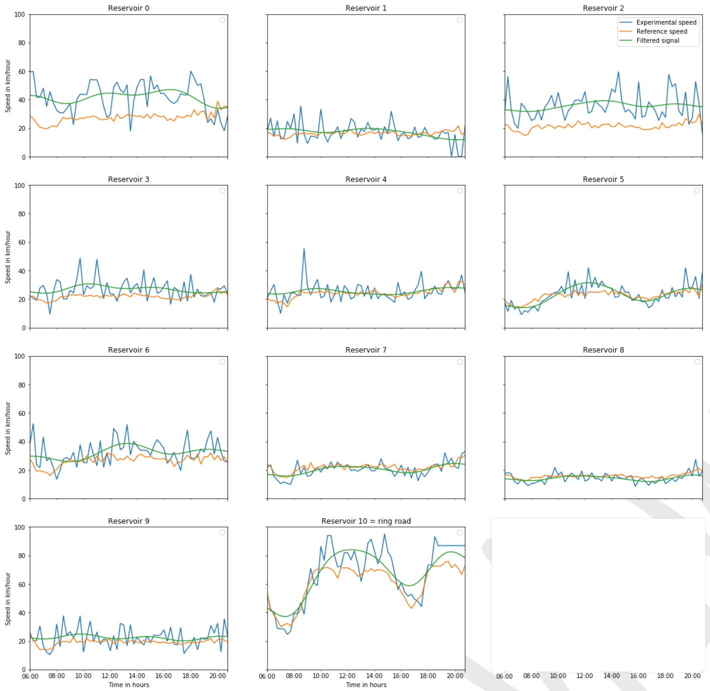


FIGURE 5 : Results of methodology applied to dataset T_2 of macro-trips with bias correction

the speed dynamics on Lyon's ring road. On each subplot, the ground truth traffic speed is represented in orange. The blue curves characterises the results obtained from the system resolution. As it happens that the resolution does not converge, the returned speed signal can return aberrant values. We apply a physical filter to remove the few diverging speed values. A speed of 60 km/hour is selected as the threshold for detecting aberrant values in every reservoir except on the ring road where the limit is raised up to 90km/hour. Values above those thresholds are arbitrarily replaced by the previous consistent one. This explains the step that we can particularly observe for the reservoir 10 during the evening. This physical filtering removed the aberrant speed estimations. However the obtained signals still present high volatility. In order to smooth the results, we apply a second filter represented in green on the plots. Assuming that the signal noise corresponds to higher speed peaks, we apply a low-pass frequency filter. The Butterworth filter is selected and applied with a forward-backward filtering method (once in each direction).

From figure 3 we observe that the system generally renders correctly the speed dynamics. This is especially true for reservoirs 5, 7, 8, 9 and 10. We also notice an overestimation of the traffic speed in particular for the reservoirs 0 and 2, which we do not explain yet.

In figure 4 and figure 5 we display the results of the method applied on dataset T_2 respectively with and without the correction of the temporal bias. Figure 5 highlights the very negative effect of the temporal bias introduced by the downsampling process on the results. This evidences that the temporal bias induced by human uneven and large communication times makes CDR data unusable as they are to estimate the traffic speed, even at a regional level. After correcting the system however, we can observe in figure 5 that the estimation tend to better fit the reference curve again.

Table 1 summarises the achieved results, by comparing the root mean square error over the full day in each reservoir, in-between the filtered results (green curves) and the reference values.

The results obtained for T_1 and de-skewed T_2 are sensibly similar. For some reservoirs, the results obtained with T_2 after bias correction are slightly better than with T_1. Those results should be further explored in order to explain this slight improvement, which can maybe come from the biased arrival time used in T_2 for the system resolution.

Reservoirs	Original macro-trips (T_1)	T_2 without bias correction	T_2 with bias correction
0	16.58	18.56	15.16
1	3.38	13.53	3.58
2	14.25	14.99	14.46
3	6.81	16.30	5.53
4	3.66	19.66	2.91
5	3.11	18.70	3.58
6	7.58	22.17	6.57
7	2.89	18.58	2.68
8	2.16	12.40	1.97
9	4.89	14.35	3.66
10	9.04	48.55	9.06

TABLE 1 : Synthesis of the RMSE for each reservoir according to the dataset used

5. CONCLUSION

In this article we have proposed 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 reservoirs, and on the identification of groups of sampled trips sharing common macro-paths and arrival time period. This clustering of macro-trips provides a robust estimation of the travel times along each path. Cross-referencing the macro-path travel time estimations within a linear system allows us to estimate the traffic speed dynamics, 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.

Thanks to 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. Removing the average bias permitted to correct the travel time estimation and to obtain satisfactory speed results compared to the ground truth data.

In our future works, we plan to explore the sensibility of our method the different parameters such as the size of the reservoirs, the time period duration, or the significance threshold of the system equation. About the specific question of the reservoir size, it is clear that the larger the reservoirs are, the more likely they are to encompass very different micro-paths under a same macro-path. In such cases, it might be interesting to divide a macro-path into several sub-macro-paths, analysing the potential modes in travel distance and time distributions. 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] Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabási, Understanding individual human mobility patterns. *Nature*, Vol. 453, 2008, pp. 779 EP –.

[2] Iqbal, M. S., 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*, Vol. 40, 2014, pp. 63 – 74.

[3] Çolak, S., L. P. Alexander, B. G. Alvim, S. R. Mehndiratta, and M. C. González, Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transportation research record: Journal of the transportation research board*, , No. 2526, 2015, pp. 126–135.

[4] Alexander, L., 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*, Vol. 58, 2015, pp. 240 – 250, big Data in Transportation and Traffic Engineering.

[5] Toole, J. L., S. Çolak, B. Sturt, L. P. Alexander, A. Evsukoff, and M. C. González, The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 162 – 177, big Data in Transportation and Traffic Engineering.

[6] Derrmann, T., R. Frank, F. 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)*, 2017, pp. 1–7.

[7] Bar-Gera, H., 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*, Vol. 15, No. 6, 2007, pp. 380–391.

[8] Janecek, A., D. Valerio, K. A. Hummel, F. Ricciato, and H. Hlavacs, The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, 2015.

[9] Shang, J., Y. Zheng, W. Tong, E. Chang, and Y. Yu, Inferring Gas Consumption and Pollution Emissions of Vehicles throughout a City, 2014.

[10] Li, Q., Y. Cheng, F. Ding, X. Wan, and B. Ran, Citywide Hourly Traffic Emissions Estimation Using Cellular Activity Data. In *TRB 95th Annual Meeting Compendium of Papers*, 2016.

[11] Chen, G., S. Hoteit, A. Carneiro Viana, M. Fiore, and C. Sarraute, Enriching sparse mobility information in Call Detail Records. *Computer Communications*, 2018.