



HAL
open science

Estimation of urban zonal speed dynamics from user-activity-dependent positioning data and regional paths

Manon Seppecher, Ludovic Leclercq, Angelo Furno, Delphine Lejri, Thamara Vieira da Rocha

► **To cite this version:**

Manon Seppecher, Ludovic Leclercq, Angelo Furno, Delphine Lejri, Thamara Vieira da Rocha. Estimation of urban zonal speed dynamics from user-activity-dependent positioning data and regional paths. *Transportation research. Part C, Emerging technologies*, 2021, 129, pp1-28. 10.1016/j.trc.2021.103183 . hal-03461907

HAL Id: hal-03461907

<https://hal.science/hal-03461907>

Submitted on 1 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Estimation of Urban Zonal Speed Dynamics from **User-Activity-Dependent Positioning** Data and Regional Paths

Manon Seppecher^{a,b,*}, Ludovic Leclercq^{b,**}, Angelo Furno^b, Delphine Lejri^b, Thamara Vieira da Rocha^a

^aCITEPA, 42 rue de Paradis, 75010 Paris, France

^bUniv. Gustave Eiffel, Univ. Lyon, ENTPE, LICIT, F-69518, Lyon, France

Abstract

Over the past few decades, **the digitalization of services and infrastructures has led to the emergence of a broad set of new information sources to characterize human mobility.** These sources usually offer valuable significant population penetration rates but may also suffer from important temporal sparsity. **Data generated by user activity, such as social networks or mobile phone data, especially fit this description.** Although this temporal sparsity might prevent estimating individual travel speeds, we state that **such low-frequency positioning** data enable estimating the average urban traffic speed dynamics when considering an adequate network partitioning. In this sense, this article proposes a new method, based on the division of the urban area **of a given city** into regions and on the analysis of a limited set of basic characteristics of individual vehicle trips, such as the regional path. Our solution first involves estimating robust travel times from travelers sharing similar trip features and then jointly analyzing these travel times to deduce the underlying regional traffic speeds, using regression analysis. We apply this methodology on a set of trips **derived** from a large GPS dataset of vehicle tracks covering the city of Lyon. **These data are purposely downsampled to reduce the sampling rate and reproduce bias and temporal features that are proper to sparser but larger-scale, mobility data sources dependent on user's communication activities.** Controlling the data downsampling process allows us to evaluate the impacts of the progressive information loss on the speed estimation, while the raw GPS data provide the ground truth speed reference against which to compare our results. Provided that the amount of observed individual trips is sufficient, the analysis returns satisfying speed estimation results, both at low and high downsampling levels. Thus, we successfully demonstrate that it is possible to estimate zonal traffic speeds from degraded trip data without evaluating individual travel speeds.

Keywords: traffic speed estimation, travel time estimation, regional paths, regions, bias model, mobile phone data, CDR data, GPS data, LBNS data

1 Highlights

- 2 • **We propose a framework for estimating traffic speeds from temporally-biased trips.**
- 3 • **We propose a temporal bias model linked to user-activity-dependent positioning data.**
- 4 • **We apply the methodology to a set of GPS trips downsampled according to this model.**
- 5 • **We investigate the effect of successive data degradation on the results.**
- 6 • **We investigate the effect of statistical representativity on the results.**

*Corresponding author. Tel.: +33 (0) 4 72 04 72 95, manon.seppecher@univ-eiffel.fr

**Corresponding author. Tel.: +33 (0) 4 72 04 77 16, ludovic.leclercq@univ-eiffel.fr

7 1. Introduction

8 Over the last two decades, the digitalization of services and infrastructures has led to the emergence of a broad set
9 of new information sources to characterize human mobility. In particular, GPS tracks from navigation systems and
10 services have become prevalent (Castro et al., 2013; Lin and Hsu, 2014). The exploitation of other sources, such as
11 anonymous geolocalized social media logs (Twitter, Foursquare) and cell phone data, has also become increasingly
12 popular (Chen et al., 2016). The collected geolocated tracks may vary significantly in both the spatial and the temporal
13 resolution depending on the technology used to generate the data (GPS, mobile telephony, wireless networks), the
14 sensing device or service (on-board or mobile navigation systems, geolocation through social networks and location-
15 based services, 2G, 3G 4G cellular networks), as well as the level of user activity (Asgari et al., 2013; Toch et al.,
16 2018).

17 GPS tracks derived from vehicular and mobile navigation systems are usually quite accurate both in space and
18 time. The navigation system generally acquires the user's position at a regular frequency, which usually ranges from
19 a very high frequency (*e.g.*, every second) to lower sampling rates (in the order of the minute). Despite these possible
20 variations and acquisition noise and errors, GPS navigation systems remain a key source to explore individual and
21 aggregated mobility patterns and monitor traffic (Castro et al., 2013; Lin and Hsu, 2014). However, the related data
22 sets often suffer from limited penetration rates.

23 Another source of information on human mobility can be found in social networks and Location-Based (Network-
24 ing) Services (LBS - LBNS). For instance, the Twitter social network allows users to share their geo-location with
25 their tweets, while the LBNS Swarm (formally called Foursquare) offers its users to "check-in" in various venues and
26 share this information with friends. GPS being the technology on which those networks and services rely, the spatial
27 accuracy of the data generated using such services is mostly the same as that of navigation systems. However, the
28 main difference with the latter lies in the data generation process. Instead of being automatic and regular, the avail-
29 ability of geolocated samples with social networks and location-based services depends on the user's communication
30 and sharing behaviors. In particular, users with little posting and check-in activities will generate fewer location data,
31 and their mobility becomes harder to estimate.

32 This is a characteristic that social networks and location-based services data share with several types of passive
33 mobile phone data, such as Call Detail Records (CDR) and network signaling data (which, in addition to calls and
34 texting events, include network control ones such as handovers). These data are passively generated by mobile phone
35 users while communicating and are collected and stored by communication data providers for billing or network
36 management purposes. CDR data register each communication event (*i.e.*, a call, message, or data browsing event
37 emitted or received by a cellular device) at the base station scale (*i.e.*, antenna), while handovers register each base
38 station involved in a call. Thus, the less a user communicates, the fewer data will be generated. Barabási (2005),
39 and later Candia et al. (2008); Gonzalez et al. (2008); Calabrese et al. (2011) and Chen et al. (2018), explored the
40 existence of patterns in mobile communication behaviors and observed that the latter are bursty. While most of
41 the users' communication events happen within short time intervals, some significant time gaps also exist between
42 successive dense communication sequences. Interestingly, Gandica et al. (2017) demonstrated that message posting on
43 Twitter presents similar temporal characteristics. Those results suggest that these user-activity-dependent positioning
44 data (UADP data further on) may be more fitted to identify and analyze the static phases (often called *stays*) of users'
45 routines than to characterize the trips in-between (Ranjan et al., 2012; Hoteit et al., 2017).

46 An extensive literature exists on the use of user-activity-dependent positioning data for mobility analysis (see
47 Blondel et al. (2015); Naboulsi et al. (2016) on mobile phone data), but it mainly focuses on the characterization of
48 mobility patterns rather than the analysis of dynamic traffic features. This literature is often based on methods to detect
49 and process communication events that take place during periods of human immobility (*e.g.*, see Jiang et al. (2013);
50 Toole et al. (2015)), which allow inferring origin and destination locations of trips, for instance. On the basis of such
51 methods, the subjects covered by the literature vary from the exploration of mobility habits and characteristics (Jurdak
52 et al. (2015) with Twitter data) and the development of realistic mobility choice models (Gonzalez et al. (2008) based
53 on CDR data) to the construction of origin-destination matrices (see Osorio-Arjona and García-Palomares (2019) with
54 Twitter data, Iqbal et al. (2014); Çolak et al. (2015); Alexander et al. (2015) with CDR data) and their use as a proxy
55 for the traditionally costly transportation surveys. However, when it comes to describing the trips themselves, the
56 irregularity of the communication behaviors and the individual data generation may result in little to no positional
57 information during trips that is therefore much harder to exploit. Even if some data are collected during a trip and can

58 help to identify the likely traveled routes, as shown in Jiang et al. (2013), this situation is far from being systematic and
59 only concerns few positions. Due to this limitation, the studies related to dynamic mobility pattern characterization
60 are less developed. In Toole et al. (2015), a CDR-based origin-destination matrix is estimated in a first step, then
61 assigned onto the road network in a second step to estimate the traffic load. Handovers and Location Area Updates
62 are used to estimate traffic speeds on highway segments (Bar-Gera, 2007; Ou et al., 2011), travel time (Janecek et al.,
63 2015), or Macroscopic Fundamental Diagrams (MFD) (Derrmann et al., 2017). However, handovers guarantee a
64 minimum frequency of location updates during calls, which is not the case for other data sources such as traditional
65 CDR or social media logs. Whether UADP data can still be used to derive dynamic traffic characteristics, such as
66 speed, remains an open question.

67 Monitoring urban network traffic speed is crucial for many applications, including traffic control, route guidance,
68 or emission calculations (Zhang et al., 2011); and targeting speeds from irregular and low-frequency positioning data
69 remains the most challenging application. In fact, the traditional bottom-up speed estimation methods from GPS
70 floating vehicle tracks (Zheng et al., 2013; Shang et al., 2014), which rely on averaging individual speeds calculated at
71 the road segment level, cannot be transposed to this kind of data. However, user-activity-dependent positioning data
72 have significant advantages with respect to more conventional traffic data sources. They are usually accessible and
73 massive. Mobile phone data have very high penetration rates among the populations (Blondel et al., 2015; Algizawy
74 et al., 2017; Bachir et al., 2017), which results in excellent spatial coverage in urban areas. Social network data are
75 massive as well. They still offer lower penetration rates (because they correspond to more specific audiences and
76 uses) than cell phone data, but their availability continues growing (Cisco, 2020), offering promising perspectives in
77 more extensive use for mobility analysis. Traffic speed estimations based on GPS floating vehicle tracks often rely on
78 complementary data sources (like surveys, loop detectors, or cameras) to implement spatial extrapolation processes
79 and compensate for the low data coverage (Shang et al., 2014; Zhan et al., 2017), leading to costly overall processes.
80 On the contrary, working with temporally sparse but massive data seems promising as it could offer cost-efficient
81 and large-scale alternative methods. Given the massive amounts in which UADP data are available, and despite their
82 temporal irregularity and sparsity, we aim to prove that they offer in an urban context a viable alternative to GPS
83 floating car data to estimate the mean traffic speed dynamics at a zonal scale. By focusing on UADP data, we consider
84 all massive mobility data related to the use of new technologies and whose temporal sampling frequency depends on
85 users' communication behaviors and activities, and therefore inherently uncertain.

86 A key point of the method we propose is that it is based on the partition of the urban network into regions
87 characterized by homogeneous traffic conditions. This partition defines a new spatial scale at which the individual trip
88 data are up-scaled and analyzed. This aggregation process allows characterizing interrelated travelers, *i.e.*, travelers
89 who simultaneously cross the same network areas, but is also more adapted to the possible raw spatial resolution of the
90 data than the road segment scale. Thanks to this new scale, our method only requires a set of elementary trip features
91 but no explicit characterization of individual local speeds. Those features are the observed departure and arrival time,
92 and the regional path (as defined in Yildirimoglu and Geroliminis (2014); Batista et al. (2019)), *i.e.*, the succession of
93 regions traveled by individuals between their origin and their destination regions.

94 We propose to fuse from the outset the travel time information of individuals traveling along the same regional
95 paths on a periodic regular basis (*e.g.*, every 15 minutes), and conduct, for each of the considered period, a combined
96 analysis of the average travel times estimates derived from this data fusion. Provided that a reliable estimation of
97 the trip lengths at the city and regional scales is available from external offline sources, this analysis allows deducing
98 a broad and consistent estimation of the regional average traffic speeds. One of the main challenges of applying
99 this methodology is the correct estimation of average travel times, despite the temporal biases inherent to the use of
100 user-activity-dependent positioning data. The method we propose relies on statistical considerations to addresses this
101 challenge.

102 We apply the method to a set of artificially temporally-biased trips derived from a real GPS dataset of tracked
103 vehicles traveling in the Great Lyon area, France. This approach allows using the original GPS dataset as a ground-
104 truth reference for traffic speed, against which to assess the methodology and determine whether the simulated data
105 are qualified for urban traffic speed estimation. Although the GPS dataset size is limited, literature works have shown
106 that GPS floating car data was a particularly reliable source for estimating zonal traffic speed. Contrary to other
107 traffic variables, the traffic speed estimation does not require scaling processes. Its estimation from vehicle probe data
108 results in very satisfactory results despite low penetration rates (Nagle and Gayah, 2014; Leclercq et al., 2014). In
109 this research, by keeping the data downsampling process under control instead of directly using UADP data, we aim

110 at better understanding how the jamming and the consequent progressive information loss could impact the quality of
111 the results. By focusing on a synthetic data context that permits clear identification of the temporal bias, this study
112 aims to assess the robustness of the proposed methodology towards its application on non-synthetic UADP data.

113 This article is organized as follows. Section 2 exposes the principle of our approach and describes the proposed
114 methodology. Section 3 presents our case study, as well as the exploited data. Section 4 focuses on the results we
115 reached. Finally, Section 5 concludes with the limits and perspectives of this work.

116 2. Methodology

117 2.1. Problem statement

118 We focus on exploiting vehicle trips extracted from a generic UADP dataset (mobile phone data, LBNS data, or
119 any similar mobility dataset) leveraging the literature stay detection methods. These methods define *stays* as locations
120 (either a specific position or a cluster of positions close to each other) where users are observed for a minimum amount
121 of time. These methods are therefore geared towards identifying static phases of individual mobility. This paper will
122 consider that such methods are reliable and that the stays detected do indeed correspond to static phases. Nevertheless,
123 this identification is dependent on the communication activity of the users. It implies that static phases may only be
124 partially identified if the user is not active at the beginning or the end of their stay. Suppose we define a *trip* as
125 the mobility phase between two consecutive stays. In that case, an important distinction must be made between the
126 observed trip departure and arrival times and the exact (but unknown) ones, as the varying communication rates of
127 users provide sparse information on their mobility. In this paper, we use the following definitions, i being an individual
128 trip:

- 129 • The *observed departure time* is defined as the time when the last static event of the origin stay is observed. By
130 definition, the observed departure time precedes the actual one. In this paper, let ϵ_d^i be the positive bias between
131 these two values (all mathematical notations in the article are listed in the notation table in Appendix A).
- 132 • Reciprocally, the *observed arrival time* is defined as the time when the first static event of the destination stay is
133 observed. By definition, the observed arrival time follows the actual one. Let ϵ_a^i the positive bias between these two
134 values.
- 135 • The *observed travel time* T_{obs}^i is defined as the time elapsed between two consecutive stays, *i.e.*, between the ob-
136 served departure and the observed arrival times. It is an overestimate of the actual travel time T^i .

137 Based on these definitions, we have:

$$\epsilon^i = T_{obs}^i - T^i = \epsilon_d^i + \epsilon_a^i \quad (1)$$

138 Intermediate trip positions can give additional information, considering that the departure time occurs between the
139 observed departure time and the first mobile event. The reasoning is symmetrical for the arrival time. Therefore, the
140 longer the delay between consecutive moving and static events, the more uncertain the departure and arrival times, the
141 greater the risk of significant overestimation of the individual travel time. These individual biases are, by nature, very
142 difficult to estimate at the trip level, and they affect the observations of the individual travel time themselves.

143 In this context, the problem we address in this paper is the following. Can we provide a method to correctly
144 estimate traffic speed at least at an aggregated regional scale despite these unknown individual biases?

145 *2.2. Overview*

146 The fundamental principle behind the speed estimation method we propose is that the overall sample size of the
147 data can compensate for the low data quality at the individual trip level. The method relies on the fusion of individual
148 trip information and statistical considerations to provide a reliable regional traffic speed estimation. It requires the
149 implementation of the following steps.

- 150 1. Network partitioning and time resolution definitions;
- 151 2. Average de-biased travel time estimation through the periodic gathering of similar trips;
- 152 3. Speed calculation through the resolution of a linear system model;
- 153 4. Speed trends smoothing.

154 These steps constitute the generic skeleton of the methodological framework we propose. However, some of these
155 steps will require adaptation to the specific properties of the input data and case study.

156 The network partitioning is the starting point of our methodology. It participates in the definition of the spatio-
157 temporal resolution of the final speed results and identifying similar trips. Fine-resolution road network data constitute
158 the primary input of such a network partitioning process. However, the spatial resolution of the analyzed UADP data
159 determines the minimal resolution of the regional segmentation of the city. For instance, the spatial resolution of cell
160 phone data generally corresponds to the underlying base station network. In this case, the partitioning of the urban
161 network must result in larger regions than the Voronoi tessellation of the base station network.

162 The regional partitioning of the network impacts the data structure required for several inputs.

163 On the one hand, our method requires that the vehicle trips database (the key input of the method) include a coarse
164 representation of the trip trajectory consistent with the previously defined scale, called a regional path. Therefore,
165 the network partitioning affects this feature of vehicle trips. The other trip features are the observed arrival time and
166 observed travel time. This minimal travel data structure corresponds to a generic intermediate format that is reasonably
167 accessible by preprocessing the raw UADP data, regardless of their specific characteristics. The implementation
168 details of this preprocessing step depend on these specific characteristics. They are not addressed in this paper to
169 preserve the generality of our framework. However, in Section 5, we shed light on the challenges linked to this step
170 and provide options to overcome them.

171 On the other hand, the regional partitioning also constraints the average trip length estimates matrix, a critical
172 input for the speed calculation phase. This matrix records local average trip lengths according to different macroscopic
173 itineraries. Section 2.3 provides more details on its structure. This matrix is computed once, offline, and before the
174 trip data analysis. This calculation can be based on the analysis of GPS data, if available, as done in this study. Those
175 data must have sufficient coverage to calculate statistically reliable distances. However, alternatives exist, such as
176 methods that exploit travel surveys or the automatic and systematic analysis of the road network topography (Batista
177 et al., 2019).

178 Finally, the travel time estimation step requires an accurate evaluation of the average travel time bias caused by
179 uneven user activity patterns. This evaluation, which relies on an analysis of the specific UADP data, is considered to
180 be an input of the method. It will allow the observed travel times to be de-skewed and the average travel times to be
181 estimated correctly.

182 Figure 1 illustrates the succession of the methodological steps and their articulation with the different inputs cited
183 above. The following sections describe in more detail each of these steps.

184 *2.3. Network partitioning and time resolution definitions*

185 One of the essential steps in the methodology is the identification and the fusion of similar trips. However, sparse
186 trips distributed over space and time are difficult to compare and relate to one another. In this section, we first propose
187 to define a new spatial and temporal scale, thus laying the ground for the definition of comparison criteria between
188 different trips. The definition of such a new scale relies on both spatial and temporal aggregation.

189 We first define a new spatial scale. The targeted urban road network is partitioned into regions. These regions
190 must mainly be characterized by homogeneous city fabric, demography, road network topology, and, most importantly,

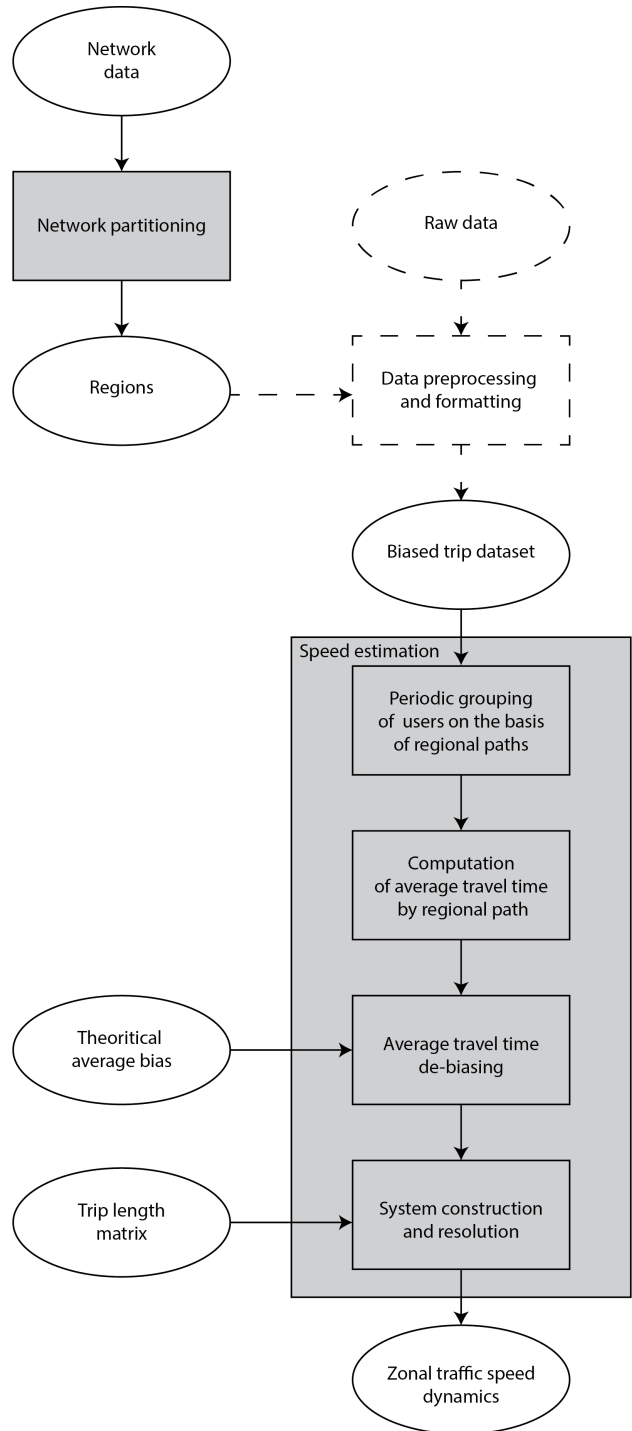


Figure 1: Methodological framework

191 traffic dynamics. Homogeneity of traffic dynamics is an essential requirement for a robust estimation of the regional
192 mean speed, as shown by the literature on the Macroscopic Fundamental Diagram (Daganzo, 2007; Geroliminis and
193 Daganzo, 2008). Following the network partitioning guidelines provided by the related literature, one can divide a
194 city into a set of regions usually ranging from 5 to 20. This new spatial scale will later determine the final spatial
195 resolution of traffic speed estimates. Therefore, it must be adapted to the precision requirements of the case study and,
196 where appropriate, to the resolution of the data, as mentioned above. This regional scale provides the background for
197 defining a fundamental notion of our method, the *regional path*:

- 198 • The *regional path* is defined as the sequence of the successive regions traveled from the origin to the trip destination.
199 Therefore, it is a coarse representation of the path followed at the road segment scale, consistent with the regional
200 partitioning of the network.

201 This up-scaling process is illustrated in Figure 2. While Figure 2a displays an individual trip at the road segment
202 network, Figure 2b represents its corresponding regional path $R_1R_4R_6$. Further on, we will consider that trips follow
203 the structure defined here:

- 204 • We call *trip* the ternary structure defined by a regional path, an observed travel time and an observed arrival time.

205 The trip length estimation that must be performed beforehand of the method is also constrained by the previously
206 defined regional scale and paths. This input shall record the average regional trip length in each region along each
207 possible regional path. Thus, it can take the shape of a distance matrix \hat{L} where rows are the different possible regional
208 paths, and columns are the different regions resulting from the spatial tessellation. The cell value at (i, j) corresponds
209 to the average distance traveled in the j^{th} region, when traveling along the i^{th} regional path P . It is equal to zero if the
210 path P does not cross the j^{th} region. This matrix is assumed to be constant over time, but time-dependent patterns can
211 be considered if they can be characterized independently on another dataset (Batista et al., 2021a).

212 Besides the change of spatial scale, we define a new temporal resolution. The evaluation period is discretized into
213 equal time intervals. This new temporal reference imposes the temporal granularity of the speed evaluation and must
214 be chosen accordingly. In particular, the temporal unit must be small enough to reproduce the rapidly changing speed
215 dynamics during peak hours. We choose 15 minutes in this study, as commonly used in the literature.

216 These processes of partitioning the temporal dimension and the studied network, and the resulting notion of the
217 regional path, provide temporal and spatial criteria for comparing different paths that are otherwise difficult to compare
218 and the basis for identifying similar trips. The following relations are defined:

- 219 • $R1$: Two trips that share the same regional path are called **spatially similar**.
- 220 • $R2$: Two trips that share the same (exact) arrival period are called **simultaneous**.
- 221 • $R3$: Two trips that satisfy both $R1$ and $R2$, *i.e.*, that are **spatially similar** and **simultaneous**, are called
222 **overlapping**.

223 These rules allow establishing a comparison between individual trips. This comparison is a crucial aspect of our
224 methodology, which relies on identifying overlapping trips and fusing their observed travel time before de-skewing it.

225 However, as previously anticipated, it is essential to remind the difference between the exact arrival time and
226 the observed one (extracted from UADP data). Such travel times might correspond to different periods if the user's
227 communication rates are low. Therefore, identifying overlapping trips theoretically requires correcting the observed
228 biased arrival time. Nevertheless, in the two following sections, we neglect this bias and consider that the exact
229 arrival time is known when presenting the core methodological framework. We will relax this favorable assumption
230 in Section 2.6.

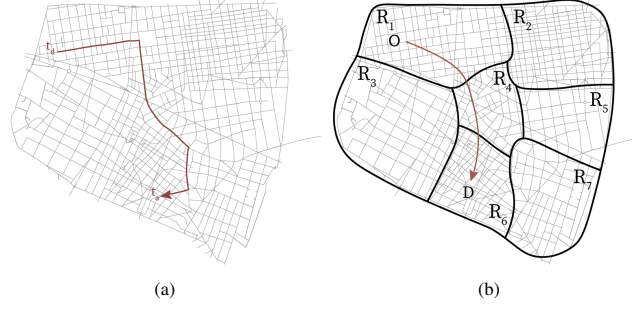


Figure 2: Representation of the different data quality for a same individual trip. (a) GPS track of an individual departing from their origin at time t_d and arriving at time t_a . (b) The scaling up of the track at a regional scales accounts for those inaccuracies and reduce the route to a core path feature: the regional path $R_1R_4R_6$

231 2.4. Average travel time estimation

232 The robust estimation of travel times over the network is a critical milestone in our methodology. The travel time
 233 observations featured in the trip data can provide, to some extent, a snapshot of the traffic conditions that individuals
 234 encounter along their regional route at a given period. However, at an individual level, these observations are not
 235 reliable enough because they are sensitive, on the one hand, to the microscopic origin, destination, and routing of trips
 236 at the network level and, on the other hand, above all, to the frequency of individual observations.

237 As anticipated in the previous section, the observed travel time of a trip can be related to its exact travel time via
 238 the introduction of an additive temporal bias. Although other (non-additive) forms of bias could be considered, this
 239 model is the simplest to start with, and *a priori* the most natural. Let P be a regional path, and let i be an individual
 240 trip traveling along P . We thus have:

$$T_{P,obs}^i = T_P^i + \epsilon^i \quad (2)$$

241 where $T_{P,obs}^i$, T_P^i and ϵ^i are, respectively, the observed travel time of i along P , the exact travel time of i along P ,
 242 and the travel time bias of trip i .

243 Although estimating this individual bias would allow de-skewing the observed travel time, this bias is, by nature,
 244 difficult to assess. However, the estimation of its average seems less challenging and can allow to de-skew on average
 245 the observed travel times. This average bias is assumed known and to be an input of our framework. This hypothesis
 246 is discussed in Section 5. To this end, we propose merging overlapping trips and averaging their observed travel times
 247 to build a unique aggregated biased travel time information by path and period.

248 Let t represent a generic period, and let I_P^t be the set of overlapping trips along P that reach destination at time t ,
 249 with $n_{t,P} = |I_P^t|$. Averaging Equation 2 over I_P^t gives:

$$\bar{T}_P^t = \bar{T}_{P,obs}^t - \bar{\epsilon}_P^t \quad (3)$$

250 where \bar{T}_P^t , $\bar{T}_{P,obs}^t$ and $\bar{\epsilon}_P^t$ are, respectively, the average actual travel time, the average observed travel time, and the
 251 average bias of trips from cluster I_P^t .

252 Assuming that the bias is independent of the trip path and time (hypothesis H_2 , discussed below), we can consider
 253 that the distribution of individual biases ϵ^i can be modeled via a unique random variable X . The construction of such
 254 a model, and the estimation of its first moment $\mu_X \equiv E(X)$, can offer an approximation of $\bar{\epsilon}_P^t$ allowing the de-skewing
 255 of $\bar{T}_{P,obs}^t$, provided that the sample of individuals associated with this period and path is large enough:

$$\bar{T}_P^t \approx \bar{T}_{P,obs}^t - \mu_X \quad (4)$$

256 One of the great advantages of merging overlapping trips data together is that it makes the estimation of the average
 257 travel time more robust, as long as μ_X can be independently characterized. This trip aggregation greatly reduces the
 258 complexity of estimating travel times and traffic speed from biased temporal data, since neither the estimation of the
 259 individual biases, nor the characterization the bias distribution are needed. Only the estimation of its average value

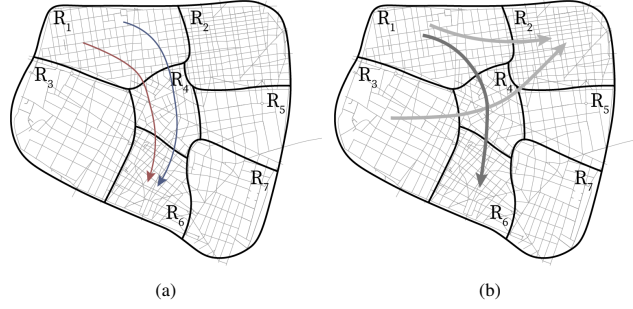


Figure 3: **Visualization** of the data merging into clusters of similar trips. (a) Two individuals traveling simultaneously along a same regional path despite following different (unknown) routes. (b) Merging those individuals into a unique average object (in dark grey) allows characterizing the average travel time needed to travel the regional path $R_1R_4R_6$. This is repeated for every regional path and helps in characterizing the travel time over the whole network.

260 is required. However, the sampling size is a key condition of the process: the larger the sample is, the better the
 261 theoretical average bias μ_X is representative of the sample's average bias.

262 The temporal independence of the bias can be discussed in light of the work of Chen et al. (2018), who showed,
 263 using CDR data, that the inter-event time distribution is sensitive to the hour of the day and. In particular, longer
 264 inter-event times are observed during nighttime and early morning. However, these results account for all individ-
 265 uals, including the ones that are static and sleeping, while we are exclusively interested in moving ones. Thus, our
 266 hypothesis comes down to considering that users' communication activities are more related to their general activity
 267 level (mobile or static) than to the hour of the day, which seems reasonable. The spatial independence of the data is
 268 similarly debatable since mobile phone or social network use is correlated with socio-demographics. Thus, it would
 269 be interesting to validate or refute our hypothesis with a study of the evolution of the inter-event time of traveling users
 270 through time and space, but this goes beyond the scope of this paper. The assumption made here allows considering a
 271 first simple de-skewing approach. Future researches on the travel time bias associated with UADP data could further
 272 complete this approach by differentiating the average bias according to time or space.

273 The systematic estimation, for each P and t , of the observed travel times, and their de-skewing using an average
 274 bias estimate results in a robust, spatially exhaustive, and dynamic evaluation of the travel times across the network
 275 at each period. In that sense, Figure 3 illustrates how two overlapping trips are jointly analyzed to build a unique
 276 representative object of the traffic conditions along $R_1R_4R_6$. Figure 3b also shows how this can be repeated for every
 277 regional path of the network. The speed estimation process relies on this systematic mean regional path travel time
 278 estimation.

279 2.5. Speed estimation

280 This section develops the mathematical foundations of the speed estimation method.

281 Starting at the individual level, we consider an individual trip i of I_P . Its exact traveled time T_P^i along P can be
 282 expressed as the sum of the traveled times $T_{P,r}^i$ over each region r of P (see Equation 5). The regional travel time terms
 283 can be in turns expressed as the fraction of the distance traveled by i in r (i.e., $L_{P,r}^i$) over the mean spatial speed of i in
 284 region r (i.e., V_r^i), as described in Equation 6.

$$\forall i \in I_P, \quad T_P^i = \sum_{r \in P} T_{P,r}^i \quad (5)$$

$$T_P^i = \sum_{r \in P} \frac{L_{P,r}^i}{V_r^i} \quad (6)$$

285 Due to the data temporal sparsity of individual tracks, $T_{P,r}^i$, $L_{P,r}^i$, and V_r^i are considered unknown.

Although vehicles may experience different local and instantaneous speeds over an area, their average speeds depend mostly on overall traffic conditions, and mainly on the accumulation (i.e., number of vehicles in the region).

These speeds show little scatter among individuals, and can be approximated by the mean spatial speed of all individuals traveling in the region. This observation has sustained the development of the MFD theory (Daganzo, 2007; Geroliminis and Daganzo, 2008). The partitioning of the network into sub-regions of consistent traffic dynamics is especially meant to enforce this assumption. On this basis, we assume that each regional speed is homogeneous and constant over the duration of each period t , so that:

$$V_r^i = V_r^t, \quad \forall i \quad (7)$$

286 where V_r^t is the regional spatial mean speed at period t .

287 In Equation 6, after summing on the $I_{t,P}$ trips, this gives:

$$\sum_{i=1}^{n_{t,P}} T_P^i = \sum_{i=1}^{n_{t,P}} \sum_{r \in P} \frac{L_{P,r}^i}{V_r^t} = \sum_{r \in P} \sum_{i=1}^{n_{t,P}} \frac{L_{P,r}^i}{V_r^t} \quad (8)$$

Equation 8 can easily be rewritten as follows:

$$\sum_{i=1}^{n_{t,P}} T_P^i = \sum_{r \in P} \frac{1}{V_r^t} \sum_{i=1}^{n_{t,P}} L_{P,r}^i \quad (9)$$

$$n_{t,P} \bar{T}_P^t = \sum_{r \in P} n_{t,P} \frac{\bar{L}_{P,r}^t}{V_r^t} \quad (10)$$

$$\bar{T}_P^t = \sum_{r \in P} \frac{\bar{L}_{P,r}^t}{V_r^t} \quad (11)$$

288 Again, a significant advantage of this averaging process over the sample I_P^t is that the characterization of individual
289 regional trip lengths $L_{P,r}^i$ for any individual i becomes unnecessary. Instead, the sample mean value $\bar{L}_{P,r}^t$ turns out to be
290 sufficient. On condition that the sampling size is large enough, this can be replaced by its static estimate $\hat{L}_{P,r}$, drawn
291 from the exogenous trip length matrix \hat{L} described above:

$$\bar{T}_P^t \approx \sum_{r \in P} \frac{\hat{L}_{P,r}}{V_r^t} \quad (12)$$

292 At this stage, the computed distance matrix is used to express, through Equation 12, a relationship between the
293 average travel time along path P at period t , and the underlying, unknown traffic speeds of the regions along the P .

294 Although the average trip duration \bar{T}_P^t is unknown, in Section 2.4 we discussed how a knowledge of the average
295 time bias μ_X could allow to estimate it. Based on Equation 4, we thus get:

$$\bar{T}_{P,obs}^t - \mu_X \approx \sum_{r \in P} \frac{\hat{L}_{P,r}}{V_r^t} \quad (13)$$

296 At each period t and for each path P , the average travel time along P , $\bar{T}_{P,obs}^t$, can be derived from the UADP
297 analysis. Conversely, the constant distance parameters $\hat{L}_{P,r}$ are drawn from the aforementioned estimated trip length
298 matrix \hat{L} . μ_X is assumed known as well. V_r^t are the only unknowns of the system. When applying in Equation 13 the
299 change of variable $x_r^t = 1/V_r^t$, we finally get the unbiased system:

$$\forall t, \quad S^t = \{ \bar{T}_{P,obs}^t - \mu_X = \sum_{r \in P} \hat{L}_{P,r} x_r^t, \quad \forall P \}. \quad (14)$$

300 In Equation 14, we name S^t the linear system composed of $|R|$ unknowns ($x_r^t, r \in R$) and as many equations as
301 the number of regional paths observed during the reference period t . The UADP data analysis and the parameters

302 extracted from the trip length matrix allow to fully characterize the system, which can be rewritten in matrix notation
 303 as:

$$\forall t, \quad S^t = \{T_{obs}^t - \mu_X = \hat{L}^t \mathbf{x}^t\} \quad (15)$$

304 where T_{obs}^t is the average observed travel time vector and \hat{L}^t is the sub-matrix of \hat{L} restricted to the regional paths
 305 observed at period t .

306 Given that the number of regional paths will generally exceed the number of regions of the adopted partitioning,
 307 S^t is very likely over-determined. Consequently, the system will probably have no exact solution, but an approximated
 308 one can be calculated using regression analysis. To this purpose, we apply a non-negative least squares regression
 309 method to the system. For a given over-determined linear system $\mathbf{Ax} = \mathbf{y}$, in which \mathbf{A} is a matrix, \mathbf{x} the unknown
 310 vector and \mathbf{y} the response one, the ordinary least square problem consists of finding the optimal x , which minimizes
 311 the sum of the squared residuals. This can be formulated as solving $\mathbf{x}_0 = \operatorname{argmin}_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{y}\|_2$, with $\|\cdot\|_2$ the euclidean
 312 norm. Additional constraints on the elements of x can be added. This is the case in the non-negative least square
 313 method, implying that the coefficient of \mathbf{x} be non-negative. In our case, such constraint allows for taking into account
 314 the non-negative nature of zonal traffic speed. We apply the non-negative least square method to S^t , by solving at each
 315 time step the following:

$$\mathbf{x}_0^t = \operatorname{argmin}_{\mathbf{x}} \|\hat{L}^t \mathbf{x}^t - T_{obs}^t + \mu_X\|_2, \quad \mathbf{x} \geq 0 \quad (16)$$

316 The non-negative least square method *nls*, implemented in Python's package *Scipy*, was used in this paper. Taking
 317 the reciprocal values of the solution vector \mathbf{x}_0^t gives the optimal speed vector \mathbf{v}_0^t . This resolution process can be iterated
 318 throughout the whole studied time span to estimate the complete temporal speed trends. It should be noted that in
 319 this paper, the intra-region trips were filtered out of the system and discarded from the analysis, as they contribute to
 320 a diagonal subsystem whose optimization seems to take precedence over the other system equations in the regression
 321 analysis.

322 2.6. Arrival time correction and data selection

323 The previous sections have considered the arrival period t to be exact. However, when extracting trips from UADP
 324 data, not only the travel time is biased, but so are the arrival time and period. Let t_0^i be the actual precise arrival time,
 325 and $t_{0,obs}^i$ the observed precise arrival time, by opposition to t^i and t_{obs}^i that refer to the actual and observed arrival
 326 periods. We have:

$$t_{0,obs}^i = t_0^i + \epsilon_a^i \quad (17)$$

327 When reducing the temporal resolution to the period level, this implies that the observed arrival period does not
 328 necessarily corresponds to the actual arrival period. This results in the following inequality:

$$t_{obs}^i \geq t^i \quad (18)$$

329 Consequently, identifying simultaneous trips based on the observed arrival period might correspond to considering
 330 together users that refer in reality to other periods, with potentially different traffic speeds. Therefore, the correct
 331 gathering of simultaneous trips ideally requires recovering for each individual their exact arrival times from their
 332 observed arrival times. This recovering cannot be done on average, as for the travel time de-skewing. Deducting the
 333 expected value of the arrival bias (*i.e.*, half of the expected value of the travel time bias μ_X) from the observed arrival
 334 times of each trip, as in Equation 19, only shifts all trips by the same amount of time, but not re-assign each trip to its
 335 correct arrival period.

$$t_0^{ii} = t_{0,obs}^i - \mu_X/2 \quad (19)$$

336 While this shift may help correct an average time offset and slightly modify individuals' grouping, it can in no
 337 way correct the massive mixing of trips together. Such a correction requires the precise estimation of individual biases
 338 separately, which seems very hard to achieve considering the nature of the data. In this paper, we abandon the idea of
 339 applying such an individual bias correction and stick to correcting the average arrival time offset by considering the
 340 new arrival period t_0^i as defined by Equation 19. Nevertheless, to fully meet with the challenge raised by this arrival
 341 bias, we also propose to enhance at each period the robustness of the linear system by implementing filtering solutions
 342 at different levels.

343 Firstly, one can consider filtering individuals according to a criterion based on their communication rates, in order
 344 to limit to some extent the mixing of individual trips corresponding to different periods. In practice, the individual
 345 overall average inter-event time can be used as an indicator of these communication rates. However, this filter must
 346 be considered with caution, as it might impact the sampling size. In our study, trips are not associated with individual
 347 inter-event times but with individual biases. We will explore the impact of filtering trips based on a criterion addressing
 348 these biases.

349 Second, we suggest implementing a filter on the minimal number of trips to consider that an equation defining a
 350 regional path at a given period is valid. Setting this minimal threshold aims at ensuring the robustness of the travel
 351 time estimates despite potential shuffled trips. One could also consider setting a maximum threshold on the travel
 352 time standard deviation, which could be particularly suitable for large trip samples.

353 The criterion above focuses on the reliability of each equation independently of others. A third element we
 354 consider is the consistency of the equations with each other. As this coherence is quite challenging to evaluate, we
 355 propose a sensitive filtering approach to stabilize the results obtained from a set of indiscriminate equations. The
 356 approach we propose is based on bootstrapping, a statistical inference method based on random sampling. For a given
 357 period, for which the data processing resulted in a system S made of n distinct equations, a set of subsystems S_i is
 358 generated and solved to explore the sensitivity of the results to the structure of the system. Specifically, the generic
 359 subsystem S_i is built by sampling with replacement the same number n of equations from S . Consequently, S_i has
 360 the same number of n equations but possible redundancy for some of them. To take this redundancy into account,
 361 we resolve the system with a weighted least square optimization method. The weight of each equation is given by its
 362 number of occurrences in the subsystem. Thus, the more an equation is sampled from the original system, the higher
 363 its weight in the resolution. This process is iterated over many subsystems (we set the minimal number of iterations
 364 to 100) to explore the results' sensitivity to different sampled equations and weighting parameters. Consequently,
 365 many derived speed solutions are generated at each period, resulting in a speed distribution for each region. We apply
 366 statistical filters to these distributions to filter out the most aberrant values before averaging the remaining speeds.
 367 This process enforces the results' consistency and stabilizes them without explicitly labeling equations as reliable or
 368 not and arbitrarily filtering them out.

369 Although our method does not exclude does not exclude working with a favorable sub-population displaying the
 370 lowest communication inter-event times, the individual filtering limits the reach of the method by reducing the range
 371 of users considered. Filtering individuals according to their inter-event travel time corresponds to considering a sub-
 372 population with a reduced average bias. Therefore, among these three filtering methods, the last two are considered
 373 preferable in the evaluation of our methodology.

374 2.7. Speed trends smoothing

375 The speed estimation process described above is applied independently at each time step. The results of this
 376 recursive application of the method to consecutive periods may present sawtooth instabilities between consecutive
 377 periods due to variations of the regional paths observed, their number or the amount of travelers they represent.
 378 To smooth speed trends over time and ensure consistency of results between consecutive periods, we implement a
 379 dynamic filtering based on a rolling window method. The window size is set to a chosen number of periods n . At each
 380 period t , the smoothed traffic speed is calculated as :

$$\bar{V}_r^t = \frac{1}{n} \sum_{i=0}^{n-1} V_r^{t-\frac{n-1}{2}+i} \quad (20)$$

381 Because the speed trends can vary faster at peak hours than during the remaining periods of the day, we increase
 382 the sensibility of the filter at this time. Then, n is set to 3 (periods) during assumed peak hours, while it is set to 5 the
 383 rest of the day.

384 2.8. Discussion

385 In this section, we discuss a few insights we can retrieve from the structure of the system.

386 First, the structure of the system S^t directly explains the impact of the chosen tessellation. The more fine-grained
 387 the spatial resolution, the larger the system size. Not only does the number of unknown variables (regional speeds)
 388 increase, but so does the number of possible regional paths, and hence of equations. Consequently, an increase in the
 389 number of regions also leads to a relative decrease in regional paths' attendance level, as they are more numerous and
 390 therefore less crossed. This attendance decrease might be problematic as the methodology relies on the hypothesis
 391 of sufficient sample representativeness. Thus, determining the appropriate spatial partitioning raises the question
 392 of finding the suitable trade-off between a fine-grained traffic speed estimation and a system composed of reliable
 393 equations.

394 Additionally, the shape of the system provides an insight into the importance of the average travel time de-skewing.
 395 The speed vector resulting from the approximated resolution of the system S^t of Equation 15 is reliable under the
 396 condition that the system is properly conditioned, *i.e.*, that the average travel time vector $\mathbf{T}_{obs}^t - \boldsymbol{\mu}_X$ is correctly
 397 estimated (the trip length matrix distance factors $\hat{\mathbf{L}}$ being considered as reliable). Without accounting for the average
 398 bias generated by the users' uneven activity rhythms, characterizing the regional network travel times based on the
 399 observed travel times would result in an overstated left side of the system compared to the latent actual average travel
 400 time \hat{T}_p^t . This system would be unrepresentative of the actual traffic speeds and likely to underestimate them.

401 3. Experimental approach

402 To evaluate the performance of the proposed methodology, we apply it to a **UADP** dataset derived from **high-**
 403 **frequency** GPS data through data simplification and downsampling. This evaluation approach, **based on high-**
 404 **frequency raw data instead of low-frequency data**, presents several advantages. First, it provides control over the
 405 average data bias, which is an essential part of the methodology **that** has not been **enough** characterized by the litera-
 406 ture. Second, it allows exploring the impact of the data simplification, the temporal downsampling, and the de-skewing
 407 process on the speed estimation quality. This exploration is a necessary step to identify the strengths and weaknesses
 408 of the method and **adjust it accordingly**. It helps to understand how to increase the robustness of the method before
 409 applying it to real inaccurate and biased **UADP** data, for which the corresponding accurate ground truth GPS tracks
 410 will most likely be lacking. Last but not least, this experimental approach provides an easily accessible and reliable
 411 baseline estimation of the traffic speed dynamics, against which to compare our results. The original GPS dataset
 412 also provides valuable data for estimating the trip lengths, which are assumed in our framework to be derived from
 413 exogenous sources and produced offline before **UADP**-like data processing.

414 3.1. Bias model

415 In the previous section, we discussed how the temporal biases of the data could substantially impact the speed
 416 estimation results and why it was necessary to take into account this bias in order for the speed estimation system to
 417 be adequately conditioned. This led to Equation 15. To the best of our knowledge, no model is characterizing this
 418 bias. Thus, we propose a simplistic and generic bias distribution modeling to downsample GPS trips and simulate
 419 **the temporal characteristics of data with low-sampling rates**. The modeled bias is positive and independent both of
 420 time and space, in agreement with the considerations and assumptions developed in Section 2. The model relies on
 421 the characterization of the sample inter-event time distribution, which is **a standard indicator to measure the sampling**
 422 **rates of user-activity-dependent positioning data**. The generated bias distribution will be applied to the individual trips
 423 in order to simulate **temporal biases** and explore the performance of the method on these downsampled trips.

424 The travel time bias is modeled by a random variable X , which we aim to characterize. **To start, we express the**
 425 **individual travel time bias ϵ^i as the sum of a departure and arrival offsets ϵ_d^i and ϵ_a^i (see Equation 1):**

$$\epsilon^i = \epsilon_d^i + \epsilon_a^i \quad (21)$$

426 ε_d^i being the time difference between the actual and observed departure times from the origin stay, and ε_a^i the time
 427 difference between the observed and actual arrival times at the destination stay. They are **considered** to be positive.
 428 Hence, if the departure and arrival temporal offsets ε_d^i and ε_a^i are themselves modeled by the same random variable Y ,
 429 Equation 21 gives $X = 2Y$. Now, we consider that an individual's departure time from a stay position can occur with
 430 a uniform probability between the pre-departure communication event and the post-departure communication event.
 431 The delay between these two events follows the distribution of the user's inter-event times, which we assimilate with
 432 the population's inter-event times distribution for the sake of simplicity.

433 Mathematically, this means that departure bias follows a uniform distribution law bounded by the population's
 434 inter-event time distribution. Symmetrically, the same reasoning applies to the arrival bias.

435 Here, we model the population's inter-event time by a simple exponential law Z of parameter λ . **While the**
 436 **literature often reports inter-event time distribution closer to truncated power law distribution, we select an exponential**
 437 **distribution here out of simplicity. It requires a single parameter λ directly linked to the distribution average.**

438 **The considerations above lead to:**

$$Z \sim Exp(\lambda) \quad (22)$$

$$Y|Z \sim U(0, z) \quad (23)$$

439 Hence, the probability density function of Z , and the conditional probability density function of Y given the
 440 occurrence of the value z of Z can be written as:

$$f_Z(z) = \lambda e^{-\lambda z} \quad (24)$$

$$\text{and } f_{Y|Z}(y|z) = \begin{cases} \frac{1}{z} & 0 \leq y \leq z, \\ 0 & \text{otherwise,} \end{cases} \quad (25)$$

441 From this, we show (see the detailed calculation in Appendix B) that the probability density function of Y is:

$$f_Y(y) = \lambda \int_0^{+\infty} \frac{e^{-\lambda(y+z)}}{y+z} dz \quad (26)$$

442 and that the first two moments of Y are:

$$E(Y) = \frac{1}{2\lambda} \quad \text{and} \quad V(Y) = \frac{5}{12} \frac{1}{\lambda^2} \quad (27)$$

443 Those results characterize the random variable Y which models the departure and the arrival offsets. This gives
 444 for $X = 2Y$:

$$\mu_X \equiv E(X) = 2E(Y) = \frac{1}{\lambda} \quad \text{and} \quad V(X) = 4V(Y) = \frac{5}{3} \frac{1}{\lambda^2} \quad (28)$$

445 We already stressed the importance of the size of I_p^i to ensure that μ_X is representative of the cluster average
 446 bias. This is all the more important as with our bias model, as the variance of X increases with mean inter-event time
 447 $E(Z) = \frac{1}{\lambda}$:

$$V(X) = \frac{5}{3} E(Z)^2 \quad (29)$$

448 Equation 29 shows that the larger the mean inter-event time is, the more scattered the trip bias distribution will be,
 449 and the more data per period and per regional path will be needed to ensure a reliable de-biasing process.

450 With this model, fairly simple and realistic, we propose a way to simulate the travel time biases related to the users'
 451 variable mobile phone activity rates. The construction of the model makes it possible to approximate the average bias

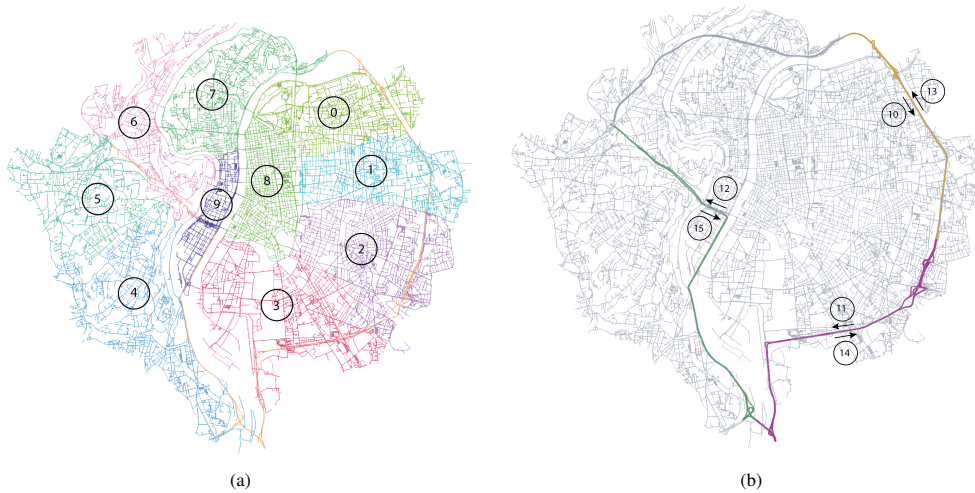


Figure 4: Maps of the regions partitioning the city of Lyon, France. (a) Map of the urban regions; (b) Map of the ring road regions. The ring road is divided into three zones, which are themselves separated into two according to the direction of traffic.

452 of the measured travel times from the analysis of the population inter-event time distribution and deduce it from them.
 453 While an exponential inter-event time distribution was chosen here, the method is transposable to any other observed
 454 distribution. Although this model was developed for the data simulation purposes and not validated against real **UADP**
 455 data, we look forward to evaluating its relevance in a real context. We also believe that the relationship between bias
 456 and inter-event time will remain a key part of a more complex bias modeling process.

457 3.2. Spatial partitioning

458 The city of Lyon, France, is chosen as our case study. The study area includes Lyon and the neighboring munici-
 459 pality of Villeurbanne, located inside Lyon’s ring road. We have parted this territory into sixteen distinct regions. Ten
 460 of them divide the urban areas, while the ring road is extracted and parted into six regions, three per each direction.
 461 Those regions are displayed in Figure 4. Urban regions were manually defined based on the natural geographical
 462 barriers (two rivers) and the major road networks. The major adopted criteria consisted of separating the main arterial
 463 roads in different regions. The traffic variables were verified to be relatively uniform in each region (Mariotte et al.,
 464 2020). We split the ring road into three main blocks based on our knowledge of daily congestion patterns: the north-
 465 east, south-east, and south-west blocks. The remaining north-west section of the ring road is mainly a tunnel. As
 466 the GPS data are lacking in this section, it was ignored in the analysis. We checked whether the two opposite travel
 467 directions could be jointly considered a homogeneous traffic area by analyzing the ring road speed profiles. As the
 468 speed profiles appeared to be significantly different, we decided to split the ring road further, regions per direction.
 469 It is important to mention here that despite the efforts to ensure the homogeneity of the traffic conditions inside each
 470 zone, some aspects of the network structure can be a limitation. In particular, many motorways serve Lyon and relate
 471 it to the neighboring cities. Those motorways cross the urban regions and cause within regions traffic heterogeneity
 472 (region 0, 2, 3, 4, and 6). One solution to limit this heterogeneity would be to isolate those motorways sections into
 473 new specific regions. However, this would **unnecessarily** increase the number of traffic speed variables. **Instead**, we
 474 propose a light and easy-to-implement adaptation of the overall methodology to take this aspect into account. Al-
 475 though this filter is specific to our case study and the chosen partitioning, it can be **applied again** in other contexts, as
 476 cities are often served by expressways passing through peripheral residential areas.

477 Based on our knowledge of the traffic in Lyon, we assume that the trips traveling along those motorways are very
 478 likely to travel along the ring road as well, as a transition to another motorway or their final destination in an urban
 479 region of the city. Consequently, the ring road is assumed to be more strongly connected to these motorways than to
 480 the rest of the urban regions. Hence, we propose to decouple our estimation equation system as follows.

481 On one side, a first subsystem S'_{RR} is built from the regional paths that travel along the ring road at one point.
 482 The system is solved and returns a first speed vector V'_0 . The corresponding equations are assumed to carry reliable

	Day 1	Day 2	Day 3	Day 4	Day 5
Number of trips	19597	20750	20951	21963	22302

Table 1: Number of trips considered per day

483 information about the traffic speed on the ring road. However, the information they carry about the dynamics in the
484 other urban regions (traveled before or after the ring road) is assumed to characterize better the traffic condition in
485 their motorways than in their urban grid. Consequently, while the solution V_0^t is considered reliable for characterizing
486 the ring road speed, it is considered as unreliable to characterize the urban regions' speeds.

487 On the other side, we build a second subsystem S_{URB}^t with regional paths that do not travel along the ring road.
488 The resolution of this new subsystem results in a second speed vector V_1^t . This solution only characterizes the urban
489 regions and is assumed to be reliable on them.

490 Both solutions are merged to build a unique speed vector V^t built from the concatenation of $V_{1|URB}^t$ the speed
491 vector V_1^t restricted to the urban components and $V_{0|RR}^t$ the speed vector V_0^t restricted to the ring road components.

492 3.3. Data description

493 The GPS dataset exploited in this study consists of **cleaned and** map-matched GPS traces over the Greater Lyon
494 area, i.e., an area larger than the perimeter selected for our study. A European navigation system provider collected the
495 data between October 2017 and September 2018. The traces are collected from **multiple** navigation system technolo-
496 gies equipping a multitude of observed floating vehicles (29,000 vehicles per day on average on the Greater Lyon).
497 Moreover, as each trace corresponds to a vehicle, there is no need to filter out pedestrian or cyclist travelers as usually
498 required **when working with mobile phone or social networks data**. This aspect slightly facilitates the problem of
499 estimating traffic speeds, since the question of detecting the mode of transport does not arise here.

500 The trips used in this study were extracted from five typical weekdays, i.e., from Monday, February 12, to Friday,
501 February 16, 2018. As few trips are observed at night-time in our dataset, the time span selected for our evaluation
502 is restrained to day-time hours, i.e., in-between 5 AM and 8 PM. The **data from the full month of February** 2018 was
503 used for the offline calculation of trip lengths.

504 3.4. Trip data preparation

505 The first phase of data processing involves filtering and further cleaning the data. As the area covered by the
506 GPS data is larger than the studied perimeter, we applied a first filter to remove from the data the segments of GPS
507 tracks outside the relevant perimeter. Moreover, the GPS tracks are additionally parsed into different trips when stays
508 are detected. Additional steps included filtering out redundant individuals, static vehicles, and GPS tracks that are
509 fragmented or do not have a spatial consistency, to obtain a clean and reliable data set. At the end of this preprocessing
510 step, the number of trips per considered day is as described in Table 1. **Although these numbers are significant, we**
511 **insist on the importance of a minimal sample size at the period and regional path level. At each time step, and for each**
512 **path, the number of trips must be large enough so that the expected value of the bias is representative of the sampled**
513 **biases. As GPS data are limited in sample size, we artificially extend the size of the dataset by duplicating each trip**
514 **100 times. This trick allows obtaining an extended sampled population, that is then downsampled and biased for each**
515 **individuals.**

516 This GPS trip dataset is then strictly reduced to the trip features needed by the methodology. The actual travel and
517 actual arrival times of each trip are directly extracted from the GPS data observation. Additionally, every GPS track
518 is down-scaled to the spatial resolution previously defined, to obtain the regional path information. Those three trip
519 features (*regional path*, *actual arrival time*, and *actual travel time*) are stored, along with the *trip id*, in a new dataset
520 that will be called DS_0 in the following. At this stage, a first downsampling level has been introduced in the spatial
521 dimension to replace the precise track information with the regional path feature. Although travel times do not yet
522 include any temporal bias at this stage, the trip representation is then already considerably simplified. This dataset
523 will be the subject of our first experiments.

524 The last processing step consists of applying to DS_0 a **temporal** downsampling process that aims to simulate the
525 temporal imprecisions of **UADP** data compared to GPS data. The idea is to simulate the travel time increase caused

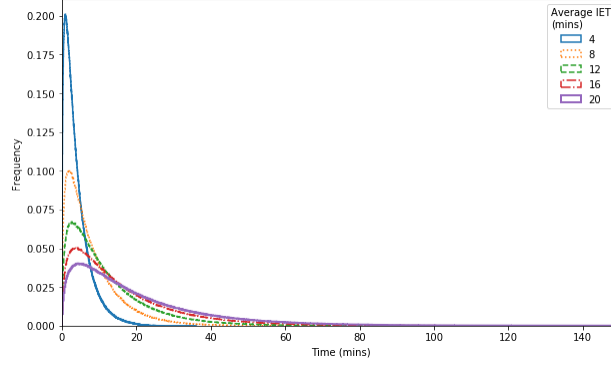


Figure 5: Bias distributions depending on the selected average inter-event time

526 by the temporal biases that the uneven inter-event communication times introduce in the departure and arrival time
527 detection, using the bias model described in Section 3.1. The average inter-event time (IET) is a crucial parameter of
528 this model. The value of this parameter may depend on the population observed **or** on the type of data chosen: **for**
529 **example, using handovers and signalization datasets will display weaker inter-event times than CDR or LBNS data.**
530 To take this inter-event time variability into account and evaluate its impact on our results, we generated, for each
531 day of data, five different downsampled dataset, one per inter-event time value. The selected average inter-event time
532 values are 4, 8, 12, 16, and 20 minutes, **to cover a large range of average communication rates.** The corresponding
533 bias distributions are displayed together in Figure 5. We observe that the larger the average bias is, the more spread is
534 the distribution, with a greater probability for high temporal biases, which was expected with Equation 29. This plot
535 allows understanding that even if the speed estimation method is statistically unbiased, the increasing dispersion of the
536 individual biases makes it necessary to have larger samples when working with important average inter-event times
537 compared to small ones. **This especially justifies the data expansion led above. Downsampling the expanded trips**
538 **sample then allows obtaining an extended bias distribution, for which the average bias will be more representative.**
539 For each trip, we sample the departure and arrival biases according to the probability density function obtained in
540 Equation 26. We generate a second dataset, referred to as DS_1 , which includes the same trips as in DS_0 but whose
541 **actual travel time information is biased by the sum of the sampled departure and arrival biases to obtain the observed**
542 **travel time. This dataset records partially biased trips. It will be the subject of our second analyses to assess our ability**
543 **to correct for travel time bias. In a final downsampling step, we generate a dataset DS_2 in which the actual arrival**
544 **time is additionally biased with the sampled arrival bias. DS_2 records synthetic UADP-extracted trips: individual trips**
545 **characterized by fully biased temporal features and low-quality path information.**

546 One last step before applying the methodology to any of those two datasets consists of grouping the data by
547 regional paths and 15-minutes periods, and averaging the travel time on the resulting groups.

548 3.5. Speed baseline

549 We divide the experiment duration into equal periods of 15 minutes. At each time step t , the method, applied to
550 one of the datasets describe before, returns a vector \mathbf{V}^t whose dimension is equal to the number of regions, in this
551 instance 16. A speed reference is needed to validate our method and estimate the impact of the data downscaling
552 and downsampling processes on the reliability of the results. The spatial mean speed $V_{r,t}^t$ in region r over a period t
553 is defined as the ratio of the total traveled distance in region $TTD_{r,t}$ and the total travel time $TTT_{r,t}$ in region r during t :

$$V_{r,t} = \frac{TTD_{r,t}}{TTT_{r,t}} = \frac{\sum_i d_{r,t}^i}{\sum_i t_{r,t}^i} \quad (30)$$

554 $TTD_{r,t}$ corresponds to the sum of the individual travel distances in region r during t , i.e., $d_{r,t}^i$. $TTT_{r,t}$ corresponds to
555 the sum of the individual travel times in region r during t , i.e., $t_{r,t}^i$. As the GPS data are map matched (see Section 3.3),
556 they include not only temporal and positional data, but also the inferred sequence of road segments traveled, the

557 inferred entrance time on each road link, and the distance traveled on each of them. These characteristics give access
558 to precise individual travel times and distances, and therefore allow obtaining a reliable speed baseline to compare
559 speed estimation results from **temporally-biased trip** data.

560 To conclude, using simulated data obtained through downscaling and the downsampling of **high-frequency** GPS
561 data presents the strategic advantage of offering substantial control over the experimental environment while providing
562 easy access to the necessary distance parameters and the ground truth speed data. The next section exposes the results
563 of the approach on datasets DS_0 , DS_1 and DS_2 .

564 3.6. Trip length estimation

565 Finally, the GPS data is used to estimate the trip length estimation matrix. The entire month of February is used to
566 generate this matrix and ensure robust estimation of trip lengths. As the mobility behaviors are characterized by high
567 redundancy, the trip lengths are mostly unvarying from one month to the other. This means that the trip length matrix
568 can be calibrated using data from a period of time that does not necessarily overlap the time span of the study. This
569 assumption was confirmed by comparing trip length matrices computed in February with a similar matrix computed
570 with March's data. Appendix C exposes the result of this comparison. We will explore the results of the method
571 when estimating travel distance with automatic network analysis in later work. In the meantime, one can refer to the
572 comparison of such trip lengths estimation with GPS data in the work of Batista et al. (2021b).

573 4. Results

574 The speed estimation method that we propose presents the advantages of relying on few mobility features and
575 hence of being easily applicable to **UADP** data. However, it is essential to evaluate the extent to which the low data
576 quality impacts the accuracy of the speed estimation. To this purpose, we proceed in **three** steps.

577 First, we intend to evaluate the impact of working with mobility data of coarser space and time resolution on the
578 results by assessing the errors when working on the DS_0 dataset. The significant degrading of the GPS data might
579 impact the results. Evaluating this impact is essential to understand the overall potential of the method on **temporally-**
580 **biased trip data**.

581 In a **second step**, the method is applied to the **partially** biased dataset DS_1 . First, we estimate the speed dynamics
582 without de-biasing the temporal system, hence with erroneous travel time information, **to evaluate to what extent it**
583 **is necessary to estimate and remove the travel time bias**. Second, we solve the de-biased system, and measure the
584 effectiveness of the de-biasing process to obtain satisfactory speed results.

585 In a **third step**, we will apply the method to the fully biased dataset DS_2 . We compare the results obtained when
586 correcting only the travel time bias with the results obtained when correcting both the arrival and travel time biases.
587 While the data expansion of our trip sample has no impact on the evaluation when using dataset DS_0 , because it does
588 not change the average travel times, we will see that this step is of importance when dealing with both biased datasets.
589

590 4.1. Method application to trip data with exact travel time

591 We start by applying the proposed methodology to dataset DS_0 , **to evaluate in a first step the impact of the spatial**
592 **aggregation and of the speed estimation method**.

593 By using our methodology, we obtain a speed profile in kilometers per hour, per region, and per 15-minutes slots
594 for each day of the evaluation. These speed profiles are compared to the corresponding speed baseline to compute
595 errors. We begin by measuring daily error indicators that characterize the global results of the methodology for
596 the overall regions and periods of the day. We evaluate the mean absolute error (MAE), the root mean absolute
597 error (RMSAE), the mean absolute percentage error (MAPE), and the root mean square absolute percentage error
598 (RMSAPE). Those daily indicators are displayed in Table 2.

599 It is interesting to observe that the daily errors are substantially similar from one day to another.

600 To better assess the performance of the methodology, we now focus on one specific day from our day-set, *e.g.*,
601 Day 1 (Monday, February 12, 2018). Comparable results were obtained for the other days. Figure 6 illustrates the
602 speed estimations dynamics obtained for this day. Each of the subplots corresponds to a region of our partitioning of

	MAE (km/h)	RMSAE (km/h)	MAPE (%)	RMSAPE (%)
Day 1	4.727834	6.548142	13.111325	16.875215
Day 2	4.820239	6.725782	13.169344	16.739542
Day 3	4.906018	7.048133	13.486273	17.370823
Day 4	4.781254	6.909102	14.909294	23.519969
Day 5	4.876119	6.538737	14.013252	17.780414

Table 2: Daily speed errors when applying method to DS_0

603 Lyon. Time throughout the day is represented on the x-axis in hours while the y-axis is for average traffic speed, in
604 kilometer per hour. The ground truth traffic speed, calculated based on the raw GPS dataset, is represented in blue.
605 The orange line corresponds to the raw speed estimation results after bootstrapping. The green line is the result of the
606 moving average filtering that smooths the speed trends. The first ten plots (from Region 0 to Region 9) correspond to
607 the urban regions, while the last six ones characterize the speed dynamics on ring road.

608 For most of the inner regions, we observe that both the raw and smoothed speed trends well match our ground
609 truth profiles, both during peak hours and off-peak periods. Regions 0, 1, 2, 5 and 6 display the most accurate speed
610 estimations. Regions 3, 4 and 7 are those among urban regions where the raw results are the least stable, and for which
611 the smoothing process is the least efficient. The traffic speed in these regions are slightly overestimated.

612 When it comes to the ring road regions (Region 10 to Region 15), we observe a more significant variability of the
613 raw results, with saw-tooth raw speed estimations. For those regions, we observe that the speed estimates reproduce
614 well the speed trends, generally following the ground-truth speed surges and drops during peak hours and matching
615 the faster speeds in-between. Regions 12, 14 and 15 display the most accurate speed estimations. During peak hours,
616 the raw results display important speed drops, and the increased sensitivity of the filter during assumed peak hours
617 (6 AM to 9 AM and 3 PM to 7 PM) proves efficient to reproduce these dynamics while smoothing the results. In
618 Regions 12, 14 and 15 the speed estimates reproduce well the speed drops. The speed estimates in Region 10 during
619 the morning drop are also satisfactory. Region 11 is the most concerning at the peak time, as both its morning and
620 afternoon speed trends are overestimated. After investigating this issue, we suppose that the divergence of the southern
621 end of this section of the ring road, on the one hand, towards Region 12, and on the other hand, towards a freeway, with
622 distinct road behaviors, could be the cause of this anomaly. In the future, we aim at looking more thoroughly at the
623 characteristics of this region and explore how its network features might impact our results in this way. Although some
624 smoothed speed trends in other regions miss reproducing the speed drops to their full magnitude (see in particular:
625 Region 13, morning peak or Region 15, afternoon peak), the deviation from the baseline is much lower than the one
626 related to Region 11, and the results remain satisfactory. As the raw speed estimates reach the lower speeds (Region
627 11, Region 15), modifying the filter during this time window to make it even more sensitive to lower speeds can be a
628 way to reduce this gap and further improve results.

629 In-between the peak periods, the raw results follow the speed baseline and reproduce its speed dynamics. Regions
630 10 and 13 display the largest deviations from baseline with a general under-estimation of the speed during this time
631 window. It is interesting to notice that those regions correspond to the opposite directions of the same section of the
632 ring road: section North-East. Region 10 corresponds to the clockwise direction, while Region 13 corresponds to the
633 counterclockwise direction. The reason for this under-evaluation of the speed is that both those regions are strongly
634 connected to the north-western part of the ring road, which was not considered in this study as it mostly corresponds
635 to tunnels. Hence, the process of filtering the scattered tracks related to this north-western section impacted the
636 number of available trips in regions 10 and 13 more than the other ring road regions. In fact, we observe that the
637 frequentation in those regions is, on average, 10% lower than in the other ring road regions This low frequentation
638 leads to poor representativeness of the average travel time and generates a distance bias between actual and estimated
639 travel distance per region and path. The other ring road regions display satisfactory results during this time window,
640 for which the moving average succeeds in smoothing the raw results and their saw-tooth shape (Regions 11, 12, 14,
641 and 15). However, this filter may be unsuitable if sudden and unexpected speed drops occur outside of peak periods.
642 Despite this limitation, this filter was fast and easy to implement choice. In future work, we will explore other filtering
643 techniques that both allow filtering the small saw-tooth instabilities of the results without neglecting the unexpected

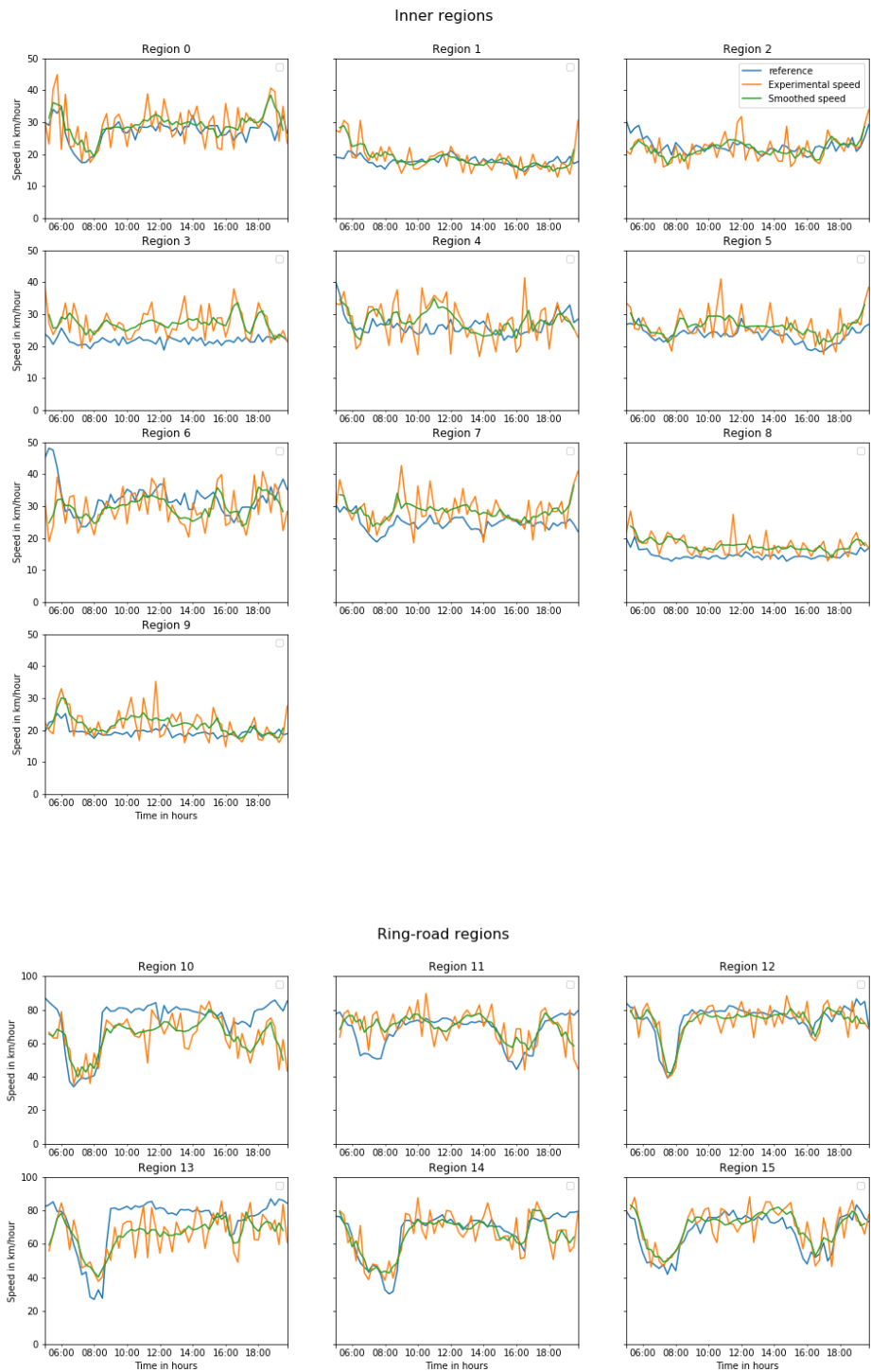


Figure 6: Speed estimation method applied to dataset DS_0 (trips with exact travel times))

	MAE (km/h)			MAPE (%)		
	All regions	Inner regions	Ring road	All regions	Inner regions	Ring road
Full day	4.727834	3.157841	7.344489	13.111325	14.274299	11.173036
Off peak	4.641103	3.090886	7.224797	12.040770	13.794607	9.117708
Peak hours	4.803374	3.216156	7.448737	14.043745	14.692095	12.963161
Morning peak	5.354016	3.665842	8.167639	16.491855	16.713291	16.122797
Afternoon peak	4.287147	2.794576	6.774766	11.748641	12.797223	10.001003

Table 3: Speed MAE and MAPE detailed by region and time window for Day 1

	MAE (km/h)	RMSAE (km/h)	MAPE (%)	RMSAPE (%)
Day 1	3.780724	5.308405	10.711582	14.277904
Day 2	3.698802	5.013979	10.599366	13.624562
Day 3	3.706125	5.343630	10.516593	13.799747
Day 4	4.061264	5.876281	12.163860	18.042898
Day 5	4.078924	5.792345	11.312538	14.590598

Table 4: Daily speed errors when applying method to DS_0 , using the actual trip lengths instead of static trip length estimates

644 speed drop that may occur at any time of the day.

645 In Table 3, we detail the MAE and MAPE errors by period and region type for Day 1. Those errors are computed
646 from the smoothed results. We observe that while the daily error is bigger in absolute value in the ring road regions
647 than in urban regions, the absolute percentage error is smaller for the ring road. Generally speaking, the percentage
648 errors are higher in peak hours than during the off-peak period. However, we also notice that its value is smaller for
649 ring road regions than in inner ones, showing that the method is quite efficient in reproducing the fast-changing speeds
650 of this particular kind of region.

651 Those results are interesting as they give a first insight into the potential of the method. Despite significantly
652 lowering the information carried by individual trips (from GPS tracks to regional paths, and from exact arrival time to
653 arrival period), the method reproduces the speed trends with limited errors. From the perspective of estimating traffic
654 speed from **temporally sparse** data, this is a promising step.

655 However, we can identify several potential improvements. We already mentioned the improvements concerning
656 the smoothing filter. The specific characteristics of Region 11 are also under investigation to understand how they
657 impact the results. More generally, we can only stress the importance of the sample size. In fact, the number of
658 individuals traveling along a regional path at each step must be large enough for the exogenously computed mean
659 travel distance to represent the sample and for the sample’s average travel time to represent the instantaneous dynamics
660 along the path. When working with **massive** data, the amount of data available will ensure this representativeness and
661 compensate for the low data information level. However, working with GPS data present the drawback of having to
662 deal with a limited amount of tracks, and therefore, even more, a limited amount of tracks by regional path and period.
663 This likely results in distance biases between the estimates and the actual average traveled distance, destabilizing the
664 results. Hence, this case study can be considered a worst-case scenario in which the method requires us to work with
665 limited access to trip information.

666 We explored the same speed estimation process from dataset DS_0 when replacing the static trip length estimates
667 by the actual travel distances to validate those considerations. The speed trends for Day 1 are displayed in Figure 7,
668 while the corresponding daily errors can be found in Table 4. The important improvement we observe, especially for
669 Regions 10 and 13, confirms that a finer representativity of the trip length estimates should allow for more accurate
670 results, thus limiting the gaps to the baseline. For this reason, and despite the limitations we mentioned, the method
671 is very promising for an application to a way larger dataset.



Figure 7: Speed estimation method applied to dataset DS_0 (trips with exact travel times), using the **dynamic** trip lengths

Avg IET	MAE (km/h)	RMSAE (km/h)	MAPE (%)	RMSAPE (%)
4	7.894771	18.576767	10.773218	21.648403
8	12.970168	31.347989	16.233049	33.484246
12	16.918677	41.133133	20.600466	42.693440
16	19.920772	48.522253	24.008746	49.748655
20	22.204453	54.189049	26.608480	55.214964

Table 5: Speed errors on average over the week for each mean inter-event time selected as downsampling parameter

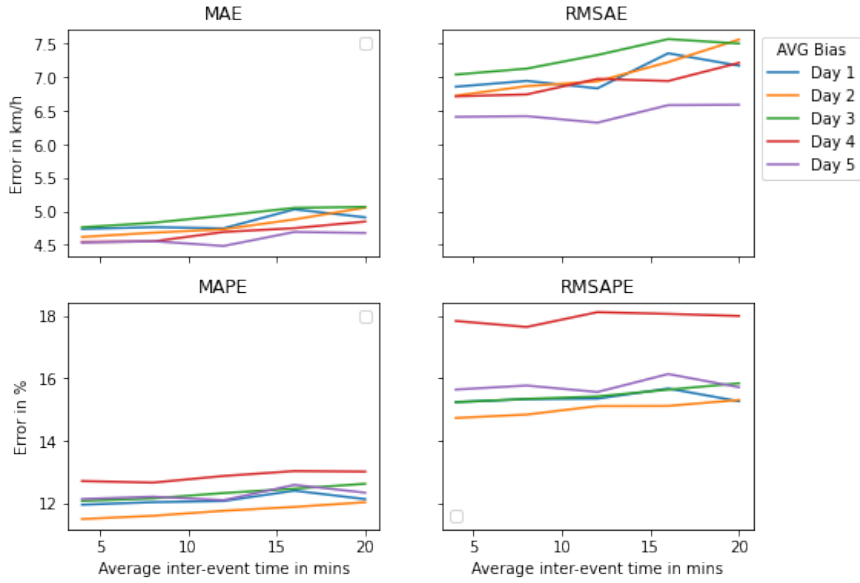


Figure 8: Evolution of daily errors with increase of average bias

672 4.2. Method application to trip data with biased travel time

673 In this section, we evaluate our method on the trip dataset DS_1 made of trips with biased travel times. First, we
674 compute the zonal traffic speeds without applying the temporal bias removal. This allows evaluating the impact of
675 the temporal imprecisions on the results. The variations of average errors over the week with average inter-event time
676 are gathered in Table 5. We observe a significant increase in the errors, compared with Table 2. This shows how a
677 bad estimation of travel times deteriorates the results' quality, even with a short average inter-event time and a limited
678 travel time increase. It justifies the need for **de-biasing in average the travel times**. The following results are computed
679 **applying this de-biasing process**.

680 We display in Figure 8 the evolution for each day of the different daily error indicators as a function of the average
681 inter-event time. We observe that the error indicators are **quite stable and rise slowly with the average inter-event time**.
682 On the contrary, when not expanding the data, the error increases quickly due to the increased dispersion of the bias
683 distribution with the average inter-event time. This shows how important the sample size is and proves the capacity
684 of a large dataset to compensate for the individual biases and imprecision and keep the bias removal process useful
685 despite a large bias dispersion.

686 Figure 9 displays the smoothed results of the speed estimation for the five different average inter-event time values.
687 The results for each value of average inter-event time almost fall into the same line, which confirms the aforementioned
688 results. We observe that we are able, once again, to reproduce the traffic trends and dynamics.

689 In urban regions, the results are satisfactory in Regions 2, 5, 6, 8 and 9. In Regions 0, 1, 3, 4 and 7, the results are
690 less consistent with the speed baseline. While the Region 3, 4 and 7 were already identified in the previous section as

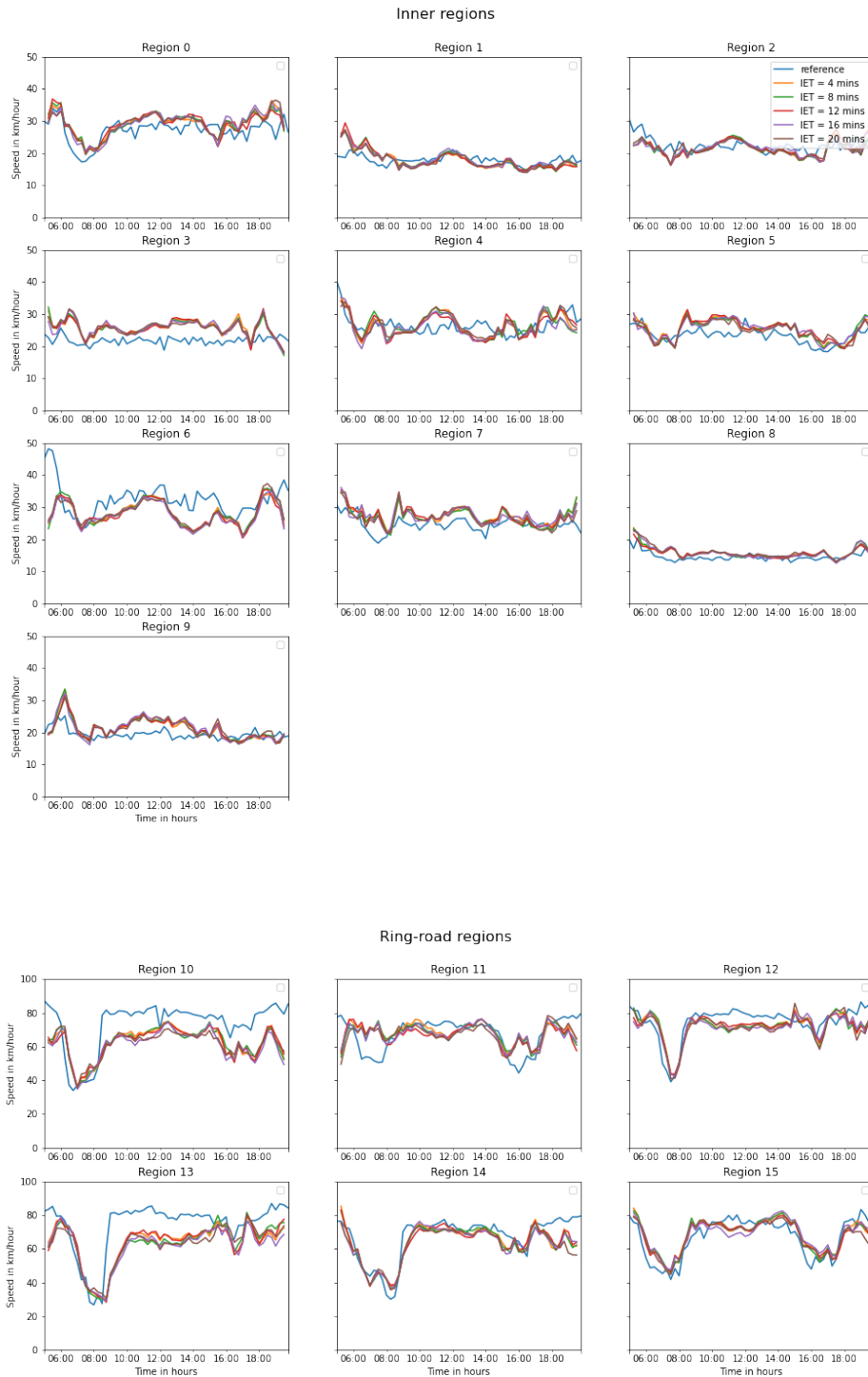


Figure 9: Speed estimation method applied to dataset DS_1 (downsampled trips) after average bias removal.

	MAE (km/h)	RMSAE (km/h)	MAPE (%)	RMSAPE (%)
Day 1	4.911907	7.174379	12.124498	15.264945
Day 2	5.057217	7.564531	12.024699	15.306312
Day 3	5.067316	7.505059	12.617445	15.839542
Day 4	4.848885	7.216612	13.006559	18.000039
Day 5	4.678821	6.591109	12.331655	15.714796

Table 6: Daily speed errors when applying our method to DS_1 in the worst bias scenario (IET = 20 mins)

	MAE (km/h)			MAPE (%)		
	All regions	Inner regions	Ring road	All regions	Inner regions	Ring road
Full day	4.911907	2.794057	8.441657	12.124498	12.113224	12.143288
Off peak	5.181909	2.790098	9.168261	11.769294	11.884125	11.577910
Peak hours	4.676744	2.797506	7.808808	12.433869	12.312762	12.635715
Morning peak	4.961260	3.236535	7.835803	14.194493	14.174446	14.227904
Afternoon peak	4.410010	2.385915	7.783501	10.783285	10.567434	11.143037

Table 7: Speed MAE and MAPE detailed by region and time window for Day 1 in the worst bias scenario (IET = 20mins)

691 displaying less satisfactory results, the increases of the errors for the Regions 0 and 1 can be related to the introduction
692 of the bias.

693 In the ring road regions, most estimated speed trends follow the speed baseline. The speed trends are particularly
694 similar to the baseline in Regions 12, 14 and 15, although some speed drops are not reproduced with their full magni-
695 tude (Region 15, especially), but it was mostly already the case when working with unbiased data. Unsurprisingly, the
696 results in Regions 10 and 13 remain underestimated in-between the peak periods, similarly to the case with unbiased
697 data, but the speed drops are clearly observed. The estimation errors that we had already observed for Region 11
698 during peak hours in the case of unbiased data are increased when using biased data.

699 Finally, we further analyze the worst-case scenario results with an average inter-event time of 20 minutes. Table 6
700 displays the average errors observed for each day in this case, while Table 7 details the precise errors by region type
701 and time window. Compared to the previous section results, we notice a general increase in the errors, although
702 limited. The errors remain under a 20% limit when considering the daily RMSAPE, which is acceptable even though
703 there is room for improvement here.

704 Overall, taking into account the errors previously introduced by upscaling of the GPS tracks to the regional path,
705 working with biased trips seems to have a limited negative impact on the result. Despite the low quality of the trip
706 information at this stage, the results are very encouraging. Therefore, the room for improvement includes the reduction
707 of errors at each stage of the process. This ranges from the representativeness of trip length and time estimates, to the
708 filtering process, to a more refined understanding of the impact of internal speed dynamics in the results.

709 4.3. Method application to trip data with both biased arrival and travel time

710 In the preceding section, we have analyzed our results when using trips with a biased travel time information.
711 However, UADP data not only display biases in the travel time, but on the arrival time as well. In this section, we
712 therefore consider this additional bias on the trips by exploiting the DS_2 dataset, and explore the impact of the methods
713 we propose on such results.

714 First, we compute the results of our method on dataset DS_2 when handling the travel time bias only. The results
715 are displayed in Table 8. Compared to Table 6 for instance, which represents the average errors we obtained for each
716 day in the worst case scenario, those results display a new significant increase of the errors. Since the arrival times
717 are de-skewed, this increasing of the errors is related to the arrival time bias only, which results in mixing together
718 users traveling at different periods and in erroneous travel time estimations. Without surprise, we can observe that the

Avg IET	MAE (km/h)	RMSAE (km/h)	MAPE (%)	RMSAPE (%)
4	6.402842	14.587325	9.344040	17.957539
8	10.950956	24.768931	14.947251	28.044663
12	14.044534	32.729041	18.393624	35.744618
16	16.329243	38.268281	21.118529	41.143766
20	17.901683	42.022657	23.004366	44.958863

Table 8: Speed errors on average over the week for each mean inter-event time selected as downsampling parameter

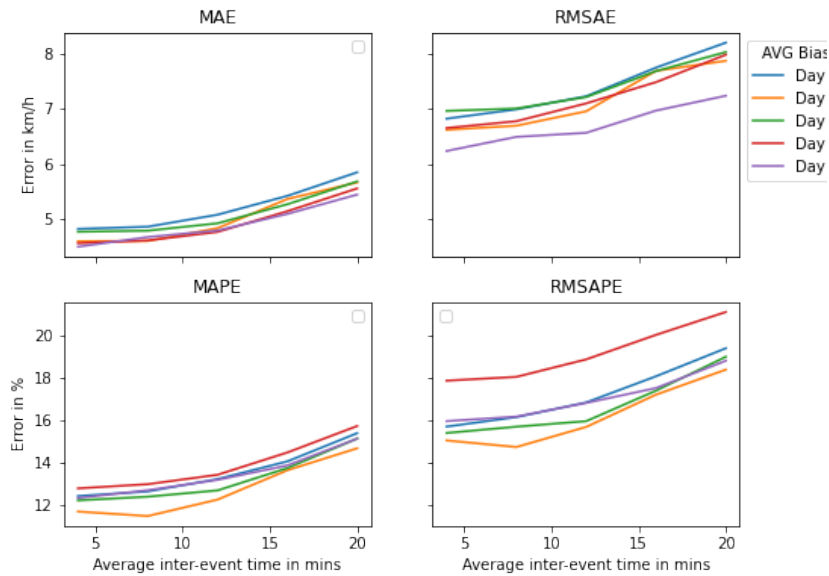


Figure 10: Evolution of daily errors with increase of average bias

larger the average inter-event time is, the larger the errors are, because trips are shifted further away from their actual travel time period.

Therefore, handling this arrival time uncertainty seems necessary, as it was previously done with the travel time. This is what we address in the second part of this section. Figure 10 displays the results obtained once we shift back each trip's arrival time by $\mu_X/2$, remove users with bias larger than twice the average bias and filter regional paths that represent less than 30 individuals.

Although we still observe a sensibility to the average inter-event time (and average bias), the results are contained within much lower bounds than the ones observed in Table 8, showing the filters' efficiency in limiting the arrival time bias impact on the results. However, compared to Figure 5, we observe a larger increasing of the error with the average bias, which can be explained by the fact that the larger the average bias is, the larger the variance, resulting in an increased data shuffling.

In Figure 11, we display the speed estimation results obtained for each inter-event time value on Day 1. These plot display results that can reasonably be compared to the ones exposed in 9. Table 9 precise the daily results in the worst case scenario, while Table 10 precise the errors by region and time period. Overall, the increasing of the errors compared to Tables 6 and 7 is limited, which confirms the viability of our method for estimating regional traffic speeds despite low-quality path information and fully biased temporal features. In particular, these latest analyses demonstrate the utility of implementing filters at the individual and equation levels to compensate for the temporal biases of the data. This suggests that these filters will have great potential when it comes to handling large amounts

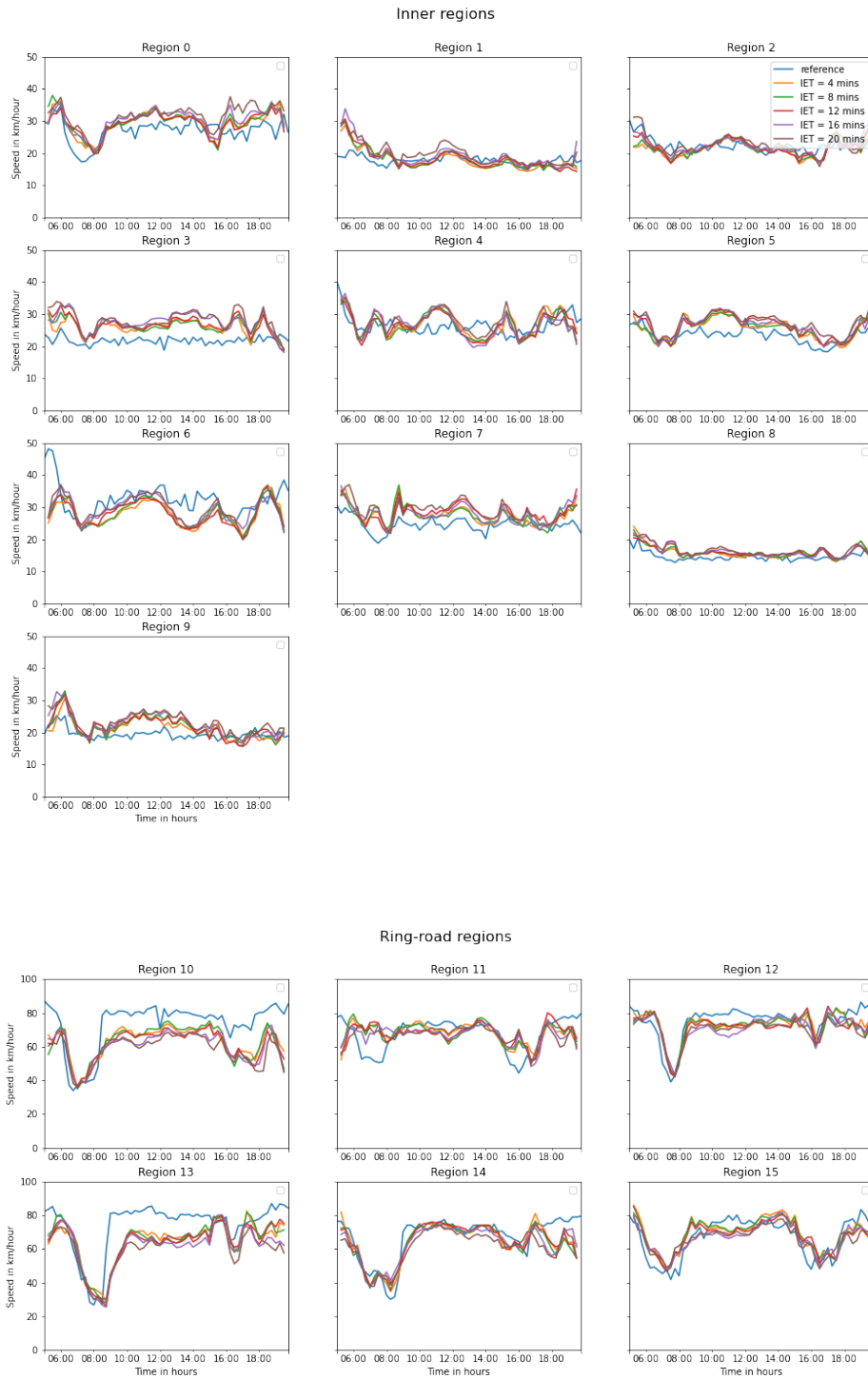


Figure 11: Speed estimation method applied to dataset DS_2 (fully biased trips) after average bias removal and arrival time correction.

	MAE (km/h)	RMSAE (km/h)	MAPE (%)	RMSAPE (%)
Day 1	5.843820	8.214792	15.377055	19.398781
Day 2	5.659251	7.883027	14.660121	18.385150
Day 3	5.676346	8.042682	15.119238	18.995601
Day 4	5.549316	7.994296	15.714570	21.116936
Day 5	5.436428	7.243572	15.116918	18.807797

Table 9: Daily speed errors when applying method to DS_2 , when in the worst bias scenario (IET = 20 mins)

	MAE (km/h)			MAPE (%)		
	All regions	Inner regions	Ring road	All regions	Inner regions	Ring road
Full day	5.843820	3.720738	9.382290	15.377055	16.573524	13.382939
Off peak	5.999818	3.771147	9.714271	14.918967	16.531425	12.231538
Peak hours	5.707951	3.676834	9.093144	15.776034	16.610191	14.385773
Morning peak	5.621608	4.077899	8.194454	17.050427	18.541036	14.566079
Afternoon peak	5.788897	3.300836	9.935666	14.581290	14.800023	14.216735

Table 10: Speed MAE and MAPE detailed by region and time window for Day 1, when in the worst bias scenario (IET = 20mins)

of data, which we are eager to verify.

5. Conclusion and discussion

This paper has proposed a new methodology for estimating the dynamics of regional traffic speeds from user-activity-dependent positioning data. The trips extracted from these data present the challenging issue of being temporally biased, making the individual traffic speed difficult to estimate. To address this issue, the method we propose first relies on the definition of a proper data resolution scale, both on the temporal and spatial dimension, which is used to group and aggregate the user-activity-dependent positioning data. It especially requires the partitioning of the studied area in sub-regions characterized by homogeneous traffic. Such partitioning allows defining basic trip features, such as regional paths, which allow the identification and aggregation of similar trips. This aggregation allows a systematic, exhaustive, and robust estimation of average travel times throughout the network and at each time step through the fusion and de-skewing of individual travel times. Finally, provided that estimates of the regional trip lengths have been performed beforehand, the travel time estimations are jointly analyzed to deduce the underlying regional traffic speeds. This structure of the method is particularly fitted to any massive but temporally sparse data input, as it requires very little temporal or itinerary information at the individual level and considers the inherent temporal bias that characterizes trips extracted from those data.

Applying this approach to downsampled GPS data offers a controlled environment to evaluate the different degrading steps of our approach. First, despite reducing the available GPS trips to minimal temporal and path information, the method could reproduce the speed trends throughout the day, especially the fast-changing dynamics observed in the ring road regions. The moving average smoothing filter that we implemented in this article was proved to be efficient to smooth the period-to-period instabilities of the results. More elaborated filters can replace this one in future work. Despite these satisfactory results, two regions, in particular, displayed underestimated speeds, which we related both to their representation level and inner dynamics. After the introduction of individual temporal biases on the travel time, we repeated the experiment. Different bias models were explored by making the average inter-event bias vary between 4 (best case) to 20 (worst case) minutes. We showed that, provided that the amount of data was sufficient for the mean bias to be representative of the individual sample, the system could be de-biased and return satisfactory results, although we noted that the error slightly increased. In the last step, the method was applied on trips for which both the observed travel and arrival times were biased. At each step, we have identified methodological options that could help to reduce these errors. Working with a large amount of data was identified as an essential requirement

765 of the method. Indeed, it ensures both good reliability of the systems' equations and a correct de-biasing process,
766 especially when working with datasets characterized by long average inter-event time. The sample size was an issue
767 in our case study, in which we had to deal with both low data availability (related to the GPS data source) and the poor
768 quality we imposed on the data to replicate the characteristics of the data UADP. This problem was circumvented by
769 artificially increasing the size of the trip data set by duplicating each displacement 100 times. When working with
770 massive UADP datasets, this problem should no longer arise because the amount of data per regional path will be
771 much more significant, allowing for greater representativeness of displacement lengths and adequate management
772 of bias dispersion. To further investigate the first aspect, we also showed how more accurate dynamic trip length
773 estimates could reduce errors. It is a promising research direction as several studies in the literature have shown that
774 regional trip lengths are relatively stable from day to day but can experience variations within days related to the
775 congestion spreading (Batista et al., 2021a; Paipuri et al., 2021).

776 In future works, we first would like to explore the sensibility of our method to the different parameters such as
777 the size of the regions or the period duration. We also look forward to testing the robustness and the portability of
778 our method in other geographical contexts. Most importantly, we plan to apply the methodology to user-activity-
779 dependent positioning data, tackling the challenges that GPS data have allowed us to leave aside so far. Despite the
780 promising results of our method, the gaps to fill for reaching this objective are still significant. A significant effort
781 will have to be put into the data preprocessing. This step has not been considered in this work because of the high
782 adaptation to the input data it requires. Although the requirements concerning the trip database are limited, this
783 preprocessing step must not be undermined. Besides data cleaning and smoothing, it will have to especially handle
784 mode detection and reconstruction of regional paths from sparse positioning data. The former question is an essential
785 aspect of the preprocessing phase because it distinguishes vehicle movements from other users who do not contribute
786 to road traffic. However, it also corresponds to a significant scientific challenge considering the sparse temporal
787 resolution of UADP data. The review produced by Huang et al. (2019) provides insight on this issue. Methods such
788 as trip reconstruction, individuals' habits completion from (Chen et al., 2019), or matching with prevalent itineraries
789 at the population scale may be helpful (Batista et al., 2021b) to address the latter question on regional path detection.

790 Above all, we would like to address the critical assumptions concerning the travel time bias we made in this study.
791 The average value was assumed to be known and considered static in time and invariant to space. While the temporal
792 characterization of inter-event times in UADP data has been explored in several works, the specific question of the
793 bias existing between observed and actual travel times when working with trips derived from these data has, to the
794 best of our knowledge, never been explored by the literature. Therefore, the assumed characteristics of the temporal
795 bias are difficult to confirm or invalidate. The leads for the estimation of such a bias are also limited. Although our
796 method only requires an estimate of the average bias and not a full characterization of its distribution, the lack of
797 literature on the subject limits the immediate application of this paper. In this paper, we have proposed a simplistic
798 model relating this bias with the inter-event time distribution. Although the objective of this model was mainly to
799 provide a methodological context for the sub-sampling of data, we believe that the characterization of this bias does
800 indeed require relating it to the inter-event time distribution. We would like to investigate this question further.

801 **Acknowledgements**

802 This study is part of a larger research and development project, the Green City Big Data project, carried out by the
803 Citepa, Paris, France.

804 **Authors contributions**

805 M. Sepecher contributed to the conceptualization, methodology, data curation, results analysis, validation, and
806 writing of the original draft of the paper. L. Leclercq contributed to the conceptualization, methodology, results
807 analysis, and review & editing of the paper. A. Furno contributed to the methodology, results analysis, and review &
808 editing of the paper. D. Lejri contributed to the methodology, results analysis, and review & editing of the paper. T.
809 Vieira da Rocha contributed to the funding acquisition, project administration and review of the paper. All the authors
810 have approved the final version of this paper submitted to publication.

Appendices

812 A. Table of notations

813 Table A.1 summarizes the notation used in this paper.

Table A.1: Nomenclature used in this paper.

General notations:

P	Generic regional path
r	Generic region
t	Generic time period

Individual trip characteristics:

i	Generic individual trip
t_0^i	Actual arrival time of trip i
$t_{0,obs}^i$	Observed arrival time of trip i
t^i	Actual arrival period of trip i
t_{obs}^i	Observed arrival period of trip i
$T_{(P)}^i$	Actual travel time of trip i (along P)
$T_{(P),obs}^i$	Observed travel time of trip i (along P)
ε_d^i	Temporal bias of trip i existing between observed departure time and actual one
ε_a^i	Temporal bias of trip i existing between actual arrival time and observed one
ε^i	Travel time bias on trip i
V_r^i	Average speed of i in region r
$L_{P,r}^i$	Distance traveled in region r of P by i

Travel time estimation:

I_P^t	Overlapping trips along P reaching destination at t
$n_{t,P}$	Number of trips in I_P^t
\bar{T}_P^t	Average actual travel time of trips in I_P^t
$\bar{T}_{P,obs}^t$	Average observed travel time of trips in I_P^t
$\bar{T}_{P,obs}^t$	Average travel time bias of trips in I_P^t
$T_{P,r}^i$	Actual travel time of trip i in region r
$\bar{\varepsilon}_P^t$	Average bias of trips in I_P^t

Speed estimation:

V_r^t	Mean spatial speed in region r
$\bar{L}_{P,r}^t$	Average distance traveled in r along P during period t
$\hat{L}_{P,r}$	Regional trip length estimate in region r along P
x_r^t	Reciprocal of V_r^t
S^t	Equation system at period t
\hat{L}	Trip length matrix estimate
\hat{L}^t	Sub-matrix of \hat{L} made of the regional paths observed at period P
T_{obs}^t	average observed travel time vector
x_0^t	Solution vector of S^t doing a least square regression

Continued on next page

Table A.1 – Continued from previous page

<i>Bias modeling:</i>	
<i>Z</i>	random variable modeling the inter-event time distribution
<i>Y</i>	random variable modeling the arrival and departure biases distributions
<i>X</i>	random variable modeling the travel time bias distribution
μ_X	Average travel time bias estimate

814

815 **B. Bias characterization**

816 We detail here the calculation leading to the results in Section 3.1. In that section, we defined :

$$Z \sim Exp(\lambda) \tag{B.1}$$

$$Y|Z \sim U(0, z) \tag{B.2}$$

817 Marginalizing over Z , the probability density function of Y can be expressed as:

$$f_Y(y) = \int_0^{+\infty} f_{Y|Z}(y|z) \cdot f_Z(z) dz \tag{B.3}$$

$$= \int_y^{+\infty} \frac{1}{z} \cdot \lambda e^{-\lambda z} dz \tag{B.4}$$

$$= \lambda \int_0^{+\infty} \frac{e^{-\lambda(y+z)}}{y+z} dz \tag{B.5}$$

818 The expected value of Y is then calculated as follows:

$$E(Y) = \int_0^{+\infty} E(Y|Z=z) \cdot f_Z(z) dz \tag{B.6}$$

$$= \int_0^{+\infty} \frac{z}{2} \cdot f_Z(z) dz \tag{B.7}$$

$$= \frac{1}{2} \int_0^{+\infty} z \cdot f_Z(z) dz \tag{B.8}$$

$$= \frac{1}{2} E(Z) \tag{B.9}$$

$$= \frac{1}{2\lambda} \tag{B.10}$$

819 While the variance of Y is given by:

$$V(Y) = E(Y^2) - E(Y)^2 \tag{B.11}$$

820 Yet:

$$E(Y^2) = \int_0^{+\infty} E(Y^2|Z=z) \cdot f_Z(z) dz \quad (\text{B.12})$$

$$= \int_0^{+\infty} \frac{z^2}{3} \cdot f_Z(z) dz \quad (\text{B.13})$$

$$= \frac{1}{3} E(Z^2) = \frac{1}{3} (V(Z) + E(Z)^2) \quad (\text{B.14})$$

$$= \frac{1}{3} \left(\frac{1}{\lambda^2} + \frac{1}{\lambda^2} \right) \quad (\text{B.15})$$

$$= \frac{2}{3\lambda^2} \quad (\text{B.16})$$

821 Thus:

$$V(Y) = E(Y^2) - E(Y)^2 \quad (\text{B.17})$$

$$= \frac{2}{3\lambda^2} - \frac{1}{4\lambda^2} \quad (\text{B.18})$$

$$= \frac{5}{12} \frac{1}{\lambda^2} \quad (\text{B.19})$$

822 C. Trip Length Matrix Variation with time

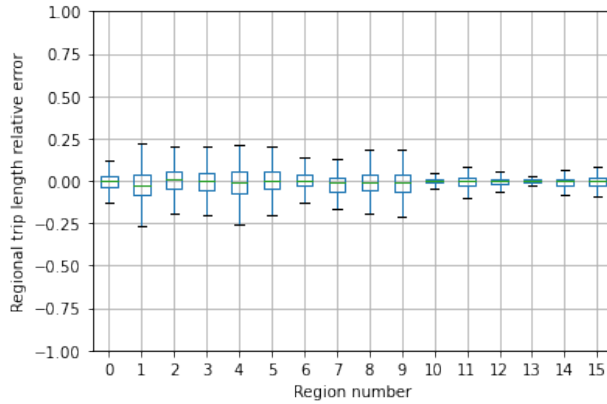


Figure C.1: Box plot of the relative errors of the regional trip lengths by region.

823 In Figure C.1, we display the boxplot describing for each region the distribution of the relative errors between
824 regional trip lengths computed on February's data and March's data. We observe larger errors in urban regions
825 (Regions 0 to 9), while the ring road regions display lower ones. This observation is related to the fact that the
826 trip lengths distances on the ring road are very constrained by the ring road linear structure. On the contrary, a
827 given regional path has a more extensive range of regional trip lengths in the city center, explaining the larger errors.
828 However, the errors are still bounded in the urban regions, which confirms a regularity of regional average trip lengths
829 overtime. This observation supports our framework, as it guarantees that average trip lengths estimated from another
830 period of time, possibly from an independent dataset, will still provide a reliable database for the speed estimation
831 process.

832 References

- 833 Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data.
834 Transportation Research Part C: Emerging Technologies 58, 240 – 250. URL: [http://www.sciencedirect.com/science/article/pii/](http://www.sciencedirect.com/science/article/pii/S0968090X1500073X)
835 [S0968090X1500073X](http://www.sciencedirect.com/science/article/pii/S0968090X1500073X), doi:<https://doi.org/10.1016/j.trc.2015.02.018>. big Data in Transportation and Traffic Engineering.
- 836 Algizawy, E., Ogawa, T., El-Mahdy, A., 2017. Real-time large-scale map matching using mobile phone data. ACM Trans. Knowl. Discov. Data
837 11. URL: <https://doi.org/10.1145/3046945>, doi:10.1145/3046945.
- 838 Asgari, F., Gauthier, V., Becker, M., 2013. A survey on human mobility and its applications. [arXiv:1307.0814](https://arxiv.org/abs/1307.0814).
- 839 Bachir, D., Gauthier, V., El Yacoubi, M., Khodabandelou, G., 2017. Using mobile phone data analysis for the estimation of daily urban dynamics,
840 in: ITSC 2017 : 20th International Conference on Intelligent Transportation Systems, IEEE Computer Society, Yokohama, Japan. pp. 626 –
841 632. URL: <https://hal.archives-ouvertes.fr/hal-01745767>, doi:10.1109/ITSC.2017.8317956.
- 842 Bar-Gera, H., 2007. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel.
843 Transportation Research Part C: Emerging Technologies 15, 380–391. URL: <http://dx.doi.org/10.1016/j.trc.2007.06.003>, doi:10.
844 [1016/j.trc.2007.06.003](http://dx.doi.org/10.1016/j.trc.2007.06.003).
- 845 Barabási, A.L., 2005. The origin of bursts and heavy tails in human dynamics. Nature 435, 207–11. doi:10.1038/nature03459.
- 846 Batista, S., Leclercq, L., Geroliminis, N., 2019. Estimation of regional trip length distributions for the calibration of the aggregated network traffic
847 models. Transportation Research Part B: Methodological 122, 192 – 217. URL: [http://www.sciencedirect.com/science/article/](http://www.sciencedirect.com/science/article/pii/S0191261518311603)
848 [pii/S0191261518311603](http://www.sciencedirect.com/science/article/pii/S0191261518311603), doi:<https://doi.org/10.1016/j.trb.2019.02.009>.
- 849 Batista, S.F.A., Leclercq, L., Menendez, M., 2021a. Dynamic traffic assignment for regional networks with traffic-dependent trip lengths and
850 regional paths. Transportation Research Part C: Emerging Technologies .
- 851 Batista, S.F.A., Seppecher, M., Leclercq, L., 2021b. Identification and characterizing of the prevailing paths on a urban network for mfd-based
852 applications. Transportation Research Part C: Emerging Technologies .
- 853 Blondel, V., Decuyper, A., Krings, G., 2015. A survey of results on mobile phone datasets analysis. EPJ Data Science 4. doi:10.1140/epjds/
854 [s13688-015-0046-0](https://doi.org/10.1140/epjds/s13688-015-0046-0).
- 855 Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C., 2011. Estimating origin-destination flows using opportunistically collected mobile phone location
856 data from one million users in boston metropolitan area. IEEE Pervasive Computing 10, 36–44.
- 857 Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.L., 2008. Uncovering individual and collective human dynamics
858 from mobile phone records. Journal of Physics A: Mathematical and Theoretical 41, 224015. URL: [http://stacks.iop.org/1751-8121/](http://stacks.iop.org/1751-8121/41/i=22/a=224015?key=crossref.97d23b44de724a7398482cd45c7fe01a)
859 [41/i=22/a=224015?key=crossref.97d23b44de724a7398482cd45c7fe01a](http://stacks.iop.org/1751-8121/41/i=22/a=224015?key=crossref.97d23b44de724a7398482cd45c7fe01a), doi:10.1088/1751-8121/41/22/224015.
- 860 Castro, P.S., Zhang, D., Chen, C., Li, S., Pan, G., 2013. From taxi gps traces to social and community dynamics: A survey. ACM Comput. Surv.
861 46. URL: <https://doi.org/10.1145/2543581.2543584>, doi:10.1145/2543581.2543584.
- 862 Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis.
863 Transportation Research Part C: Emerging Technologies .
- 864 Chen, G., Hoteit, S., Carneiro Viana, A., Fiore, M., Sarraute, C., 2018. Enriching sparse mobility information in call detail records. Computer
865 Communications .
- 866 Chen, G., Viana, A.C., Fiore, M., Sarraute, C., 2019. Complete trajectory reconstruction from sparse mobile phone data. EPJ Data Science 8, 30.
867 URL: <https://doi.org/10.1140/epjds/s13688-019-0206-8>, doi:10.1140/epjds/s13688-019-0206-8.
- 868 Cisco, 2020. Cisco Annual Internet Report (2018-2023). Technical Report. Cisco.
- 869 Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C., 2015. Analyzing cell phone location data for urban travel: current
870 methods, limitations, and opportunities. Transportation research record: Journal of the transportation research board 2526, 126–135.
- 871 Daganzo, C.F., 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. Transportation Research Part B: Methodological 41, 49 –
872 62. URL: <http://www.sciencedirect.com/science/article/pii/S0191261506000282>, doi:[https://doi.org/10.1016/j.trb.](https://doi.org/10.1016/j.trb.2006.03.001)
873 [2006.03.001](https://doi.org/10.1016/j.trb.2006.03.001).
- 874 Derrmann, T., Frank, R., Viti, F., Engel, T., 2017. Estimating urban road traffic states using mobile network signaling data, in: 2017 IEEE 20th
875 International Conference on Intelligent Transportation Systems (ITSC), pp. 1–7. doi:10.1109/ITSC.2017.8317718.
- 876 Gandica, Y., Carvalho, J., Sampaio dos Aidos, F., Lambiotte, R., Carletti, T., 2017. Stationarity of the inter-event power-law distributions. PLOS
877 ONE 12, 1–10. URL: <https://doi.org/10.1371/journal.pone.0174509>, doi:10.1371/journal.pone.0174509.
- 878 Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. Transportation
879 Research Part B: Methodological 42, 759 – 770. URL: <http://www.sciencedirect.com/science/article/pii/S0191261508000180>,
880 doi:<https://doi.org/10.1016/j.trb.2008.02.002>.
- 881 Gonzalez, M.C., Hidalgo, C.A., Barabási, A.L., 2008. Understanding individual human mobility patterns. Nature 453, 779 EP –. URL: <https://doi.org/10.1038/nature06958>.
- 882 Hoteit, S., Chen, G., Viana, A.C., Fiore, M.C., 2017. Spatio-Temporal Completion of Call Detail Records for Human Mobility Analysis, in:
883 Rencontres Francophones sur la Conception de Protocoles, l'Évaluation de Performance et l'Expérimentation des Réseaux de Communication,
884 Quiberon, France. URL: <https://hal.archives-ouvertes.fr/hal-01516717>.
- 885 Huang, H., Cheng, Y., Weibel, R., 2019. Transport mode detection based on mobile phone network data: A systematic review. Trans-
886 portation Research Part C: Emerging Technologies 101, 297 – 312. URL: [http://www.sciencedirect.com/science/article/pii/](http://www.sciencedirect.com/science/article/pii/S0968090X1831369X)
887 [S0968090X1831369X](http://www.sciencedirect.com/science/article/pii/S0968090X1831369X), doi:<https://doi.org/10.1016/j.trc.2019.02.008>.
- 888 Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data.
889 Transportation Research Part C: Emerging Technologies 40, 63 – 74. URL: [http://www.sciencedirect.com/science/article/pii/](http://www.sciencedirect.com/science/article/pii/S0968090X14000059)
890 [S0968090X14000059](http://www.sciencedirect.com/science/article/pii/S0968090X14000059), doi:<https://doi.org/10.1016/j.trc.2014.01.002>.
- 891 Janecek, A., Valerio, D., Hummel, K.A., Ricciato, F., Hlavacs, H., 2015. The cellular network as a sensor: From mobile phone data to real-time
892 road traffic monitoring. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS .
- 893 Jiang, S., Fiore, G.A., Yang, Y., Ferreira, J., Frazzoli, E., González, M.C., 2013. A review of urban computing for mobile phone traces: current
894 methods, challenges and opportunities, in: UrbComp@KDD.
895

896 Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., Newth, D., 2015. Understanding human mobility from twitter. PLOS ONE 10, 1–16.
897 URL: <https://doi.org/10.1371/journal.pone.0131469>, doi:10.1371/journal.pone.0131469.

898 Leclercq, L., Chiabaut, N., Trinquier, B., 2014. Macroscopic fundamental diagrams: A cross-comparison of estimation methods. Transportation
899 Research Part B: Methodological .

900 Lin, M., Hsu, W.J., 2014. Mining gps data for mobility patterns: A survey. Pervasive and Mobile Computing 12, 1–16. URL: <https://www.sciencedirect.com/science/article/pii/S1574119213000825>, doi:<https://doi.org/10.1016/j.pmcj.2013.06.005>.

901
902 Mariotte, G., Leclercq, L., Batista, S., Krug, J., Paipuri, M., 2020. Calibration and validation of multi-reservoir mfd models: A case study in
903 lyon. Transportation Research Part B: Methodological 136, 62 – 86. URL: <http://www.sciencedirect.com/science/article/pii/S0191261519306769>, doi:<https://doi.org/10.1016/j.trb.2020.03.006>.

904
905 Naboulsi, D., Fiore, M., Ribot, S., Stanica, R., 2016. Large-scale mobile traffic analysis: a survey. IEEE Communications Surveys Tutorials 18,
906 124–161.

907 Nagle, A.S., Gayah, V.V., 2014. Accuracy of networkwide traffic states estimated from mobile probe data. Transportation Research Record 2421,
908 1–11. URL: <https://doi.org/10.3141/2421-01>, doi:10.3141/2421-01, arXiv:<https://doi.org/10.3141/2421-01>.

909 Osorio-Arjona, J., García-Palomares, J.C., 2019. Social media and urban mobility: Using twitter to calculate home-work travel matrices. Cities 89,
910 268 – 280. URL: <http://www.sciencedirect.com/science/article/pii/S0264275118312976>, doi:[https://doi.org/10.1016/](https://doi.org/10.1016/j.cities.2019.03.006)
911 [j.cities.2019.03.006](https://doi.org/10.1016/j.cities.2019.03.006).

912 Ou, Q., Bertini, R.L., van Lint, J.W.C., Hoogendoorn, S.P., 2011. A theoretical framework for traffic speed estimation by fusing low-resolution
913 probe vehicle data. IEEE Transactions on Intelligent Transportation Systems 12, 747–756. doi:10.1109/TITS.2011.2157688.

914 Paipuri, M., Barmounakis, E., Geroliminis, N., Leclercq, L., 2021. Linear regression analysis of regional mean speed of athens city network using
915 drone data: A multi-modal approach, in: 100th TRB Annual Meeting.

916 Ranjan, G., Zang, H., Zhang, Z.L., Bolot, J., 2012. Are call detail records biased for sampling human mobility? ACM SIGMOBILE Mobile
917 Computing and Communications Review 16, 33–44. doi:10.1145/2412096.2412101.

918 Shang, J., Zheng, Y., Tong, W., Chang, E., Yu, Y., 2014. Inferring gas consumption and pollution emissions of vehicles throughout a city.
919 Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining .

920 Toch, E., Lerner, B., Ben-Zion, E., Ben-Gal, I., 2018. Analyzing large-scale human mobility data: a survey of machine learning meth-
921 ods and applications. Knowledge and Information Systems URL: <https://doi.org/10.1007/s10115-018-1186-x>, doi:10.1007/
922 [s10115-018-1186-x](https://doi.org/10.1007/s10115-018-1186-x).

923 Toole, J.L., Çolak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: Travel demand estimation using big
924 data resources. Transportation Research Part C: Emerging Technologies 58, 162 – 177. URL: [http://www.sciencedirect.com/science/](http://www.sciencedirect.com/science/article/pii/S0968090X15001631)
925 [article/pii/S0968090X15001631](http://www.sciencedirect.com/science/article/pii/S0968090X15001631), doi:<https://doi.org/10.1016/j.trc.2015.04.022>. big Data in Transportation and Traffic Engi-
926 neering.

927 Yildirimoglu, M., Geroliminis, N., 2014. Approximating dynamic equilibrium conditions with macroscopic fundamental diagrams. Transportation
928 Research Part B: Methodological 70, 186 – 200. URL: <http://www.sciencedirect.com/science/article/pii/S0191261514001568>,
929 doi:<https://doi.org/10.1016/j.trb.2014.09.002>.

930 Zhan, X., Zheng, Y., Yi, X., Ukkusuri, S.V., 2017. Citywide traffic volume estimation using trajectory data. IEEE Transactions on Knowledge and
931 Data Engineering 29, 272–285. doi:10.1109/TKDE.2016.2621104.

932 Zhang, J., Wang, K., Lin, W.H., Xu, X., Chen, C., 2011. Data-driven intelligent transportation systems: A survey. IEEE Transactions on Intelligent
933 Transportation Systems 12, 1624–1639. doi:10.1109/TITS.2011.2158001.

934 Zheng, Y., Liu, F., Hsieh, H.P., 2013. U-air: When urban air quality inference meets big data. URL: [https://www.microsoft.com/en-us/](https://www.microsoft.com/en-us/research/publication/u-air-when-urban-air-quality-inference-meets-big-data/)
935 [research/publication/u-air-when-urban-air-quality-inference-meets-big-data/](https://www.microsoft.com/en-us/research/publication/u-air-when-urban-air-quality-inference-meets-big-data/).