On Evaluating Floating Car Data Quality for Knowledge Discovery

Cerqueira, Vitor; Moreira-Matias, Luis; Khiari, Jihed; van Lint, Hans

DOI

[10.1109/TITS.2018.2867834](https://doi.org/10.1109/TITS.2018.2867834)

Publication date

2018

Document Version

Final published version

Published in

IEEE Transactions on Intelligent Transportation Systems

Citation (APA)

Cerqueira, V., Moreira-Matias, L., Khiari, J., & van Lint, H. (2018). On Evaluating Floating Car Data Quality for Knowledge Discovery. *IEEE Transactions on Intelligent Transportation Systems*, 19(11), 3749 - 3760. <https://doi.org/10.1109/TITS.2018.2867834>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' – Taverne project

<https://www.openaccess.nl/en/you-share-we-take-care>

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

On Evaluating Floating Car Data Quality for Knowledge Discovery

Vitor Cerqueira¹, Luis Moreira-Matias, *Member, IEEE*, Jihed Khiari, and Hans van Lint

Abstract—Floating car data (FCD) denotes the type of data (location, speed, and destination) produced and broadcasted periodically by running vehicles. Increasingly, intelligent transportation systems take advantage of such data for prediction purposes as input to road and transit control and to discover useful mobility patterns with applications to transport service design and planning, to name just a few applications. However, there are considerable quality issues that affect the usefulness and efficacy of FCD in these many applications. In this paper, we propose a methodology to compute such quality indicators automatically for large FCD sets. It leverages on a set of statistical indicators (named Yuki-san) covering multiple dimensions of FCD such as spatio-temporal coverage, accuracy, and reliability. As such, the Yuki-san indicators provide a quick and intuitive means to assess the potential “value” and “veracity” characteristics of the data. Experimental results with two mobility-related data mining and supervised learning tasks on the basis of two real-world FCD sources show that the Yuki-san indicators are indeed consistent with how well the applications perform using the data. With a wider variety of FCD (e.g., from navigation systems and CAN buses) becoming available, further research and validation into the dimensions covered and the efficacy of the Yuki-San indicators is needed.

Index Terms—Floating car data, GPS, data quality, traffic control, origin-destination matrix, travel time estimation, trajectory mining.

I. INTRODUCTION

Nowadays, there are multiple providers of raw global positioning system (GPS) measurements, ranging from public transport vehicles to individual pedestrians through their private smartphones. Typically, such measurements - when made on-board a given road vehicle - are known as floating car data (FCD). This digital era empowers intelligent transportation systems (ITS) by enabling the automatic extraction of valuable mobility information through distinct data mining processes.

Manuscript received December 11, 2017; revised June 29, 2018; accepted August 12, 2018. Date of publication September 26, 2018; date of current version November 9, 2018. The Associate Editor for this paper was R. Trasarti. (Corresponding author: Vitor Cerqueira.)

V. Cerqueira was with NEC Laboratories Europe, 69115 Heidelberg, Germany. He is now with Department of Computer Science, Faculty of Sciences, University of Porto, 4099-002 Porto, Portugal, and also with LIAAD-INESC TEC, 4200-465 Porto, Portugal (e-mail: vitor.cerqueira@fe.up.pt).

L. Moreira-Matias and J. Khiari are with NEC Laboratories Europe, 69115 Heidelberg, Germany (e-mail: luis.matias@neclab.eu; jihed.khiari@neclab.eu).

H. van Lint is with Transport and Planning Department, Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2600 GA Delft, The Netherlands.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TITS.2018.2867834

FCD provides multiple sources of information, whose applications may range across different industries [1]. Successful examples in ITS are car sharing [2], mass transit [3] and taxis [4]. However, little attention has been paid to the quality of FCD sources. FCD quality evaluation can be critical to estimate how suitable a dataset is for a particular data mining task, such as road map generation [5], demand estimation [6] or typification [7].

The quality of a mobility data source is inversely proportional to the effort necessary to extract meaningful and yet reliable mobility-related information. As popularity of data science grows across multiple industries, so does the cost of both professionals and software/hardware frameworks. Consequently, the evaluation of the data source quality is key in planning data mining projects across industries. Moreover, the noise usually associated with raw GPS data (e.g. [8]) raises an uncertainty flag on the results which is undesirable to researchers and industrial practitioners in general.

This evaluation process can be split in two distinct dimensions: (i) value and (ii) veracity. The first one addresses how representative a dataset is regarding its original *population* (e.g. how safely can we infer a travel pattern on a city based on such dataset). The second one relates to how reliable a dataset is, which includes GPS error measurements and missing data (e.g. periods of signal absence largely superior to the sampling rate). Additionally, this dimension also includes sample size and rate, city spatial coverage as well as the presence of additional types of data (e.g. weather-based).

Typical work on FCD quality evaluation is often concentrated on the accuracy of GPS measurements with respect to the real location of vehicles [4], [9]–[11]. By focusing on a single dimension, these evaluation approaches are biased, taking little advantage of the large computational power and sample sizes that are available nowadays. The main objective of our research is to overcome the limitations of the existing state of the art on this topic by creating a set of FCD quality indicators with the following characteristics:

- Multi-dimensional: FCD sources are evaluated from different dimensions;
- Data-driven: indicator values are essentially driven by the input data;
- Normalized: the results are quantified in the same scale;
- Scenario and application-oriented: indicators are configurable to different applications and scenarios.

To accomplish this goal we propose *Yuki-san*: a set of statistical indicators to evaluate the quality of FCD. These indicators rely on a series of statistics, unsupervised learning techniques

(e.g. clustering) and external data sources (e.g. commercial road maps) that are proposed throughout this paper. *Yuki-san* is applied to two publicly available probe car data sets collected from taxi fleets running in two cities: Nanjing, China [12] and San Francisco, USA [13].

We evaluate the credibility of such indicators by conducting two simple machine learning experiments over an origin-destination (O-D) matrix: (a) apriori travel time prediction and (b) flow count estimation. By doing so, we aim to demonstrate the insights about the knowledge that we can extract from such datasets on an apriori fashion. *Yuki-san* can have multiple application in the transportation industry ranging from setting the prices of datasets and data sources [14] to filter unreliable sources from feeding advanced traffic visualization/inference frameworks in traffic management centers [15]–[17].

Our main contribution is a unified set of statistical indicators that evaluate the quality of the data broadcast by an FCD source in an automated way. To the authors' best knowledge, *Yuki-san* meets no parallel in the existing literature on FCD quality evaluation.

This paper is structured as follows: Literature review is described in section 2. The proposed methodology for evaluating FCD quality is formally explained in section 3. In section 4, two case studies are presented, followed by the respective experimental evaluation in section 5. In section 6 we outline some important potential applications of the proposed method. The paper is concluded in section 7.

II. LITERATURE REVIEW

There is a wide variety of applications within the mobility domain that benefit from FCD, and there are many different ways to categorize these. In this paper we choose the following taxonomy: (1) digital route/map inference [5], [8], [9], (2) traffic sensing and prediction [15], [16], [18]–[20], (3) trajectory reconstruction and mining [21]–[23], (4) O-D matrix and demand inference [6], [24]. In this section, we briefly overview each of these approaches. Then, we examine the state of the art on FCD quality, pointing out its drawbacks as well as the main contributions of our work.

A. Mining FCD for Different Mobility Applications

Digital route/map inference addresses the problem of inferring road topology from a set of raw GPS measurements. Reconstructing road topology using FCD can be regarded as the inverse problem of map-matching, i.e. reconstructing FCD trajectories using an existing graph representation of the network, which we discuss further below. Cao and Krumm [5] presented a two-step algorithm for digital map inference - first, they cluster similar GPS traces (i.e. likely to be from the same road), and then generate a graph to represent the road network with a simple heuristic. The main contribution of this work is the first step: the aggregation reduces the *noise* (unreliable data derived from inaccurate GPS measurements). On the other hand, Lou *et al.* [8] focus on performing map-matching from low-granularity FCD. The authors point to some empirical

assumptions to deal with such low sampling rate and highlight the drawbacks of the traditional neighborhood-based incremental map-matching approaches in such scenarios. Recently, Hunter *et al.* [9] proposed a conditional random field where the filtering process assumes both a stochastic model for the vehicle dynamics and probabilistic drivers' preferences learned from data. The authors point that the model is robust to low granularity and/or missing data due to an adequate parameter tuning.

Raw GPS data has also been used to (2) estimate the current and/or future traffic status. The central idea in such estimation and prediction applications is that although FCD may not disclose information about flow or density (unless 100% of the vehicles are equipped), they are informative of speed and travel time [19], [20], [25], [26]. To capitalize on this information one first needs to match FCD trajectories to an existing graph of the network. Many algorithms have been proposed to this end, see e.g. [27], [28].

Herrera *et al.* [20] presented "Mobile Century," a famous FCD-based experiment where 100 GPS-equipped vehicles cruised a particular highway stretch in loops to sense its traffic status. Their results show that even a low penetration rate of FCD-enabled vehicles are able to accurately estimate link travel times. In [26] it is shown that, even on the basis of coarse and unreliable GSM positioning data, link speeds can be estimated. Miller *et al.* [19] propose a simple traffic flow model to forecast the link travel time. Their results show that the forecasting accuracy increases logarithmically with the sampling frequency/granularity. Work *et al.* [18] use this dataset to infer regular speed measurements of such segment. The non-linear characteristics of the noise are handled with a non-deterministic sampling technique (Monte Carlo integration). Hofleitner *et al.* [15], [16] proposed a bayesian network to infer the joint probabilities of the number of vehicles to traverse each node as well as the travel times required to do it so based on the states of the remaining nodes. The authors proposed to constrain this graphical model using physical laws from traffic flow theory in order to maintain reliable estimates under noise and/or missing data. More potential for monitoring and estimation arises when FCD are combined with data from other sensors (e.g. loops, camera's) to reconstruct the traffic state [29], [30].

(3) Trajectory reconstruction is the topic of inferring realized routes from data. Typically, (a) the objects of research are typically either on route choice behavioral analysis - a relevant topic in transportation planning [31], [32] - or, in alternative, (b) on getting information on the network - namely, which routes are (not) used, when and by whom - which may be an information of interest for road authorities and transport operators in general [21]–[23], [33]. In [21] a method is proposed to perform intelligent routing using a hierarchical multi-level graph. This hierarchical graph is not based on the well-know highway hierarchies [33]. Instead, each road is assigned to one out of three categories based on the usage of the link. Zhang *et al.* [22] present a framework to detect anomalous taxi trips in three steps: spatial discretization, sequence analysis and trajectory classification.

In the first step, the authors discretize the space into a grid of equally sized cells as a way to reduce the measurements noise. Xue *et al.* [23] go a step further by decomposing the trajectories into a set of sub-trajectories between neighboring locations (i.e. landmarks). This approach tries to deal with data sparsity, both on time and space.

Finally, FCD can also be instrumental in estimating origin-destination (O-D) flow matrices - a matrix which describes the flow counts/mobility demand quantities between major points of interest of a given geographical area. Early studies with probe vehicles highlighted the importance of the spatial and temporal coverage of the traces collected by our probe vehicle fleets in providing unbiased estimations for the vehicle's locations along time [24]. Moreira-Matias *et al.* [6] provided a machine learning framework able to predict the spatio-temporal distribution of taxi-demand demand for short-term horizons. The authors propose a straightforward drift-aware learning approach which is able to produce accurate results even under uncommon traffic conditions (e.g. derived from car accidents or a major soccer match). Despite the accuracy of the proposed method under such conditions, the authors also leverage on the assumption of high sampling granularity/frequency as a backbone aspect of their approach. This issue hinders severely its applicability to FCD sources with low granularity.

The above-mentioned approaches follow one out of two paths with respect to data quality: (a) pose strong assumptions such as a high sampling frequency/granularity or Gaussian noise on the raw GPS measurements or (b) address some of those limitations with techniques tailored for a particular application.

B. FCD Quality Evaluation

Little work has been done to correct/complete or evaluate missing or inaccurate FCD sources. When referring to data imputation, the term is used for different tasks in different fields and applications, making it difficult to analyze its state of the art. In the map generation (1) literature terminology this refers to graph completion, which is usually accomplished with graph logic [34] or simple map data [31]. Regarding (2) traffic state information and O-D matrix estimation, imputation usually means to resort to additional data sources to turn your information more accurate and/or to increase your penetration rate [35]. Finally, for (3) trajectory reconstruction, data imputation typically refers to completing realized paths, which is achieved, for instance, using logic [36] or topological information [37]. In this paper, we are interested on a more classical definition of data imputation - coming from the statistical sense, where missing observations are estimated by resorting to the available data.

Linear interpolation and regression are some of the state of the art techniques for the the data imputation process in FCD. In a landmark paper, Zhang *et al.* [38] propose a framework for both FCD cleaning and imputation. Firstly, unreliable measurements are cleaned based on a statistical approach where data is assumed to be follow a normal distribution. To impute missing data, the authors propose a moving

average method where the coefficients are computed using the Holt-Winters Exponential Smoothing. To correct noisy measurements, the authors use principal component analysis (and the respective reconstruction of the original data using the principal components).

As we analyzed, the literature includes multiple usages for FCD-based knowledge discovery for mobility mining. Yet, the evaluation of the quality of FCD is a relatively new topic. All the above-mentioned approaches follow one out of two paths with respect to data quality: (a) pose strong assumptions such as a high sampling frequency/granularity or Gaussian noise on the raw GPS measurements or (b) address some of those limitations with techniques tailored for a particular application. From the latter, we can highlight that the works typically evaluate one (or a small subset) of the criteria on FCD quality such as granularity [8], [10], [11], [39]; reliability [10], [11], [20]; accuracy [11]; spatial coverage [10], [39] and missing data [10].

References [10] and [39] are the works more similar to our own. The latter elaborates over the work of Hunter *et al.* [9] and Hofleitner *et al.* [15], [16] to study which is the impact of changing the penetration rate (i.e. fleet size) and the sample rate (granularity) over the forecasting error on the link travel time prediction problem. The technical report in [10] characterizes large scale GPS data with respect to the geographical range and time span as well as for the above-mentioned penetration/sample rates. However, it does not provide a quantitative and fully interpretable normalized indicator set focusing attentions instead on providing improved map-matching algorithms and scalable FCD processing pipelines. Moreover, they do not address the dimension of the temporal coverage of the datasets (e.g. the diversity of the temporal timespans covered by them in terms of times of the day or days of the week) neither the relevance of the geographical coverage (e.g. land usage, landmarks, population density) as we propose to in this paper.

By failing to provide a quantitative measurement of the value of a given FCD source, the above-mentioned works end up by just providing contextual and/or partial evaluations which are not reproducible or not comparable to others performed using other sources. Yuki-san aims to fill such gap by providing a multi-criteria statistical evaluation schema whose output is quantified and normalized to a certain range. To the authors' best knowledge, this is first work to do it so in the related literature.

III. METHODOLOGY

This section formalizes Yuki-san: an automatic and interpretable methodology to evaluate the quality of FCD sources for knowledge discovery. Yuki-san settles on a set of statistical metrics that can be divided in two dimensions: (i) value, which focus on evaluating the generalization ability of a dataset, i.e., the value of the quantity of data provided w.r.t. space and time; and (ii) veracity, which is related to the reliability of a FCD source. An illustrative example of the potential relevance of analyzing FCD quality is provided in Fig. 1. In the graphic, two sets of GPS traces with different characteristics are shown, where each colored dot represents a GPS trace.

Fig. 1. Example of different granularities in FCD. Note the loss of routing information on the lower granularity case.

We denote a GPS trace as $x_i = \{lat_i, lon_i\}$, where lat_i and lon_i represent the latitude and longitude at time instance i , respectively. The GPS traces in the right side are captured with a high sampling rate and it is possible to pinpoint the vehicle in space and time with precision. Conversely, in the left is shown an example where fewer data points render loss of information and increased uncertainty.

The indicators encompassing (i) value can be enumerated as follows:

- (A) Granularity, which analyzes the frequency of transmission of GPS traces;
- (B) Macro temporal coverage, for evaluating the time-span and temporal diversity of the data;
- (C) Micro temporal coverage, for analyzing the temporal coverage within each part of day (e.g. daytime);
- (D) Spatial coverage, which provides knowledge on the spatial information present in the data.

On the other hand, (ii) veracity is covered by the following three indicators:

- (E) Missing data, which evaluates any signal gaps in the data;
- (F) Reliability, which aims at measuring the logical precision of the data;
- (G) Accuracy, for inferring the spatial precision of the GPS devices comprising the data.

Our motivation is to propose a set of statistical indicators, which may be configurable from application to application and still provide an invariant interpretation. Consequently, all the indicators were formulated to result in a scalar $\in [0, 1]$ where 1 denotes optimal quality.

A. Granularity

Granularity (G) provides insight about the frequency of the GPS traces transmitted by a given vehicle (sampling rate). A dataset with a high sampling rate is valuable in the sense that we can retrieve information on a vehicle with an higher temporal precision, enabling a better tracking of that same vehicle. This is particularly important for several tasks in ITS, such as map-matching [8] or congestion prediction [15], [16]. We evaluate G by measuring the sampling rate across all vehicles comprising the dataset. This is formalized as follows:

$$\delta_G = \sum_{v=1}^V \tilde{\delta}_v \cdot T_v : T_v \in [0, 1] \quad (1)$$

$$G = \begin{cases} 1, & \text{if } \delta_G < \delta_{opt} \\ \text{erfc}(\delta_G - \delta_{opt}), & \text{otherwise} \end{cases} \quad (2)$$

where V denotes the number of distinct vehicles in the dataset, $\tilde{\delta}_v$ is the median sampling rate of vehicle v , T_v is the percentage of trips of vehicle v in the dataset and erfc denotes the complementary Gaussian error function. δ_G represents the global sampling rate as we formalize it. δ_{opt} denotes the optimal value of sampling rate, which is a user defined parameter. δ_{opt} is configurable since it can differ strongly according to application, the protocol of GPS devices (3G/4G), or travel mode (e.g. road or rail).

The intuition behind such formulation is to find a global sampling rate in the dataset. To do so we perform a weighted average, where the median sampling rate of each vehicle is balanced by the representativeness of that vehicle in the data. The reason for this is because a dataset probably has several vehicles with different GPS devices. Moreover, the use of the median as centrality measure (as opposed to the typical arithmetic mean) is motivated by its greater robustness to outliers, which the sampling rate is prone to.

Despite its insights, Granularity lacks temporal and spatial context. To complement that measuring we start by analyzing the range and diversity of those GPS traces both on space and time - as proposed by the indicators below.

B. Macro Temporal Coverage

Macro temporal coverage ($MaTC$) evaluates the temporal coverage of FCD. This is accomplished by measuring its time-span and diversity in a time scale of a day. This is particularly relevant when addressing demand forecasting tasks (e.g. [6]). In such scenarios, we want FCD as diverse as possible and comprising a large time span. Time span ts is related to the raw size of the dataset and is computed by the following equation: $ts = 1 - \text{erfc}(ndays)$, where $ndays$ is the number of days elapsed from the first to the last GPS trace. The more days covered, the greater ts value is. On the other hand, diversity is related to the spread of weekdays covered: $dv = (1 - \sqrt{\sigma(\rho_{wd})}) \cdot \theta_{wd}$, where $\sigma(\rho_{wd})$ is the standard deviation of the relative frequency of GPS traces in each weekday and θ_{wd} is the ratio of unique weekdays covered (1 if all weekdays are covered). Finally, the value of $MaTC$ is computed taking the arithmetic mean of the ts and dv , along with a penalty Φ , which stands for the ratio of missing days in the dataset (i.e. days without any GPS trace):

$$MaTC = \frac{ts + dv}{2} \cdot \Phi \quad (3)$$

Intuitively, we consider FCD to have good macro temporal coverage if it comprises a large time span with a uniform distribution of weekdays. The main drawback of $MaTC$ arises from its high level formulation. As such, a single GPS trace is enough to consider a day as covered. Nonetheless, we take this issue into account in the indicator formulated below.

C. Micro Temporal Coverage

The main difference between macro and micro temporal coverage ($MiTC$) is that the latter is computed in a finer time

scale, i.e. within the day. Here we aim at understanding how well the data is covering all parts of the day (e.g. morning, evening). The absence of one of these components is critical to understand one key component on transportation-related data mining tasks: the in-day seasonality [40]. Some examples of related phenomena are rush hours (e.g. [41]) or demand peaks generated by a given event (e.g. soccer match) [42]. To formalize *MiTC*, we propose the following formulation:

$$MiTC = \frac{\sum_{d=1}^D \theta_d \cdot (1 - \sqrt{\sigma(\rho_d)})}{D} \quad (4)$$

where θ_d is the ratio of hours covered in part of day d . $\sigma(\rho_d)$ denotes the standard deviation of the relative frequency of the GPS traces in the hours within part of day d . The final value for the indicator is computed through the mean value across all parts in D .

Similarly to the previous indicator, we leverage on the expressiveness of the deviance to the relative frequency of GPS traces to measure its diversity.

D. Spatial Coverage

Indicators III-A, III-B and III-C are mostly related to the temporal component of the data. We start addressing the spatial attributes of FCD by devising a Spatial Coverage (*SC*) indicator. *SC* measures the spatial diversity of FCD. Particularly, its value increases as the spread of the GPS traces across the map increase. However, some areas of the city have greater demand than others, for example, downtown. Consequently, absolute trace frequencies cannot be applied to solve this problem. Alternatively, we devised a schema which takes into consideration the **relevance** (Υ) of each zone.

This procedure starts by discretizing a city spatially through a grid decomposition technique [4], thus obtaining a grid of equally sized cells of $nblocks$ by $nblocks$ blocks.

Then, we quantify the relevance of each cell based on three dimensions: (i) **road density**: One naive way to estimate the importance of a part of a map in terms of mobility is by the number of possible ways there is to cruise that part. The road density of a grid cell gives a rough estimation of how many roads it covers. We start by assigning relevance Υ_{cc} to the grid cell containing the city center (cc) as an *anchor*. The relevance of all other cells is given according to this *anchor*, w.r.t. their road density.

Secondly, we evaluate each cell's (ii) **proximity to landmarks** and other hot-spots (e.g. city center, hospitals, airport), denoted hereby as lm . Empirical observation suggests that most trips start/end in the whereabouts of such hot-spots/landmarks. The relevance of this dimension is computed through Algorithm 1 - in which all grid cells containing at least one landmark are assigned the maximum value for relevance.

Thirdly, we evaluate each cell's relevance given its (iii) **neighborhood**, i.e. their proximity to other relevant cells. E.g.: A grid cell may be relevant, even if it does not have a reasonable road density or is not close to any landmark, for example roads connecting cities to airports. We consider a grid cell to be of some relevance if it is adjacent to any important grid cell. The Algorithm 2 improves the relevance

Algorithm 1 Landmark Importance Imputation

```

1: Input: Grid Cell  $gc \in GRID$ ,  $lm$  locations,  $\Upsilon$ 
2: Output: Updated Relevance of  $gc$ ,  $\Upsilon_{gc}$ ,  $\forall gc \in GRID$ 
3: if  $gc$  contains any landmark
4:   then  $\Upsilon_{gc} = \max(\Upsilon)$ 
5: return  $\Upsilon_{gc}$ 

```

Algorithm 2 Influence Propagation

```

1: Input:  $gc \in GRID$ ,  $\Upsilon$ 
2: Output: Updated Relevance of  $gc$ ,  $\Upsilon_{gc}$ ,  $\forall gc \in GRID$ 
3: for each adjacent grid cell to  $gc$ ,  $gc_{adj}$  do
4:   if  $\Upsilon_{gc} < \Upsilon_{min}$  then  $\Upsilon_{gc} \leftarrow \Upsilon_{gc} + \eta \cdot \Upsilon_{gc_{adj}}$ 

```

Algorithm 3 Spatial Coverage Indicator

```

1: procedure  $GRID = GRIDDECOMP(City\ Map, nblocks)$ 
2: for  $gc \in GRID$  do
3:    $S_{gc} = \sum_{x \in X} x_i : x_i \in gc$ 
4:    $\Upsilon_{gc} = \text{Algorithm 1}(gc, cc, \varphi_{gc})$ 
5:    $\Upsilon_{gc} = \text{Algorithm 2}(gc, lm, \Upsilon)$ 
6:    $\Upsilon_{gc} = \text{Algorithm 3}(gc, \Upsilon, gc_{adj})$ 
7: return  $SC = \frac{\sum_{gc \in GRID} (S_{gc} \cdot \Upsilon_{gc})}{\sum_{gc \in GRID} \Upsilon_{gc}}$ 

```

of the neighborhood cells considering the that of each cell. In other words, if a given grid cell gc has Υ_{gc} below some threshold Υ_{min} , the relevance of its adjacent grid cells gc_{adj} positively influences its relevance by a factor of η .

The combination of these three variables provides a reasonable notion of which parts of the city are more important in terms of urban mobility. The whole procedure for *SC* is described in Algorithm 3, where the total number of GPS traces in a given grid cell gc (S_{gc}) is weighed according to its relevance φ_{gc} .

E. Missing Data

The notion of missing value is a well defined concept in most knowledge discovery applications. However, in FCD this idea is more ambiguous. Generally, a GPS device transmits signals to the data center at a well defined rate, an aspect evaluated in the Granularity indicator III-A. However, there may be significant gaps of time between two transmitted signals within a trip, that we treat as *missing data* (*MD*). This may be caused by device malfunction or misuse.

Formally, we consider that one or more data points are missing - here denoted as *black-hole* - if the time elapsed since the last transmission falls above two times the median sampling rate of the vehicle v in question. This concept is formalized below, where bh_i represents a *black-hole* in a GPS trace $x_i \in X$:

$$bh_i = \begin{cases} 1 & \Delta t_{(i,i-1)} \geq 2 \cdot \tilde{\delta}_v; \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

where $\Delta t_{(i,i-1)}$ denotes the time elapsed between the transmission of x_i and x_{i-1} .

From equation 5 we can then compute the average number of black-holes by trip, r_{bh} . However, black-holes are typically of different durations, which must be taken into account. To address this, we introduce the notion of missing packets (P). Given the global sampling rate δ_G introduced in section III-A, missing packets represent the number of samples or GPS traces that are missing, on average. This is formally defined as follows:

$$\beta = \begin{cases} \lceil r_{bh} \rceil, & \text{if } r_{bh} \leq \xi. \\ \xi, & \text{otherwise.} \end{cases} \quad (6)$$

$$n = \begin{cases} 1, & \text{if } \alpha \leq 1 \text{ or } \beta = 1. \\ (\alpha - 1) \frac{\beta \sqrt{\beta}}{\xi \sqrt{\xi}}, & \text{otherwise.} \end{cases} \quad (7)$$

$$\alpha = \frac{r_{bh} \times \tilde{b}h}{\lceil r_{bh} \rceil \times \delta_G} \quad (8)$$

$$P = \frac{\Omega_d \times \alpha}{n \times \Omega_t} \quad (9)$$

$$MD = \frac{\text{erfc}(P) + G}{2} \quad (10)$$

where r_{bh} denotes the ratio of *black-holes* per trip.

In equation 8, $\tilde{b}h$ stands for the median duration of black-holes. In equation 9, α gives a raw estimate of how many packets are lost. A normalization factor n is used to smooth that value for the cases where those lost samples are spread across the data. A practical example on its *modus operandi* can depart from the supposition that p packets are lost. MD tones down its value as more of those p missing packets are spread across the time-span and not just into a single *black-hole* episode. Furthermore, we add a penalty Ω_d that takes into account the deviance of the *black-holes* duration. Conversely, we smooth P by a factor Ω_t with respect to the median duration of trips.

Finally, in equation 10, this value is normalized according to the granularity indicator (section III-A).

F. Reliability

Another issue when analyzing the veracity of a dataset is related to its logical sense. Some counter examples (i.e. illogical observations) are: i) GPS positions thousands of miles away from the city where one is performing mobility analysis; ii) A vehicle in a given position at a given timestamp and then be marked 10 miles away after 10 seconds. The Reliability indicator covers the objectivity of the dataset. Particularly, we aim at certifying that the GPS traces we are logically possible, both in spatial and temporal terms. Firstly, we introduce the following definitions:

Definition 1: A GPS trace x_i is *awake* if the traveled distance from the previous transmitted signal ($\Delta d_{(t_i, t_{i-1})}$) is greater than the respective sampling rate:

$$x_i \text{ awake} \iff \Delta d_{(t_i, t_{i-1})} > \delta_i$$

As an illustrative example, a given vehicle with a sampling rate of 10 seconds is awake if it traveled at least 10 meters from its last position. From Def. 1, we compute two values: *awake* trace ratio (at), which is the ratio of *awake* traces across

the data; and *awake* trip ratio (aT), standing for the ratio of trips that have a percentage of *awake* traces greater than a given threshold ϵ .

Secondly, we impose one more definition to compute the notion of *reachability*. It is used for uncovering dubious data, from malfunctions of GPS devices or synthetically inputted data.

Definition 2: A GPS trace x_i is *reachable* if the traveled distance from the previous transmitted signal is within a given threshold Ψ , where Ψ is given by a estimate of the average speed times the respective sampling rate:

$$x_i \text{ reachable} \iff \Delta d_{(t_i, t_{i-1})} > \delta_i \cdot \Psi$$

Here, we compute two more ratios: *reachable* trace ratio (rt), which is the ratio of *reachable* traces across the data; and *reachable* trip ratio (rT), which is the ratio of trips with all its comprising points *reachable*.

Finally, the Reliability indicator is computed as follows

$$Reliability = \frac{at + aT + rt + rT}{4} \cdot \kappa \quad (11)$$

where κ denotes the proportion of points that lie inside our *bounding box* - a user-defined *meta*-rectangle that delimits our Geographical Area of Interest.

A drawback of this indicator is the lack of context. For example, it would be simple to generate synthetic data that trick these ratios, e.g. vehicles traveling in loops. This issue is addressed in the Accuracy indicator, where the data is compared to a digital roadmap network (DRN).

G. Accuracy

Accuracy (**A**) operates by measuring the discrepancy between positions given by a GPS device and the true position of the vehicle. The true position is estimated via a Map-matching procedure of the GPS device positions while compared to a DRN. Let $e_i \forall t_i \in X$ be the point-wise error of GPS measurements where t_i represent the GPS traces within a trip t , which belong to the set of all GPS traces X . The error measurement of a trip t is estimated by taking the arithmetic mean of the errors of each GPS trace comprising t . In turn, the general error measurement of the dataset is computed by taking the median value of e_T , the vector containing the error of each trip. We use the median to combine the scores across trips in the interest of robustness (e.g. different GPS devices). Finally we use *Acc* function to transform the estimated value to the interval $[0, 1]$. This is formalized in Algorithm 4.

Algorithm 4 Accuracy Indicator

- 1: **Input:** Set of trips T , DRN
 - 2: **Output:** value of **A**
 - 3: **for** each trip t **do**
 - 4: **procedure** MAP-MATCHING(t, DRN)
 - 5: Return e_t , estimated error measurement of the GPS traces containing t
 - 6: $e_T = \langle e_t \rangle, \forall t \in T$
 - 7: **return** $A = \text{Acc}(\text{median}(e_T))$
-

TABLE I
DATASETS DESCRIPTION

	Nanjing	San Francisco
No. GPS traces	18 million	11 million
No. Trips	432.899	959.025
No. vehicles	7648	536
Timespan	1 day	23 days
Median Trip Duration	10 min	7.5 min

There are several approaches for the map-matching task in the literature (e.g. [5], [8], [9]). This is a particularly tricky problem for FCD in urban environments, where in a small range we may have many candidate roads as matching possibilities. We remark that the computation of the Accuracy indicator is orthogonal to the method of map-matching.

IV. CASE STUDIES

Two FCD sets were collected from taxi fleets operating in two case studies briefly described in Table I: Nanjing (NJ) [12] and San Francisco (SF) [13].

The datasets have several domain-dependent attributes, e.g. fare type. In order to use only information that we can generalize to other FCD (i.e. not only taxi fleets) the attributes used throughout our analysis are the following five: **Timestamp, Vehicle Id, Trip Id, Latitude** and **Longitude**.

V. EXPERIMENTS

This section presents the experimental evaluation carried to validate the proposed methodology. We start by introducing the testbed that we used to evaluate the two FCD sets, followed by the setup of two mobility-related data mining experiments. These experiments carry out two different predictive analytics tasks over the datasets. Finally, we also introduce an additional experiment where the potential benefits of assessing FCD quality are quantified through the gradual introduction of artificial noise in the data. The obtained results in terms of generalization error are then used to briefly discuss the insights provided by *Yuki-san*.

A. Experimental Setup

Table II summarizes the hyperparameter setting used in the experiments. This configuration was chosen in the interest of applying predictive modeling methodologies to the datasets.

1) *Optimal Sampling Rate*: The parameter δ_{opt} was set to 15. This means that datasets with an average sampling rate below or equal to 15 seconds are considered to be of optimal value in terms of granularity (section III-A).

2) *Parts of Day*: Regarding the *MiTC* indicator (Section III-C) we split each day into 4 equally sized parts ($D = 4$).

3) *Spatial Coverage Setup*: In section III-D, the city map is decomposed into a grid of 500×500 cells ($nblocks = 500$). The relevance of the city center is set to 3 (Υ_{cc}). Υ_{min} is set to 2, which means that the influence propagation algorithm (Algorithm 2) is triggered for grid cell with relevance below 2 (with a propagation factor η is set to 0.3).

TABLE II
HYPERPARAMETER SETTING USED IN THE EXPERIMENTS

Hyperparameter	Value(s)
δ_{opt} in III-A	15
$erfc$ in III-A	$a = 0.01$
$erfc$ in III-B	$a = 0.02$
D in III-C	4
$nblocks$ in III-D	500
Υ_{cc} in III-D	3
Υ_{min} in III-D	2
η in III-D	0.3
$erfc$ in III-E	$a = 0.2$
Ψ in III-F	20
$erfc$ in III-G	$a = 0.025$
ξ in III-E	5
$cc_{nanjing}$	$lat = 32.05, lon = 118.76667$
$cc_{sanfrancisco}$	$lat = 37.78333, lon = -122.41667$

4) *Missing Data Indicator Setup*: The penalty factors Ω_d and Ω_t in Section III-E are set according to equations 12 and 13, respectively. Essentially, we smooth the value of P according to dev and $mdur$ with a maximum value of 1.3 and 1.5, respectively.

$$\Omega_d(dev) = \begin{cases} \frac{dev}{500} + 1, & \text{if } dev < 150 \\ 1.3, & \text{otherwise} \end{cases} \quad (12)$$

$$\Omega_d(mdur) = \begin{cases} \frac{mdur}{1800} + 1, & \text{if } mdur < 150 \\ 1.5, & \text{otherwise} \end{cases} \quad (13)$$

Where dev is the Inter-Quartile Range of the black-holes duration and $mdur$ is the median duration of a trip.

5) *Erfc Setup*: We resorted to the $erfc$ function to normalize some results into a standardized range of values. The general form for $erfc$ is:

$$erfc(x) = \frac{2 \cdot a}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt \quad (14)$$

We used different parameter values for a according to the application. For example, in Section III-B we consider $a = 0.02$ which yields a optimal value for dataset with a timespan of one year.

6) *Map-Matching Approach*: Algorithm 5 formalizes our approach to the map-matching problem. Essentially, we created an *ad-hoc* procedure to estimate the average GPS error measure of each trip $-e_t$. Moreover, we use a Monte Carlo approximation to estimate e_t , using $nreps$ repetitions. For each repetition we proceed as follows: We pick a random points t_i along with its next $s \cdot L_t - 1$ GPS traces, where L_t is the number of GPS traces comprising trip t and s is the sample size. This yields a contiguous random sample of t, t_s . We then extract the candidate roads from the *DRN* for t_s . A road is considered candidate if it lies inside the bounding box of t . Then, we compute the Haversine distance d_i of each GPS trace t_i to three points in each candidate road: the initial point r_1 , the mean point r_2 and the end point r_3 . The road that minimizes that distance is the one we choose as the one the vehicle is traversing. The error measurement of the GPS trace t_i is its Haversine distance to that road.

Algorithm 5 Map-Matching Algorithm

```

1: Input: GPS traces of trip  $t$ ,  $DRN$ 
2: Output: Distance of each GPS trace  $t_i$ ,  $t_i \in t$ , to the
   respective predicted road.
3:  $nreps \leftarrow$  Monte Carlo repetitions
4:  $L_t \leftarrow$  no. of GPS traces in  $t$ 
5: for each rep in  $nreps$  do
6:   procedure SAMPLE( $t$ )
7:      $t_s \leftarrow$  contiguous random sample of  $t_i$  of size  $s \cdot L_t$ 
8:   procedure CANDIDATEROADS( $t$ ,  $DRN$ )
9:     Return  $R_c$ : candidate roads for  $t$ 
10:  for each GPS trace  $t_i$  in  $t_s$  do
11:    for each  $r$  in  $R_c$ : do
12:       $r_1 \leftarrow$  Initial point of  $r$ 
13:       $r_2 \leftarrow$  Mean point of  $r$ 
14:       $r_3 \leftarrow$  End point of  $r$ 
15:       $d_i^r = \min(\text{Dist}(t_i, r_1), \text{Dist}(t_i, r_2), \text{Dist}(t_i, r_3))$ 
16:       $r_i^s = \{r \in R_c : d_i^r = \min(d_i^r)\}$ 
17:       $e_i \leftarrow \text{Dist}(t_i, r_i^s)$ 
18:       $e_t^{rep} \leftarrow \text{mean}(e_i), \forall i \in t_i \in t_s$ 
19:  $e_t = \text{mean}(e_t^{rep}), \forall rep \in nreps$ 

```

The error measurement of t in the Monte Carlo repetition (e_t^{rep}) is estimated by averaging each e_i , $\forall t_i \in t_s$. Using the Monte Carlo approximation we average all e_t across all repetitions to estimate the e_t , the error measurement of trip t . This Monte Carlo approximation procedure is important in big datasets to keep the computations tractable, providing a way to analyze all the trips for an Accuracy measure. In our experiments, $nreps$ is set to 10, while s is set to 5%.

This is a simple and highly heuristic approach to the map-matching problem. We remark that any more sophisticated approach to map-matching can be used in the proposed methodology. The Acc function used in the Section III-G is formalized in the Equation 15:

$$Acc(e) = \begin{cases} 1, & \text{if } e < 15 \\ -\frac{3(e-15)}{100} + 1, & \text{if } 15 < e \leq 45 \\ \text{erfc}(e), & \text{otherwise} \end{cases} \quad (15)$$

The a parameter for the $erfc$ is set to 0.025 (see Equation 14).

7) *Other Parameters:* The estimated speed Ψ is set to 20 mph. Note that this estimate also incorporates eventual stops (e.g. traffic lights). We used OpenStreetMap [43] as DRN.

B. Setup of Data Mining Experiment I

The first studied experiment (DM-I) is Time Travel Prediction (TTP). The objective is to predict the duration of a given trip.

We address this task using the predictive framework provided by Hoch [44]. The presented technique is the winning solution of a discovery challenge for the TTP task hosted by the well-known Kaggle platform. The method is based on an ensemble of regression models. The experimental setup is the

TABLE III
RESULTS FROM THE DATA MINING EXPERIMENT I

Error	Nanjing	San Francisco
RMSE	1619.34	624.96
MAE	700.7	455.2

same one proposed in [44]. The basic information used about each trip is the sequential geographic positions as well as the corresponding timestamps. The position of the city center is also used to derive some attributes related to the positioning of the trip w.r.t. the downtown of the city.

As pre-processing, we also excluded trips with less than 4 GPS traces, for numerical computation issues. However, we did not cut very long trips like was initially suggested by Hoch [44]. We do it so in order to minimize to pre-processing to the one strictly necessary to conduct the experiments. The current position of a vehicle was estimated by randomly cutting the full trajectories using a uniform distribution.

The performance of the method was estimated using the root mean squared error (RMSE) and mean absolute error (MAE) on a 5-fold cross validation procedure.

C. Setup Data Mining Experiment II

The second data mining experiment (DM-II) is related to an O-D flow estimation (see [41] to know more about this topic).

We tackle this task using popular approaches in the literature. We start by making a spatio-temporal discretization and then fitting an ARIMA model to each decomposed block. This idea is formalized according to the following steps: (i) spatial discretization: we decompose the data into several equally sized cells. Unlike in Section III-D where we fixed the number of grid cells, here we fix the size of each cell, so we can draw comparisons between the two case studies. We fixed each cell to be 500×500 meters; (ii) Top- K O-D pairs: we extract the origin and destination cells from each trips and select the top- K pairs, i.e. the pairs with most trips; (iii) temporal discretization: each top- K O-D pair is discretized by hour. This yields hourly time series for each selected O-D pair. As pre-processing step we discarded near-zero variance time series. K was set to 10.

The method was evaluated using the RMSE in a prequential evaluation strategy (e.g. [42]). We take into account the different time-spans of each dataset. In SF, we start testing in the last day (i.e. last 24 hours for testing) whereas in NJ we used the last 4 hours.

D. Results

The results from the application of *Yuki-san* are presented in Fig. 3. These are shown as radar charts. In each graphic, the larger the magnitude of a variable, the larger the respective indicator value is. Table III and Fig. 2 display the results obtained from Data Mining Experiments I and II, respectively.

1) *Discussion:* The indicator G is higher for NJ. Even though the raw value of NJ is almost half the one of SF, the nonlinearity of G yields a smoother result.

The **MaTC** represents the indicator with second highest discrepancy between case studies: 0 to NJ and 0.56 of SF.

Fig. 2. MSE bias-variance decomposition for DM-II, averaged across the discretized cells.

Fig. 3. *Yuki-san* results for NJ (left) and SF (right).

These values are mostly driven by the time-span of the datasets as NJ is only 23 hours long. In effect, we hypothesize that the short time-span presents some generalization problems for some of the data mining tasks. This is empirically demonstrated in Table III, where SF shows significantly lower errors than NJ in the experiment. TTP is a task that leverages the spatio-temporal coverage of the dataset. Naturally, this capitalizes on the knowledge of typical destinations from the current position at the present time.

Another relevant insight regards the spatial coverage indicator (SC): it marks 0.653 for NJ, around the double of the value for SF, 0.318. Relative to SF, NJ covers more ground in a less time-span, though it has significantly more operating vehicles (according to Table I). The combination of the values of MaTC and SC may increase the sensitivity to outliers of concept, i.e. $p(y|x)$. Due to the lack of diverse observations in terms of timespan, the outliers are taken as a representative part of the joint distribution that we are trying to model (e.g. a soccer match occurring in some specific days which largely increases the typically low cruising times on given OD pair). This may lead to overfitting issues. Yet, the data in SF is concentrated in a smaller part of the town (according to the results from SC). A low SC leads to a more sparse feature space in space-related features. Such sparsity may lead to outliers in $p(x)$ which, in many cases, must be addressed before modeling tasks.

Fig. 2 illustrates the poor (overfit) model obtained to NJ per opposition to a relatively good generalization provided to SF. The main reason to do it so lies on the base learners chosen were ensembles of decision trees. Decision trees are not sensitive to outliers in $p(x)$ since the splitting criteria are based on the proportion of samples that fall in each newly created leaf rather than on the size of the feature space [45]. The small time-span of NJ obliged to a setup where the

TABLE IV
INTERMEDIATE RESULTS OF MD

Measure	Description	NJ	SF
r_{bh}	average black-holes per trip	1.21	0.45
bh	median black-hole duration	598 secs.	448 secs.
δ_G	global sampling rate	60 secs.	34 secs.

TABLE V
THE AVERAGE ACCURACY ERROR ACROSS TRIPS – INTERMEDIATE RESULTS OF A

Measure	NJ	SF
Median	40.3	15.4
Mean	15042.8	66.8

training of the model for O-D estimation and its testing to happen in different parts of the day. Because of this issue, the learner is unable to do an adequate generalization and capture the intra-day seasonality.

The results of the MD indicator are comparable, though SF shows a slight upper-hand. Although NJ has a better sampling rate compared to SF, it also has higher average black-holes per trip (as presented in the intermediate results depicted in Table IV).

The results of the A indicator are significant: 0.24 to NJ and 0.98 to SF. SF has a noticeable better GPS measurement accuracy relative to NJ, with a median error of around 15 meters, compared to approximately 40 meters in NJ (c.f. Table V for the intermediate results). Furthermore, the difference between the mean as median values suggest that there are severe outliers in NJ.

As discussed above, we were also able to link the results of *Yuki-san* to the results of DM-I and DM-II. These comparisons provided a way to reason about the differences in performance of the same model in different case studies. Moreover, we connected these differences in performance to the properties of each dataset, uncovered by the application of *Yuki-san*. The authors would like to stress that the obtained results demonstrate empirically the potential that *Yuki-san* has to assess the potential knowledge broadcast by a given FCD in a fully automated way and without conducting any manual data mining procedure. The carried out experiments suggest that NJ need pre-processing while SF can be used as is for data mining, specifically the tasks studied in this paper.

E. Adding Synthetic Noise

To further our argument relative to the importance of FCD quality, we artificially added different levels of noise to the data. Specifically we removed a percentage of random GPS traces from the training data – this process affects the FCD quality in multiple aspects, mainly the spatial and temporal coverage, and missing data. In the interest of conciseness we focused this analysis on the TTP task (DM-I).

The results of this analysis are shown in Fig. 4. This graphic shows that the generalization error grows consistently with the amount of artificially uncertainty introduced, for both cases. This is an aspect that can be captured by *Yuki-san* before applying any predictive modeling process.

Fig. 4. TTP with different levels of missing data.

VI. POTENTIAL DEPLOYMENT

This Section contains a brief discussion of the potential applications of Yuki-san, followed by some guidelines for scenario-oriented parametrization.

A. Scope and Impact

Reliable FCD is at the backbone of operations planning of ITS. The proposed framework could be useful in several applications in real-world and in this section we describe three potential deployment scenarios.

The first scenario is the definition of the **market value** of FCD. A generalized data market is emerging. For example, in telecommunications call detail records (CDR) are widely used as mobility probes [46]. FCD is as detailed and even more pervasive than CDR data. In a short future, FCD will potentially be commercialized, e.g. as a feed, where its price is set by volume. In this context, users or clients can leverage Yuki-san as a tool to quantify the quality of these FCD sources.

Another potential application is **troubleshooting** through the detection of malfunctions on the sensors of FCD providers.

Finally, Yuki-san can be useful for **saving communication broadband**. This can be accomplished by pruning unreliable or irrelevant FCD from the workflow of, for example, traffic status visualization tools, or traffic control actions centers. Besides enhancing the reliability of FCD sources, this process leads to decreased storage demands and computational power required for predictive analytics operations.

B. Guidelines for Scenario-Oriented Configuration

While some rules of thumb can be applied, the best configuration for the parameter setting is dependent on the application.

In our experiments the value for the optimal sampling rate (δ_{opt}) was set to 15 seconds. This configuration is geared towards the studied tasks, i.e. O-D flow estimation and TTP. However, there are scenarios where such set up would be inadequate. For instance map-matching, where high spatio-temporal precision is fundamental (see Figure 1). In this case, a lower value of δ_{opt} would be more appropriate.

Another example is the parameter set up for the MaTC indicator in section III-B. We selected the parameter for the erfc function (see Table II) so that one year of data renders an

optimal value. Again, for the tasks we studied in this paper, covering a large period of data is important to capture seasonal effects. However, there are tasks less sensitive to these issues, e.g. map-matching. In the latter, more emphasis is placed on the spatial coverage of the data. For example, the FCD source for the NJ case study is able to achieve a reasonable SC score with a single day of data.

VII. FINAL REMARKS

This paper presents *Yuki-san*: an automatic framework for multi-criteria evaluation of FCD quality. As acquiring FCD grows cheaper, its quality assessment becomes an increasingly important task. *Yuki-san* is based on a set of statistical indicators that cover different properties of the data, from its granularity, temporal and spatial coverage to its robustness, reliability and accuracy.

We applied *Yuki-san* to two real-world case studies of data collected by taxi fleets running in Nanjing and San Francisco to validate the proposed method. We evaluate the insights provided by this indicator set by conducting two data mining experiments regarding common supervised learning problems in mobility tasks: travel time prediction and O-D flow prediction. The results clearly illustrated the potential that *Yuki-san* has to uncover value in FCD sources in a fully automated fashion. Specifically, the analysis of the results enabled the discovery of underfitting/overfitting potential issues before conducting any supervised learning task. As future work, we intend to deploy *Yuki-san* in real-world applications to select reliable FCD sources in real-time on modern traffic management centers.

REFERENCES

- [1] F. Giannotti *et al.*, "Unveiling the complexity of human mobility by querying and mining massive trajectory data," *VLDB J. Int. J. Very Large Data Bases*, vol. 20, no. 5, pp. 695–719, 2011.
- [2] D. Jorge and G. Correia, "Carsharing systems demand estimation and defined operations: A literature review," *Eur. J. Transport Infrastruct. Res.*, vol. 13, no. 3, pp. 201–220, 2013.
- [3] L. Moreira-Matias, J. Mendes-Moreira, J. Freire de Sousa, and J. Gama, "Improving mass transit operations by using AVL-based systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1636–1653, Aug. 2015.
- [4] P. S. Castro, D. Zhang, C. Chen, S. Li, and G. Pan, "From taxi GPS traces to social and community dynamics: A survey," *ACM Comput. Surv.*, vol. 46, no. 2, p. 17, 2013.
- [5] L. Cao and J. Krumm, "From GPS traces to a routable road map," in *Proc. GIS 7th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, New York, NY, USA, 2009, pp. 3–12.
- [6] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "On predicting the taxi-passenger demand: A real-time approach," in *Proc. Portuguese Conf. Artif. Intell.* Berlin, Germany: Springer, Sep. 2013, pp. 54–65.
- [7] L. Moreira-Matias, J. Mendes-Moreira, M. Ferreira, J. Gama, and L. Damas, "An online learning framework for predicting the taxi stand's profitability," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 2009–2014.
- [8] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Map-matching for low-sampling-rate GPS trajectories," in *Proc. 17th ACM SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, New York, NY, USA, 2009, pp. 352–361.
- [9] T. Hunter, P. Abbeel, and A. Bayen, "The path inference filter: Model-based low-latency map matching of probe vehicle data," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 507–529, Apr. 2014.
- [10] Y. Wang, Y. Zhu, Z. He, Y. Yue, and Q. Li, "Challenges and opportunities in exploiting large-scale GPS probe data," HP Lab., Palo Alto, CA, USA, Tech. Rep. HPL-2011-109, 2011.

- [11] Z. Zheng, S. Rasouli, and H. Timmermans, "Evaluating the accuracy of GPS-based taxi trajectory records," *Procedia Environ. Sci.*, vol. 22, pp. 186–198, Aug. 2014.
- [12] Datamall. (Jan. 2015). *Nanjing*. Accessed: Apr. 2015. [Online]. Available: <http://dublinked.com/datastore/datasets/dataset-254.php> and <http://www.datatang.com/data/44060>
- [13] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. (2009). *CRAWDAD Dataset Eptf/Mobility (V. 2009-02-24)*. [Online]. Available: <http://crawdad.org/eptf/mobility/20090224>
- [14] A. Muschalle, F. Stahl, A. Löser, and G. Vossen, "Pricing approaches for data markets," in *Proc. Int. Workshop Bus. Intell. Real-Time Enterprise*. Berlin, Germany: Springer, 2012, pp. 129–144.
- [15] A. Hoffleitner, R. Herring, and A. Bayen, "Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning," *Transp. Res. B, Methodol.*, vol. 46, no. 9, pp. 1097–1122, 2012.
- [16] A. Hoffleitner, R. Herring, P. Abbeel, and A. Bayen, "Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1679–1693, Dec. 2012.
- [17] L. Moreira-Matias and F. Alesiani, "Drift3Flow: Freeway-incident prediction using real-time learning," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2015, pp. 566–571.
- [18] D. B. Work, O.-P. Tossavainen, S. Blandin, A. M. Bayen, T. Iwuchukwu, and K. Tracton, "An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices," in *Proc. 47th IEEE Conf. Decision Control (CDC)*, Dec. 2008, pp. 5062–5068.
- [19] J. Miller, S.-I. Kim, M. Ali, and T. Menard, "Determining time to traverse road sections based on mapping discrete GPS vehicle data to continuous flows," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2010, pp. 615–620.
- [20] J. C. Herrera, D. B. Work, R. Herring, X. Ban, Q. Jacobson, and A. M. Bayen, "Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century field experiment," *Transp. Res. C, Emerg. Technol.*, vol. 18, no. 4, pp. 568–583, Aug. 2010.
- [21] Q. Li, Z. Zeng, B. Yang, and T. Zhang, "Hierarchical route planning based on taxi GPS-trajectories," in *Proc. 17th Int. Conf. Geoinform.*, Aug. 2009, pp. 1–5.
- [22] D. Zhang, N. Li, Z.-H. Zhou, C. Chen, L. Sun, and S. Li, "iBAT: Detecting anomalous taxi trajectories from GPS traces," in *Proc. ACM 13th Int. Conf. Ubiquitous Comput.*, 2011, pp. 99–108.
- [23] A. Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang, and Z. Xu, "Destination prediction by sub-trajectory synthesis and privacy protection against such prediction," in *Proc. IEEE 29th Int. Conf. Data Eng. (ICDE)*, Apr. 2013, pp. 254–265.
- [24] B. R. Hellinga and L. Fu, "Reducing bias in probe-based arterial link travel time estimates," *Transp. Res. C, Emerg. Technol.*, vol. 10, no. 4, pp. 257–273, 2002.
- [25] Q. Ou, J. van Lint, and S. Hoogendoorn, "Piecewise inverse speed correction by using individual travel times," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2049, pp. 92–102, Sep. 2008.
- [26] Q. Ou, R. Bertini, J. W. C. van Lint, and S. Hoogendoorn, "A theoretical framework for traffic speed estimation by fusing low-resolution probe vehicle data," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 3, pp. 747–756, Sep. 2011.
- [27] T. Miwa, D. Kiuchi, T. Yamamoto, and T. Morikawa, "Development of map matching algorithm for low frequency probe data," *Transp. Res. C, Emerg. Technol.*, vol. 22, pp. 132–145, Jun. 2012.
- [28] M. Rahmani and H. N. Koutsopoulos, "Path inference from sparse floating car data for urban networks," *Transp. Res. C, Emerg. Technol.*, vol. 30, pp. 41–54, May 2013.
- [29] J. W. C. Van Lint and S. Hoogendoorn, "A robust and efficient method for fusing heterogeneous data from traffic sensors on freeways," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 25, no. 8, pp. 596–612, 2010.
- [30] M. Treiber, A. Kesting, and R. E. Wilson, "Reconstructing the traffic state by fusion of heterogeneous data," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 26, no. 6, pp. 408–419, 2011.
- [31] D. Sun, C. Zhang, L. Zhang, F. Chen, and Z. Peng, "Urban travel behavior analyses and route prediction based on floating car data," *Transp. Lett.*, vol. 6, no. 3, pp. 118–125, 2014.
- [32] T. Morikawa and T. Miwa, "Preliminary analysis on dynamic route choice behavior: Using probe-vehicle data," *J. Adv. Transp. Banner*, vol. 40, no. 2, pp. 140–163, 2006.
- [33] P. Sanders and D. Schultes, "Engineering highway hierarchies," in *Proc. Eur. Symp. Algorithms*. Berlin, Germany: Springer, 2006, pp. 804–816.
- [34] T. Feng and H. J. P. Timmermans, "Extracting activity-travel diaries from GPS data: Towards integrated semi-automatic imputation," *Procedia Environ. Sci.*, vol. 22, pp. 178–185, Aug. 2014.
- [35] T. Djukic, J. van Lint, and S. Hoogendoorn, "Application of principal component analysis to predict dynamic origin-destination matrices," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2283, no. 1, pp. 81–89, 2012.
- [36] E. Ospina, F. Moreno, and I. A. Uribe, "Using criteria reconstruction for low-sampling trajectories as a tool for analytics," *Procedia Comput. Sci.*, vol. 51, pp. 366–373, Nov. 2015.
- [37] L. Montero, M. Pacheco, J. Barceló, S. Homoceanu, and J. Casanovas, "A case study on cooperative car data for traffic state estimation in an urban network," in *Proc. TRB 95th Annu. Meeting Final Program*, 2016, pp. 1–17.
- [38] Z. Zhang, D. Yang, T. Zhang, Q. He, and X. Lian, "A study on the method for cleaning and repairing the probe vehicle data," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 419–427, Mar. 2013.
- [39] A. D. Patire, M. Wright, B. Prodhomme, and A. M. Bayen, "How much GPS data do we need?" *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 325–342, Sep. 2015.
- [40] J. Khiari, L. Moreira-Matias, V. Cerqueira, and O. Cats, "Automated setting of bus schedule coverage using unsupervised machine learning," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2016, pp. 552–564.
- [41] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Time-evolving O-D matrix estimation using high-speed GPS data streams," *Expert Syst. Appl.*, vol. 44, pp. 275–288, Feb. 2016.
- [42] L. Moreira-Matias, J. Gama, and J. Mendes-Moreira, "Concept neurons—Handling drift issues for real-time industrial data mining," in *Machine Learning and Knowledge Discovery in Databases*. Riva del Garda, Italy: Springer, 2016, pp. 96–111.
- [43] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervas. Comput.*, vol. 7, no. 4, pp. 12–18, Oct. 2008, doi: [10.1109/MPRV.2008.80](https://doi.org/10.1109/MPRV.2008.80).
- [44] T. Hoch, "An ensemble learning approach for the kaggle taxi travel time prediction challenge," in *Proc. ECML-PKDD-DC*, 2015, pp. 52–62.
- [45] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.
- [46] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in Rome," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 141–151, Mar. 2011.

Vitor Cerqueira received the B.Sc. degree in applied mathematics from the Faculty of Sciences, University of Porto (U. Porto), in 2012, and the M.Sc. degree in data analytics from the Faculty of Economics, U. Porto, in 2014. He is currently pursuing the Ph.D. degree in machine learning with U. Porto.

He is currently a Research Fellow with LIAAD, a laboratory for artificial intelligence and decision support systems. His main research topic is related to learning in time series.

Luis Moreira-Matias received the M.Sc. degree in informatics engineering and the Ph.D. degree in machine learning from University of Porto in 2009 and 2015, respectively. He won the International Data Mining Competition held at the Research Summer School, TU Dortmund, in 2012. He served for the Program Committee of multiple high-impact research venues such as KDD, AAAI, IEEE TKDE, ESWA, ECML/PKDD, and KAIS, among others.

He is currently a Senior Researcher with NEC Laboratories Europe. His interests include machine learning, data mining, and predictive analytics in general. He has authored 40+ high-impact peer-reviewed publications on related topics. He was invited to give keynotes around the globe, ranging locations from Brisbane, Australia, to Las Palmas, Spain.

Jihed Khiari received the M.Sc. degree in computer science, with a specialization in machine learning, from University of Carthage in 2015. She is currently a Software Engineer with the Intelligent Transportation Systems Group, NEC Laboratories Europe, Heidelberg, Germany. Her main research interests include data analytics and machine learning techniques applied in the intelligent transportation systems field.

Hans van Lint received the M.Sc. degree in civil engineering informatics and the Ph.D. degree in transportation from Delft University of Technology (DUT) in 1997 and 2004, respectively. He was an Information Analyst and a Transport Engineer with various organizations. He was appointed as the Anthonie van Leeuwenhoek Full Professor (an honor reserved for only a few young talented scientists and educators) by the Executive Board, DUT, in 2013. He has co-authored 55+ peer-reviewed journal articles. His expertise lies on the interface between traffic flow theory and simulation, data analytics, and machine learning techniques. He has co-promoted eight Ph.D. students and has currently 10 Ph.D. students, five post-doctoral students, and two programmers under supervision in his lab (dittlab.tudelft.nl). He serves as an AE for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, and is active in many international projects and collaborations.