Improving VANET Simulation with Calibrated Vehicular Mobility Traces

Clayson Celes, Fabrício A. Silva, Azzedine Boukerche, *Fellow, IEEE*,
Rossana M. C. Andrade, and Antonio A. F. Loureiro

Abstract—Simulation is the most frequently adopted approach for evaluating protocols and algorithms for Vehicular Ad hoc Networks (VANETs) and Delay-Tolerant Networks (DTNs). Usually, simulation tools use mobility traces to build the network topology based on the existing contacts between mobile nodes. However, quality of the traces, in terms of spatial and temporal granularity of each entry in the logfile, is a key factor that impacts the network topology directly. Therefore, the reliability of the results depends strongly on the accurate representation of the real network topology by the vehicular mobility model. We show that five widely adopted existing real vehicular mobility traces present gaps, leading to fallible outcomes. In this work, we propose a solution to fill those gaps, leading to more fine-grained traces, which lead to more trustworthy simulation results. We propose and evaluate a data-based solution using clustering algorithms to fill the gaps of real-world traces. In addition, we also present the evaluation results that compare the communication graph of the original and the calibrated traces using network metrics. The results reveal that the gaps do indeed induce network topologies differing from reality, decreasing the quality of the evaluation results. To contribute to the research community, we have made the calibrated traces publicly available, so that other researchers may adopt them to improve their evaluation results.

Index Terms—Vehicular mobility traces, calibration, vehicular ad-hoc networks, performance evaluation, simulation

1 Introduction

Simulation is the most frequently adopted approach for evaluating protocols and algorithms for Vehicular Ad hoc Networks (VANETs) [1], [2]. The performance evaluation of VANET solutions presents a considerable challenge to researchers, given the particular characteristics of this kind of network, such as its highly dynamic topology and large-scale nature. Conducting real experiments using ordinary vehicles is a very expensive and time-consuming approach, particularly when a large-scale evaluation is required. In addition, there is no publicly available, large-scale testbed that can be readily used by researchers. Moreover, it is unlikely that a large-scale testbed will be available in the near future, due to involved deployment and maintenance costs. Simulation, on the other hand, is a cost-effective, large-scale, timely approach widely adopted by researchers. However, the reliability of the simulation results depends on the vehicular mobility models to represent the network topology.

The adopted vehicular mobility model plays a key role on the reliability of the simulation results [3], [4], [5], [6]. Existing simulation tools use mobility models to build scenarios in which vehicles move and communicate with each other. The mobility model is responsible for determining the position of vehicles at each moment in time; this information is used to build the network topology. In other words, unrealistic mobility models lead to unrealistic network topologies, and, therefore, to unreliable evaluation results [7]. Hence, it is very important to adopt realistic vehicular mobility models when evaluating VANET solutions.

One possible strategy for achieving this goal is to use records of real vehicular positions over time (i.e., traces). The availability of traces in recent years has led the research community to investigate methods for modeling vehicles and their connectivity. To this end, some studies started characterizing the mobility traces. Amici et al. [8] characterized the taxi trace from Rome, and analyze an epidemic dissemination protocol using this trace as the mobility model. The studies presented in [9], [10], [11] characterized the network topology and connectivity metrics of the taxi trace from San Francisco. Furthermore, the taxi trace of Shanghai was used to study mobility patterns [12], [13], [14], [15], network topology, and connectivity metrics [16], [17], [18]. Similarly, the trace from Beijing was also explored in mobility characterization studies [19], [20].

Those characterizations and analyses have led to important findings about mobility patterns, helping to define novel solutions related to communication and dissemination protocols for VANETs and DTNs. However, most VANET and DTN performance evaluations rely on vehicle contacts. It turns out that the network graph representing those contacts is built based on the mobility traces, which may present gaps in space and time (i.e., long periods or

- C. Celes and A.A.F. Loureiro are with the Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, MG 31270-901, Brazil. E-mail: {claysonceles, loureiro}@dcc.ufmg.br.
- F. Silva is with the Department of Informatics, Federal University of Viçosa-Campus Florestal, Florestal, MG 35690-000, Brazil. E-mail: fabricio.asilva@ufv.br.
- A. Boukerche is with the Department of Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada. E-mail: boukerch@site.uottawa.ca.
- R.M.C. Andrade is with the Department of Computer Science, Federal University of Ceara, Fortaleza, CE 60440-554, Brazil. E-mail: rossana@ufc.br.

Manuscript received 21 June 2016; revised 17 Mar. 2017; accepted 22 Mar. 2017. Date of publication 12 Apr. 2017; date of current version 1 Nov. 2017. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TMC.2017.2690636

distances between two consecutive entries of a given vehicle). Furthermore, such gaps lead to missing contacts, since all interactions that might have happened among vehicles during successive entries will not be present in the trace. Consequently, an incomplete graph denoting the network topology will not correctly represent the real contacts among vehicles. In other words, the existence of gaps leads to contact graphs that differ from reality. Hence, it turns out that finding and eliminating such gaps to build a high-fidelity mobility model is a key aspect for guaranteeing the reliability of the results. Nevertheless, this problem is not tackled properly in the literature, since most solutions focus on adding a straight path between two sparse points, instead of building a fine-grained trajectory between them.

In this work, we find and fill existing gaps appropriately by performing a process referred to as calibration [21]. Calibration consists of filling the gaps in raw mobility traces, leading to fine-grained traces. First, we demonstrate the existence of gaps in available traces. After that, we propose and evaluate a cluster-based solution to fill the gaps, following the methodology proposed in [22]. Our solution relies on the existing trajectory points, obtained from the trace itself, that are organized into clusters to represent anchor points used in the calibration. Therefore, our approach is flexible enough to be adopted in different real traces, since there is no need for looking at a map or any further information. In fact, we demonstrate this by applying our solution to calibrate five existing, widely adopted taxi traces in different scenarios [23], [24], [25], [26], [27]. We consider taxi traces in our study because they are real, publicly available, and widely adopted in the literature. However, our solution is general enough to be applied to any vehicular mobility trace. The results reveal that the gaps of indeed lead to different network topology graphs, directly affecting the results of the performance evaluation. To cooperate with the research community, we made the calibrated traces publicly available at [28].

This work is an extended version of our study published in [29]. We extend our previous work by making the following improvements: (i) we conduct experiments to validate the efficiency of the proposed calibration method; (ii) we include a detailed discussion on how our calibration method differs from the state-of-the-art; (iii) we compare our proposal against the one described in [22]; (iv) we describe the computational complexity of our proposal in comparison with the existing ones; (v) we apply our methodology to two new datasets (Beijing and Shenzhen); (vi) we add new results comparing the communication graph of the original and the calibrated traces (i.e., after the gaps were filled); (vii) we conduct simulation studies using the Network Simulator (NS-3) to assess the impact of calibration when the IEEE 802.11p protocol is used; and (viii) we make the calibrated traces available in the NS-3 format, which will facilitate their adoption by other researchers.

The key contributions of this paper are summarized as follows:

- We develop a solution to reduce, or even eliminate, gaps in real-world vehicular traces. Our solution for filling the gaps in vehicular mobility traces is divided into two stages. The first extracts a reference system from the vehicles' historical GPS trajectory dataset. The second stage applies a calibration method, using a subset of points of the previously built reference system;
- We validate our proposed solution by intentionally adding gaps to a fine-grained trace and comparing the calibrated results with those of the original. The results reveal that our calibration method leads to calibrated trajectories that are near the original ones;
- We compare our solution and the one proposed in [22]. The results reveal that our calibration approach accurately fills gaps in vehicular mobility traces, obtaining spatio-temporal results better than the baseline;
- We analyze how the gaps affect the communication network. For this, we show that existing gaps in the original traces available in the literature lead to unrealistic network topologies, which are improved with our calibration method;
- We conduct simulation experiments to assess the impact of applying the calibrated trace to a real vehicular network protocol in a realistic VANET scenario (IEEE 802.11p);
- We find and eliminate gaps of five widely adopted real vehicular mobility traces. To contribute to the research community, the calibrated traces are publicly available to other researchers, who can use them in their research studies.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 presents a detailed description of the vehicular mobility traces used in this paper. Section 4 offers some essential background information, and confirms the existence of gaps in those real mobility traces. Section 5 introduces our calibration method for filling the gap. Section 6 compares our proposal to the state-of-the-art solution. Section 7 discusses the evaluation results comparing the communication graph of the original and the calibrated traces. Section 8 analyzes the impact of calibrated traces on vehicular networking, comparing the calibrated traces and the original ones in a realistic simulation scenario. Finally, Section 9 concludes this work.

2 Related Work

The actual movement of vehicles is inherently a continuous-time function, but it is sampled at a discrete time due to different issues such as storage limitation and the ease in working with discrete data, including the availability of techniques to work in the discrete domain. Moreover, the sampling rate is generally low, and consequently, details of the movement are lost. There are a number of studies on how to reconstruct the vehicle's movement from trajectories sampled at a low rate. In this section, we present the related works that have focused on the techniques used for this purpose, highlighting their strengths and weaknesses.

Before reconstructing a trajectory, depending on how we want to reconstruct it and the quality of the data, we need to deal with many preprocessing issues [30] such as filtering to remove invalid points; trajectory compression to reduce the size of a trajectory while maintaining its significant portion; and map matching to associate each trajectory to a corresponding projection in the legitimate road network. Those preprocessing techniques have a fundamental role in the treatment of raw trajectories, but they are not enough to transform raw trajectories into meaningful trajectories. That is, we need to perform other techniques to obtain an

TABLE 1
Computational Complexity to Create a Reference System for Each Anchor Point Type

Algorithm	Resource	Input	Complexity
Space-based	None	n-dimensional grid	$O(n)$
Data-based	Trajectories	n-sample points	$O(1)$
PoI-based	PoI dataset	n-PoI points	$O(n^2)$
Feature-based	Trajectories	n-sample points and k is the number of reported points	$O(n\sqrt{n} + nk)$

TABLE 2
Time Complexity of Proposed Calibration Methods in [22]

Algorithm	Complexity
Geometry-based	$O(N_T N_a log N_a)$, where N_a is the number of anchor points close to the gap and N_T is the size of the trajectory.
Model-based	$O(N_T \lvert PP\rvert^2)$, where N_T is the size of the trajectory and $\lvert PP\rvert$ is the average number of paths connecting two consecutive anchor points of calibrated trajectories.

approximate form of the actual movement, since the methods described above do not directly address sampling issues.

The straightforward idea for reconstructing vehicle trajectories is to apply interpolation in between consecutive records. In the literature, many interpolation methods have been proposed for different applications, such as linear interpolation [31], nearest-neighbor interpolation [32], and piecewise cubic spline interpolation [33]. The linear interpolation [31], when applied to trajectories, computes straight lines between two consecutive records. However, this method is not suitable for certain urban scenarios with curved paths between each pair of consecutive records, since drivers do not always travel in a straight line. Hoteit et al. [34] evaluated the difference between real human trajectories and the ones obtained through cellphone data using these interpolation methods, showing that trajectories obtained from interpolation are far from the actual path. To overcome this problem, Liu et al. [35] proposed an interpolation method based on the shortest path between consecutive GPS points using the road network. However, the assumption of the shortest path between points may not be sufficient, since it does not represent vehicles' movements. Hoque et al. [11] interpolated adjacent points with the objective of finding an intermediary point between them. To this end, they averaged samples one minute backward and one minute forward to estimate the position of a mobile entity in each period. As previously mentioned, this simple approach works when the mobile entity travels following a straight line. However, it fails when the entity turns its direction at an intersection, a very common mobility pattern when it comes to vehicles.

With the objective of reconstructing trajectories more accurately, Su et al. [22] introduced a methodology composed of two components: a reference system and a calibration method. The reference system was built from a set of anchor points independent of the current trajectory. The calibration method used the reference system to find points to be inserted along the trajectory, making it more complete. The authors evaluated and presented results of different strategies of their methodology, as discussed in the following.

The proposed model relies on four types of anchor points obtained from different kinds of external resources: space-based, data-based, PoI-based, and feature-based anchors. Space-based anchors are centroid points of the cells retrieved from dividing the map into a grid. Data-based anchors are points from historical trajectories. PoI-based anchors are centroid points from a set of semantic locations (e.g., restaurant, hotel, shopping). Feature-based anchors are important points in trajectory data, named features, such as turning points. Each type of anchor point has strengths and drawbacks when used to build a reference system. However, the most relevant factors are the computational cost and how the reference system contributes to the quality of the calibration. Table 1 shows the computational cost of the basic operations using different anchor points. The results shown in [22] reveal that the feature-based approach contributes significantly to the quality of the calibration, when compared to the others.

The anchor points form the reference system to be used in the calibration method. Su et al. [22] presented two calibration methods: geometry-based and model-based. Table 2 compares the time complexity of both calibration methods. The geometry-based method runs faster than the model-based one. However, the model-based is the most robust method for reconstructing the input trajectory, since it considers the correlation between anchor points.

The following drawbacks motivate us to develop the current work. First, a detailed algorithm for building a reference system is not presented in [22]. Also, its geometry-based calibration method is faster than the model-based method, but ignores the relationship between anchor points in the reference system, leading to an inaccurate calibration. In addition, the taxi trace mentioned and used in their work is not described; in other words, it is not possible to reproduce their results. Finally, the calibrated data is not publicly available.

This work goes further and proposes algorithms to calibrate incomplete trajectory data, and makes calibrated traces available to the research community. Furthermore, our geometry-based calibration method performs better than [22], since it considers the relationships between the points in the reference system. Therefore, researchers can easily reproduce our results, apply our solution to other traces, and download the already calibrated traces from five different cities. Most importantly, we envision more realistic performance evaluation results of VANET and DTN solutions.

3 Mobility Traces

The vehicular mobility traces available in the literature can be classified as synthetic or real. The synthetic traces are built by mobility generator tools considering particular characteristics of the city, such as population, neighborhood (i.e., residential, commercial, industrial), and other aspects collected by the city managers. The most well-known synthetic mobility traces are from Cologne [36] and Zurich [37]. Since synthetic traces present a high granularity in terms of space and time, there is no need to fill their gaps. Moreover, this kind of trace will be very useful in our research, as it will work as the ground truth to validate our calibration method.

The real mobility traces are the ones generated by real vehicles equipped with GPS-enabled devices. Usually, the real mobility traces represent the mobility of taxis, since it is

TABLE 3
Main Characteristics of the Vehicular Mobility Traces Used in This Work

Trace	Type	Average Granularity (s)	Duration (h)	Vehicles	Availability
Cologne	Synthetic	1	24	>117,488	Fully
Rome	Real-world	7	720	320	Fully
San Francisco	Real-world	60	720	536	Fully
Shanghai	Real-world	60	720	4316	Partially
Beijing	Real-world	177	168	10,357	Fully
Shenzhen	Real-world	30	216	13,799	Partially

easier to perform this kind of experiment in vehicles of this category than in ordinary vehicles [38]. We have selected five real mobility traces to use in this work: Rome, San Francisco, Shanghai, Beijing, and Shenzhen (see Table 3). The selection was motivated by their use in the literature and their geographical locations, which represent three different parts of the world, namely Europe, North America, and Asia. Each trace was created from a different source and uses a different format. To facilitate their adoption and use, we formatted all entries as tuples ⟨*id*, *timestamp*, *lat*, *long*⟩, where *id* is the vehicle's unique identifier, *timestamp* is the date and time of the entry in the format *yyyy-mm-dd hh:mm:ss*, and *lat* and *long* are the latitude and longitude, respectively, in the WGS84 coordinate system format. In the following, we describe the main details of each trace.

3.1 Cologne

The trace of Cologne [36] is a synthetic dataset that represents car traffic in the city of Cologne, Germany. The trace contains the positions of cars throughout a 24-hour period of a workday in an area of 400 km^2. This dataset is the result of employing state-of-art tools and resources such as road topology from OpenStreetMap (OSM); microscopic vehicular mobility simulated with the software Simulation of Urban Mobility (SUMO); and macroscopic traffic flow derived from census data and surveys.

3.2 Rome

The trace of Rome [8], [23] contains the positions of taxi cabs working during the entire month of February, 2014. Each taxi had a device running Android operating system, and was equipped with a GPS receiver that periodically retrieved its position and sent it to a central server. Positions with a precision error higher than 20 m were ignored. For the entire month, this trace contains a total of 21,817,851 entries coming from 316 taxis. On average, each vehicle contributed 69,040 positions for the available period. However, a few vehicles contributed higher values up to 118,500, while others contribute values as low as 19, for example.

3.3 San Francisco

The trace of San Francisco [24], [39] contains the positions of 536 taxis working during the month of May, 2008. Each taxi had a GPS receiver installed in it, and sent location information (identifier, timestamp, latitude, longitude) periodically to a central server. This trace contains, for the entire month, a total of 11,219,955 entries. Each taxi contributed, on average, around 20,930 entries. A small minority of them contributed significantly lower values, while others contributed higher measures up to 49,370.

3.4 Shanghai

The trace of Shanghai [25], [40] contains the positions of 4,316 taxis working from February to April of 2007.[1] However, the public available dataset includes data for a single day of February, containing 6,075,587 position entries. Like the other cases, the taxis were equipped with GPS-enabled devices, which sent their position information periodically to a server. On average, each vehicle collected around 1,408 entries. A few vehicles contributed lower values, while others contributed higher values, up to 7,011.

3.5 Beijing

The Beijing dataset [26], [41], [42] contains the positions of 10,357 taxis during one week in February, 2008. In total, it includes over 15 million positions. The distance traveled, considering all trajectories, sums up to 9 million kilometers. On average, each taxi collected its position every 177 seconds. A few vehicles contributed only a small number of entries, while others collected over 10,000 positions.

3.6 Shenzhen

The Shenzhen data trace [27] is the largest real-world dataset that could be found in the literature. It contains the trajectories of 13,799 taxis recording their positions every 30 seconds on average. In total, there are over 150 million entries collected during 9 days of April, 2011. This dataset was obtained from the authors of [27], who made it available.

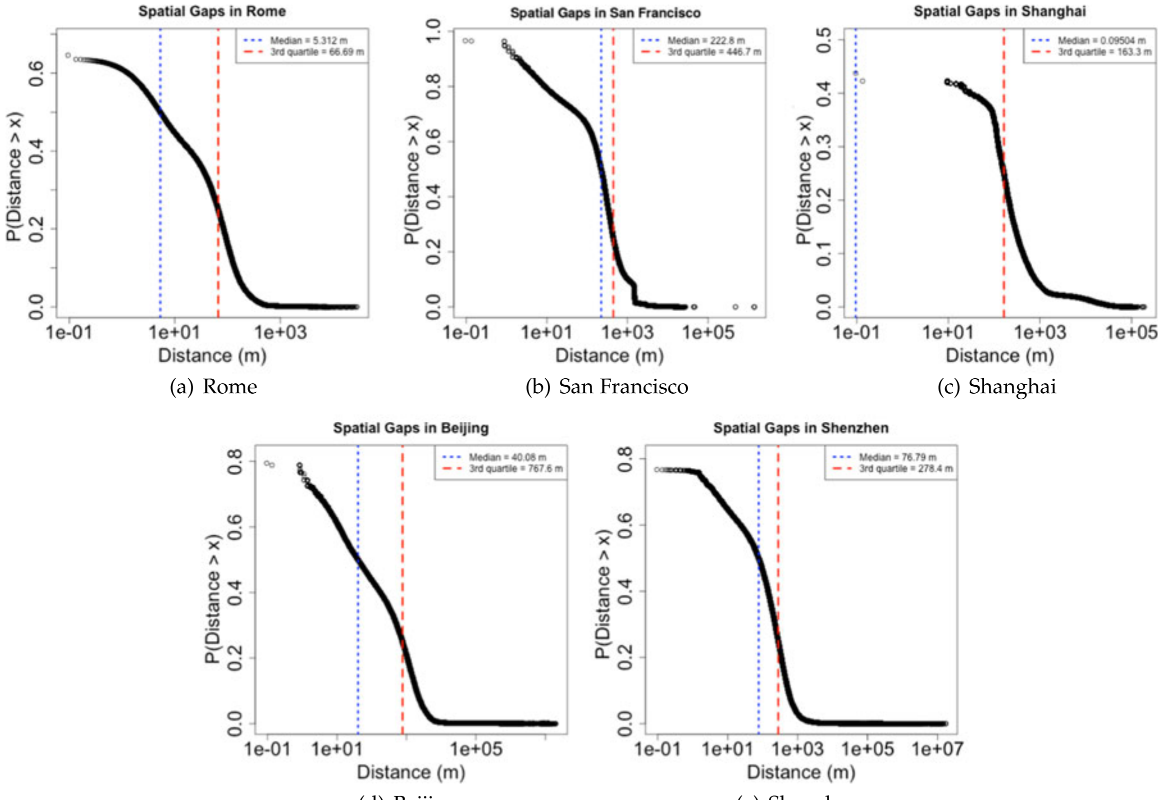
4 Identifying the Gaps

The completeness of the topology graph is a key factor for the performance evaluation of VANETs. In fact, contacts among vehicles that occurred in reality, but were not considered due to gaps in the trajectories of the traces, affect the evaluation of algorithms and protocols, since data exchange depends on these contacts. The formal definition of gap is introduced in Definition 2, where $\mathrm{d}(\cdot,\cdot)$ is the distance between two coordinates.

Definition 1 (Trajectory). *A trajectory is defined as a sequence of spatio-temporal points $T = \langle p_1, \ldots, p_n \rangle$, where $p_i = (x, y, t)$ for $i = 1 \ldots n$, and x, y are spatial coordinates, t is a timestamp, and $p_i.t < p_{i+1}.t$.*

Definition 2 (Gap). *Given the trajectory T of a vehicle and the threshold θ, a gap occurs when the spatial distance Δs between two consecutive points of T is greater than θ, i.e., $\Delta s = d(p_i, p_{i+1}) > \theta$.*

[1] These data were obtained from the Wireless and Sensor Networks Lab (WnSN), Shanghai Jiao Tong University.



(a) Rome (b) San Francisco (c) Shanghai

(d) Beijing (e) Shenzhen

Fig. 1. Complementary Cumulative Distribution Function (CCDF) of the distances between two adjacent points. These plots reveal that a significant number of entries present a distance between points that could affect the network topology.

To measure the expressiveness of the gaps in the existing original traces, we evaluate the distance between every two consecutive entries. Fig. 1 depicts the Complementary Cumulative Distribution Function (CCDF) of the distances between every two consecutive points for all original traces. As indicated by the third quartile (red vertical line), 25 percent of two consecutive points are 66.7, 446.7, 163.3, 767.6, and 278.4 m apart for Rome, San Francisco, Shanghai, Beijing, and Shenzhen, respectively. Considering those gaps and assuming a transmission range of 100 m [43], many existing contacts will be missed for the network topology graph built from the original traces. This clearly demonstrates the need for a method to calibrate the original traces with the objective of filling the existing gaps. In the next section, we describe and validate our approach for solving this problem.

5 Filling the Gaps

Our approach for filling the gaps in vehicular mobility traces is divided into two stages. The first stage extracts a reference system from the vehicles' historical GPS trajectory dataset. The second stage applies a calibration method, using a subset of anchor points of the previously built reference system. In the following, we describe both steps.

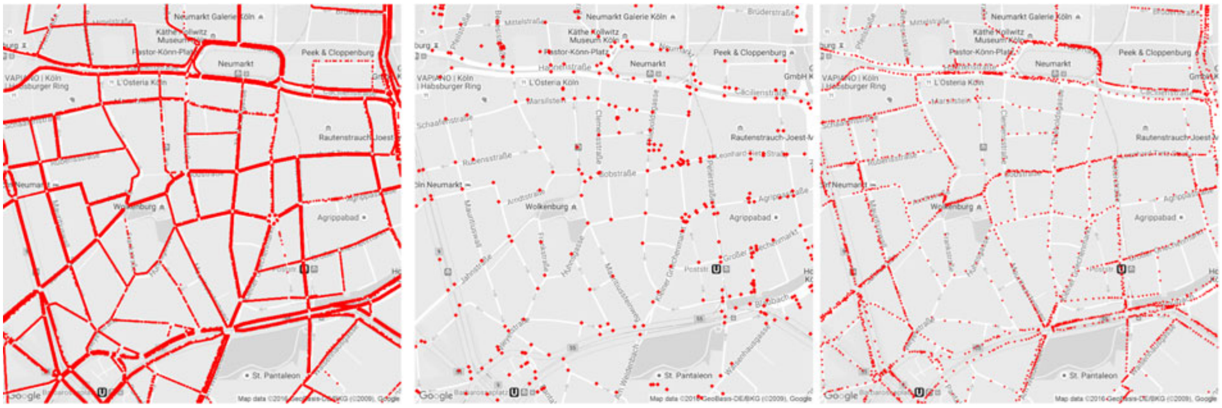
5.1 Cluster-Based Reference System

The reference system consists of a set of points resulting from a clustering process that uses historical trajectories. Each point, called centroid, represents a cluster of GPS points in close proximity to one another, recorded by all vehicles in the trace. Since GPS points represent real trajectories, it is reasonable to assume that each centroid is a potential location for a new point in a trajectory. In other words, it is very likely that a centroid represents a correct point in a road that vehicles travel through. Here, we adopt the k-means clustering method [44] for partitioning the data into k clusters, according to the density of GPS points; then, we obtain the centroid point of each cluster to form the reference system.

Algorithm 1 shows the basic steps to obtain the reference system. Initially, the k-means method partitions the data into k groups, according to the density of points (Line 2). Then, we obtain the centroid of each group and add it to the reference system (Line 3).

When using k-means, we need to choose an appropriate value of k. Thus, to overcome this problem, we apply the elbow method [45], which finds the minimum value of k that seems to give the smallest error. In other words, if we increase the value k, the error will not decrease significantly, meaning it is not worthwhile to do so. For the datasets used in this work, we find an average value of 20 percent of the total number of points. Regarding the computational complexity, the running time of the k-means clustering method is given as $O(nkdi)$, where n is the number of samples, d is the number of dimensions (two dimensions in our case, namely latitude and longitude), k is the number of clusters, and i is the number of iterations needed until the convergence of the clustering process is reached.

As mentioned in Section 2, Su et al. [22] proposed four methods to extract a reference system based on GPS data points. Here, we propose a novel cluster-based approach that outperforms those methods, considering the cost-benefit in terms of computational cost (see Table 1) and how the reference system contributes to the quality of the calibration. Fig. 2 depicts the reference systems obtained using the best cost-effectiveness method from [22] (namely, feature-



(a) Original GPS points

(b) Anchor points based on a feature

(c) Anchor points based on a clustering approach

Fig. 2. Part of the reference system created from the Cologne dataset. (a) Original GPS data from the downtown area of the Cologne dataset. (b) Example of a reference system created using a method based on a feature (turning points) [22]. (c) Example of a reference system created using our clustering approach.

based) and our cluster-based approach. Our approach leads to a finer calibration, since it does not consider only turning points, as can be seen in Figs. 2b and 2c. In addition, in road topologies where the presence of turning points are uncommon (such as highways), a method that considers only turning points will not work properly. However, the maps used in Fig. 2 are for the purposes of visualization only, and are not used by the algorithms.

Algorithm 1. Reference System Based on Clustering

```
Input: The historical of vehicles trajectories (raw_data) and number of clusters (k)
Output: Reference System (RefSys), a set of centroid points.
procedure ClusteringGpsPoints
    Clusters ← applyClustering(raw_data, k)
    RefSys ← getCentroids(Clusters)
end procedure
```

5.2 Calibration Method

In this stage, we perform the calibration following a geometric-based approach, which is an improvement to the base method described in [22]. More specifically, when there is a gap in a trajectory T, we obtain the reference system of the region, and then select the centroid points between the end-points of the gap.

The calibration method receives the following parameters as input: T, a set of n consecutive points with spatio-temporal information; *RefSys*, the reference system obtained from Algorithm 1; *min_d*, the threshold to consider the existence of a spatial gap; and *time_d*, the threshold to consider a temporal interval between two consecutive coordinates. As a result, we have a new trajectory T' with the original points from T and a set of calibrated points added to fill the existing gaps in T.

Algorithm 2 describes the calibration method. For each sequence of two points in T, we first check if there is a gap between them according to input parameters (Lines 4-8). If this is the case, we perform the calibration. Initially, we detect the set of centroid points from the reference system near the corresponding gap. For this, the *bounding_box* function finds the point halfway (midpoint) between the two end-points of the gap, and returns to the circle with its center in this midpoint (Line 9). Then, we obtain all centroid points from the reference system with coordinates inside the circle, and store them in C (Line 10). Next, we iteratively find the nearest point $a^* \in C$ to the centroid that satisfies the angular condition (Lines 14-15). The angular condition (Line 15) guarantees that only centroids in the same direction of the trajectory are considered, in order to avoid the selection of points in the opposite direction. If this is the case, we insert a^* in L between p_p and p_n. Next, we remove a^* from C and repeat this last sequence of steps while C is not empty (Lines 13-23). Finally, we insert the calibrated points of L into T'.

The algorithm described in [22] does not consider the relationship between the inserted points. In our solution, presented in Algorithm 2, we consider the relationship for choosing each new centroid based on the distance from the last selected centroid (Line 14).

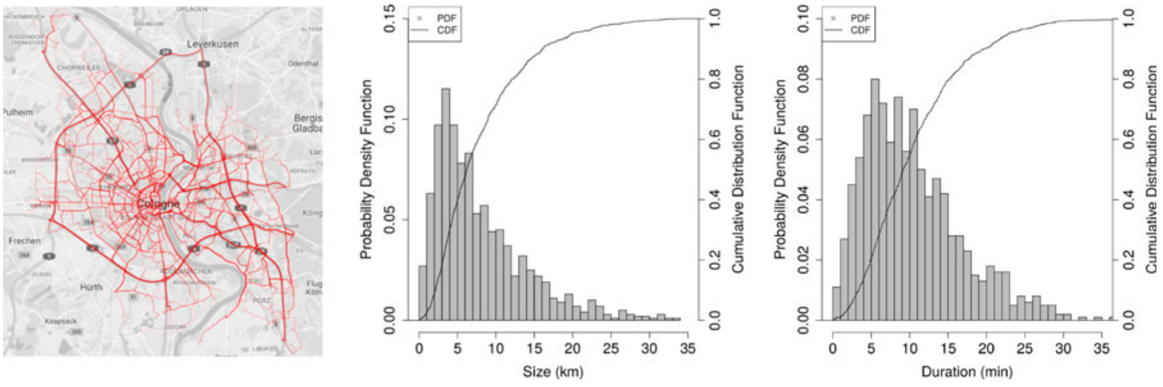
In addition to inserting the calibrated points given the spatial gap, it is important to obtain their timestamp to accurately represent the trajectory. Thus, before adding a^* to L (Line 16), we compute an estimated time for the temporal occurrence of the centroid a^* using Equation (1) [22], where $\mathrm{d}(\cdot,\cdot)$ is the distance between two coordinates

$$a^*.t = p_p.t + \frac{(p_n.t - p_p.t)\cdot d(p_p, a^*)\cdot \left|\overrightarrow{p_p a^*}\cdot \overrightarrow{p_p p_n}\right|}{d(p_p, p_n)\cdot \left|\overrightarrow{p_p a^*}\right|\cdot \left|\overrightarrow{p_p p_n}\right|}. \quad (1)$$

Regarding the computational complexity, the running time of Algorithm 2 depends on the length of T and the number of centroid points in C for each calibrated gap. As N_c is the average number of centroids for a gap and N_T is the length of the trajectory, it follows that the complexity is $O(N_T N_c^2)$. Given that the number of centroids is not high because of the adopted elbow method, and that this is an offline process that aims to calibrate the traces only once, this complexity seems to be very reasonable.

6 Validation

In this section, we perform trajectory similarity analysis to validate the impact of our method on low-sampling-rate trajectories. The goals of this validation are twofold. The first is to qualitatively evaluate the trajectories after calibration, highlighting visual differences in the shape. The second



(a) Spatial coverage (b) Size of the trajectories (c) Duration of the trajectories

Fig. 3. Characterization of the sampled dataset from Cologne. These plots reveal the spatio-temporal heterogeneity in the subset of trajectories obtained from the Cologne dataset.

goal is to compare the calibrated trajectories with those of the original using similarity measures.

Algorithm 2. Calibration Method

Input: Trajectory ($T = [P_1, P_2, \ldots, P_n]$), Reference System (*RefSys*), minimum spatial distance (*min_d*), and temporal distance (*time_d*)
Output: A new trajectory (T') without gaps.

```
Procedure CALIBRATE
    $T' \leftarrow T[1]$
    for $i \leftarrow 2$ to length($T$) do
        $p_p \leftarrow T[i-1]$                    $\triangleright$ $p_p$ is the previous point
        $p_n \leftarrow T[i]$                        $\triangleright$ $p_n$ is the next point
        $d \leftarrow$ distance($p_p, p_n$)
        $t \leftarrow$ interval($p_p, p_n$)
        if $d > min\_d$ and $t < time\_d$ then
            $bb\_coord \leftarrow$ bounding_box($p_p, p_n$)
            $C \leftarrow$ subset($RefSys$, $bb\_coord$)
            Initialize an empty list $L$
            $a' \leftarrow p_p$
            while $C$ is not empty do
                $a^* \leftarrow \arg\min_{a \in C} d(a, a')$
                if $\angle(\overrightarrow{a'a^*}, \overrightarrow{p_p p_n}) < \frac{\pi}{2}$ then
                    Add $a^*$ to $L$
                    $a' \leftarrow a^*$
                end if
                Remove $a^*$ from $C$
            end while
            Insert the centroids in $L$ into $T'$
        else
            Insert $p_n$ in $T'$
        end if
    end for
    return $T'$
end procedure
```

In this validation, we first randomly select (with a uniform distribution without replacement) 1,000 trajectories from different vehicles from the Cologne dataset. Fig. 3 shows a characterization in terms of spatial and temporal features of these selected trajectories. Fig. 3a shows the spatial coverage of the selected trajectories. We can observe that many urban roads are in red, indicating the presence of trajectories over different parts of the city (i.e., downtown, highways and peripheral areas). The intensity of red represents a high incidence of points in the same roads; this behavior is more common in the central area and in roads crossing the city. The size of the selected trajectories, depicted in Fig. 3b, varies from a few meters to about 35 kilometers. Approximately 70 percent of the selected trajectories have a size smaller than 10 km, as expected in urban scenarios and observed in the original Cologne dataset [36]. Some trajectories have a size greater than 10 km representing commuters crossing the city. Intrinsically, the size of the trajectories impacts the duration of the displacement, as can be seen in Fig. 3c, where we can observe that approximately 60 percent of the trajectories last less than 10 minutes.

6.1 Qualitative Validation

For each of the selected trajectories, we apply a sampling process that retrieves records every 10, 20, 30, 60, and 100 seconds, generating gaps in the fine-grained data.[2] Thus, T is an original trajectory with sampling rate every 1 s and T_x, where $x \in \{10, 20, 30, 60, 100\}$, is a trajectory with sampling rate x obtained from T. For instance, Fig. 4 depicts the original trajectory of a vehicle and Figs. 5a, 5c, 5e, 5g, and 5i are sampled trajectories of the same vehicle. We may see that this sampling intentionally causes gaps in the trajectory.

To validate our calibration method, we apply it to the sampling trajectories (i.e., with hand-generated gaps) to fill their gaps. Considering that the chosen trajectory presents interesting peculiarities such as straight segments, curvatures and a long distance path, we may see in qualitative terms that the calibrated trajectories are very similar to the original ones. Even when the gaps are large (e.g., for T_{60} and T_{100}), the calibration method accurately reconstructs the trajectories, leading to fine-grained traces. By using historical trajectories and applying a clustering approach, we detected potential candidate points to be inserted into the trajectories. In the reference system approaches proposed in the literature and described in Section 2, the anchor points are sparsely or irregularly distributed, except for the data-based strategy. However, the data-based approach has a high degree of redundancy in the data when there is a large number of records. In addition, our calibration method considers the relationship between anchor points, as can be

[2] These values are defined based on granularity of vehicular mobility traces described in Table 3.

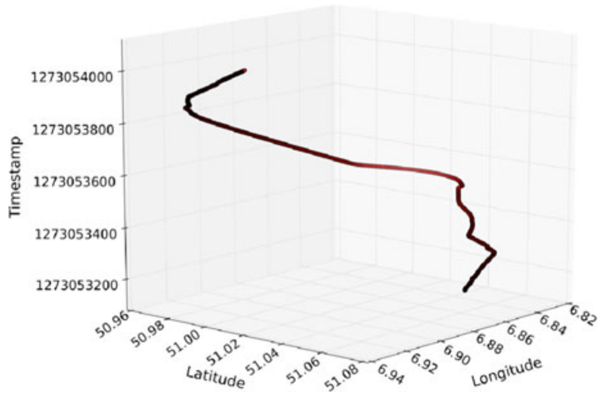


Fig. 4. Example of an original trajectory of vehicle #134 with sampling rate of 1 s.

seen in Figs. 5b, 5d, 5f, 5h, and 5j, where the calibrated trajectories have a very similar shape to the original ones (Fig. 4), whereas they respect the road topology.

6.2 Quantitative Validation

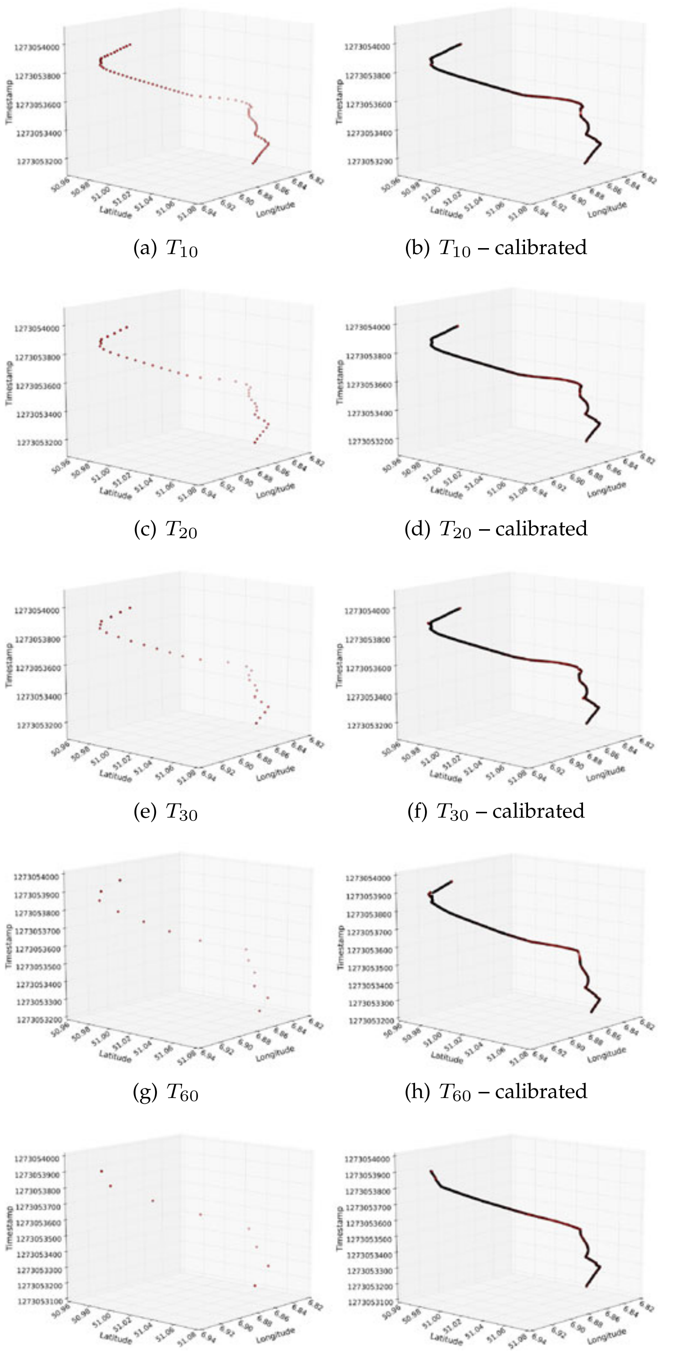
Until now, we have discussed the quality of the trajectory obtained by our calibration method. Going further, we also consider a quantitative measure that defines how a calibrated trajectory is similar to the original one, and, thus, provides a higher reliability of results. In this sense, we compare the original and calibrated traces by adopting two existing trajectory similarity measures, as described in the following:

- Dynamic Time Warping (DTW) [46]: DTW is a similarity measure that explores the matching points between trajectories. This measure presents a good performance with different sizes of trajectories and different sampling rates. However, it is highly affected by outliers, since each point in the original trajectory should have at least one associate point in the calibrated trajectory.
- Edit Distance (EditDist) [47]: This measure is relatively unaffected by the presence of outliers, because there is a parametric threshold (ϵ) that associates each point in the original trajectory to a point in the calibrated trajectory. However, if the trajectories have different sizes, the EditDist is increased.

In this way, those measures allow us to compare the original and calibrated trajectories considering outliers, differences in the size of the trajectories, and different granularities. It is worth mentioning that, when the calibrated trajectory is identical to the original trajectory, the value obtained for each measure is exactly zero.

In this validation experiment, we compute the similarity measures between the original and calibrated trajectories for each of the 1,000 trajectories initially chosen from the Cologne trace. Thus, we generate the gaps in the original trajectories with different sampling rates (i.e., 10, 20, 30, 60, and 100), then we calibrate each trajectory using our proposed method and the solution from [22], and finally we compare the calibrated and original versions. We compute the distance of the trajectories normalized by the trajectory length. It is important to know that the reference system was constructed from the original Cologne dataset.

Fig. 6 presents the results for the metrics DTW (Fig. 6a) and EditDist (Fig. 6b). For both measures, the distances



(a) T_{10} (b) T_{10} – calibrated

(c) T_{20} (d) T_{20} – calibrated

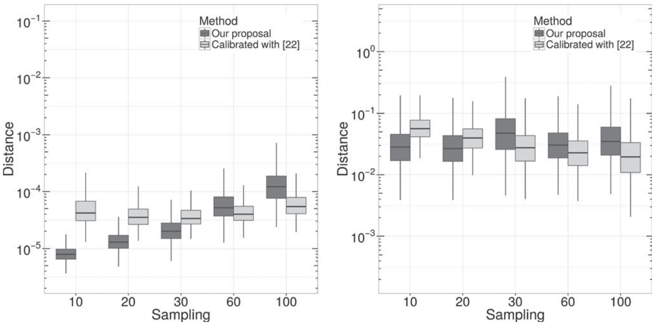
(e) T_{30} (f) T_{30} – calibrated

(g) T_{60} (h) T_{60} – calibrated

(i) T_{100} (j) T_{100} – calibrated

Fig. 5. Calibration method applied to gaps with different sizes. These plots reveal that our calibration approach could accurately fill gaps in mobility traces considering qualitative aspects.

between the original and calibrated traces are close to zero, meaning that the calibration method could accuratelly fill the gaps. The results show that the calibrated trajectories are very similar to the original ones. We stated that the calibration increases the granularity of the trajectories without entering outliers during the process. This can be seen in the case shown in Figs. 5b, 5d, 5f, and 5h in relation to Fig. 4. The high quality of the results is obtained because the calibration process uses historical data and applies the clustering method to summarize them in an anchor point. In addition, higher sampling rates generally result in greater distances, since large gaps are more



(a) DTW (b) EditDist

Fig. 6. Comparison of the original and calibrated traces in terms of two trajectory distance measures: DTW and EditDist.

difficult to fill, as expected. However, even for T_{100}, the results are very promising.

When we compare our calibration method with the work in [22], we can see that they exhibit a similar behavior regarding EditDist (Fig. 6b). The reason is that both methods do not influence the size of the calibrated trajectory and the possible outliers in the calibrated trajectories. For the DTW measure, our method generates more similar-to-the-original trajectories, mainly when the sampling rate is less than 60 seconds; this happens because there are fewer outliers in our calibrated trajectories.

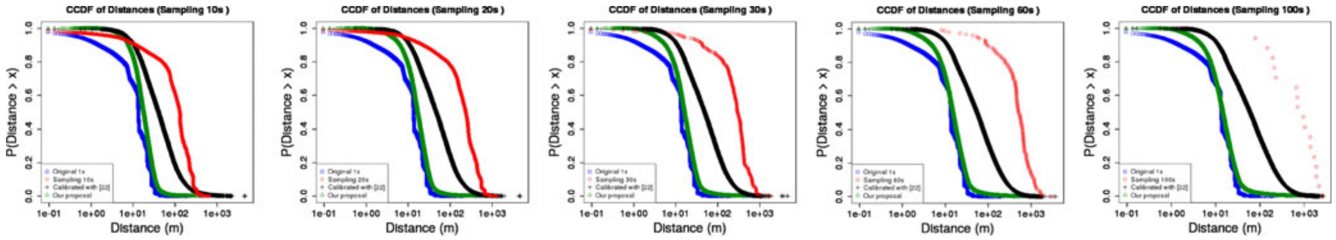
The aforementioned measures reflect only spatial aspects of the traces. To evaluate them considering a spatio-temporal perspective, we assess the distance and the time between the consecutive points of a trajectory. In Fig. 7, we have as ground truth the CCDF of the set of trajectories, with points every 1 s. As expected, both methods reduce the distance between consecutive points (gaps), as can be observed when comparing the CCDFs of the calibrated trajectories with the ground truth. Similarly, in Fig. 8, the time between consecutive points is analyzed by considering the ECDF of the trajectories calibrated with the two methods. In this case, our approach significantly reduces the sampling to approximately 1. As we can see, 90 percent of the time interval between consecutive points has less than 3 seconds. Our approach is significantly better for all analyzed cases because it considers a reference system that has well-distributed anchor points in the road segments, whereas the baseline only uses turning points.

In summary, the validation results reveal that our calibration approach could accurately fill gaps in mobility traces. In the next section, we apply our calibration method to five real mobility traces, and compare the calibrated versions with the original ones in terms of network connectivity.

7 Network Connectivity Evaluation

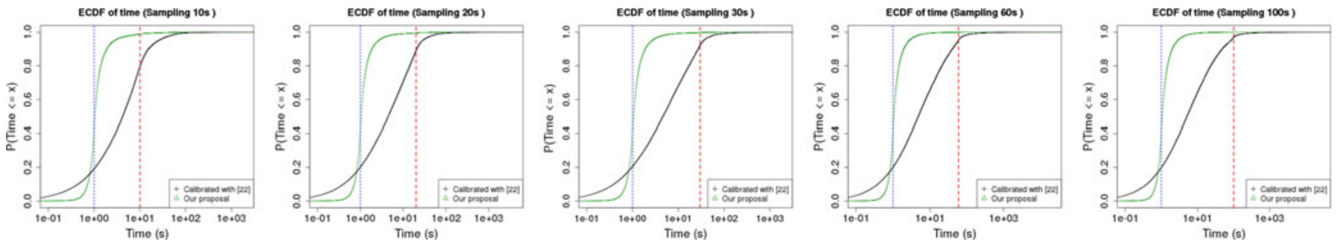
Having introduced our calibration method and validated it using the Cologne dataset, we must evaluate how possible interactions that appear in both real and calibrated traces lead to connectivity and topology in vehicular networks. An important issue here is how they differ from each other and lead to different results. This is a fundamental aspect if we want to understand the behavior of protocols and algorithms for VANETs. For this, we randomly select a day in each of the real vehicular mobility traces described in Table 3, and then apply the calibration method presented in Section 5 for all trajectories of this day in each trace. The outcome of this process therefore consists of two subsets (original and calibrated traces) for each vehicular mobility trace.

To investigate the impact of the calibration, we need to compare the communication graph of the original and the calibrated traces. The goal is to show how the gaps presented in the original traces lead to unrealistic communication graphs, which are improved with our calibration method. Results were obtained assuming a transmission range of 100 m [43]. Thus, any pair of vehicles that are, at most, 100 m apart are able to establish a communication link and, therefore, communicate. The communication topology graphs for each of the five traces, either original or calibrated, were built considering an entire period of 24 hours. Despite being a simple communication model, this strategy allows us to assess the impact of filling the gaps in the traces, which is the objective of this work, and



(a) Sampling 10s (b) Sampling 20s (c) Sampling 30s (d) Sampling 60s (e) Sampling 100s

Fig. 7. Complementary Cumulative Distribution Function (CCDF) of the distances between two adjacent points.



(a) Sampling 10s (b) Sampling 20s (c) Sampling 30s (d) Sampling 60s (e) Sampling 100s

Fig. 8. Empirical Cumulative Distribution Function (ECDF) of the time between two adjacent points. The blue dashed line represents the reference point of 1s of the original trajetories, and the red longdashed line represents the reference point of the sampling rate.

TABLE 4
Global Connectivity

Metric	Trace	Original	Calibrated
Number of connected components	Rome	7	3
	San Francisco	2	1
	Shanghai	297	141
	Beijing	3,624	780
	Shenzhen	27	12
Size of the largest component	Rome	281	285
	San Francisco	496	497
	Shanghai	3,994	4,161
	Beijing	6,203	9,293
	Shenzhen	10,844	10,859

avoids factors that may influence the assessment process, such as signal propagation and collisions. These factors are not within the scope of this work, but are part of future work, as discussed in Section 9.

7.1 Global Connectivity

An important aspect when it comes to the communication graph is to determine whether or not the graph is connected. We investigate the number of connected components and their size. These two metrics are able to summarize the connectivity of a communication graph, so that the first metric refers to the level of the network fragmentation, and the second one describes how the largest component is dominant over the whole network.

The *global connectivity* [48] measures the largest connected component of the communication graph. Therefore, the higher this value, the more connected a graph is. Table 4 presents the number of connected components and the size of the largest component. It is clear that the communication graph becomes more connected after the traces' calibration. The number of connected components decreases over 50 percent for all traces, with highlights for Beijing that decrease by 78 percent. This indicates that the gaps in the original traces cause fragmentation in the communication graph. Using the method proposed here, we obtained a less fragmented network primarily for the case of a trace with low granularity, as it is the case of Beijing. In addition, the calibration method contributes to increase the size of the largest component. This is evident from the calibration because it creates opportunities for new connections, particularly for trajectories with low sampling rates.

These results reveal that the original graphs miss important contacts that help increase the network connectivity. Moreover, these traces have been widely adopted in different studies of vehicular networks, and, thus, the calibrated traces will definitely increase the reliability of such investigations.

7.2 Transient Connectivity

The *reach* of a vehicle is the total number of vehicles to which it is transiently connected [48]. By transiently connected, we mean that a vehicle may not have a direct link with another vehicle, but can reach it through other vehicles in future contacts. This is an important metric in DTNs, since data may be delivered opportunistically to the final destination by future connections. Fig. 9 presents the Complementary Cumulative Distribution Function of the 2-hop reachability for all vehicles, that is, the proportion of other vehicles one can reach within two hops.

For all traces, the calibration method leads vehicles to reach more vehicles within two hops of distance. Again, this is due to the missing contacts existing in the original traces. Regarding the San Francisco trace, it should be noted that all vehicles in the calibrated trace reached all others within two hops, as indicated by the unique blue dot in Fig. 9b, since the probability of all vehicles to reach all others within two hops is 1.

These results have the potential for significant consequences in the evaluation of routing protocols that consider the delivery rate and overhead, as discussed in Section 8. For instance, a striking difference was noted between the reachability of the original traces and the reachability of the calibrated traces, as can be seen in Figs. 9c and 9d. In both cases, the percentage of vehicles reached in two hops increases considerably when using the calibrated trace, thereby increasing the coverage of vehicles in the network.

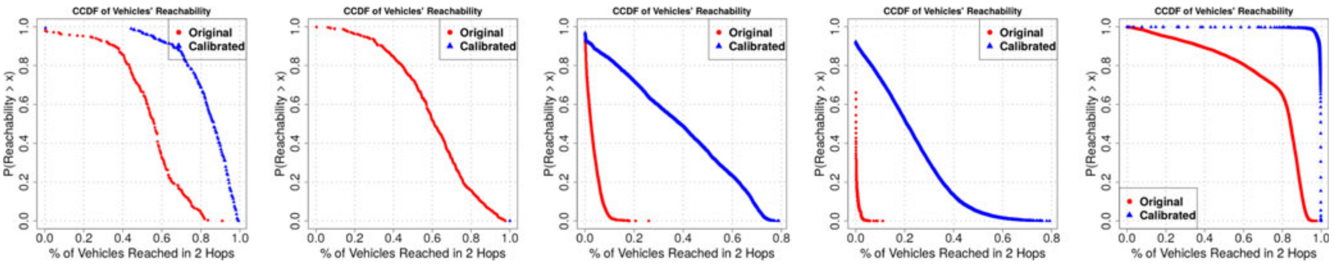
7.3 Network Density

The network density, represented by the vehicle's degree, is also an important communication metric that affects how a message is disseminated throughout the network [49]. Fig. 10 depicts the Complementary Cumulative Distribution Function of the number of contacts of all vehicles for the original and calibrated traces. It can be noted that the vehicle's degree increases with the calibration method, due to the new contacts created after filling the existing gaps.

7.4 Link Lifetime

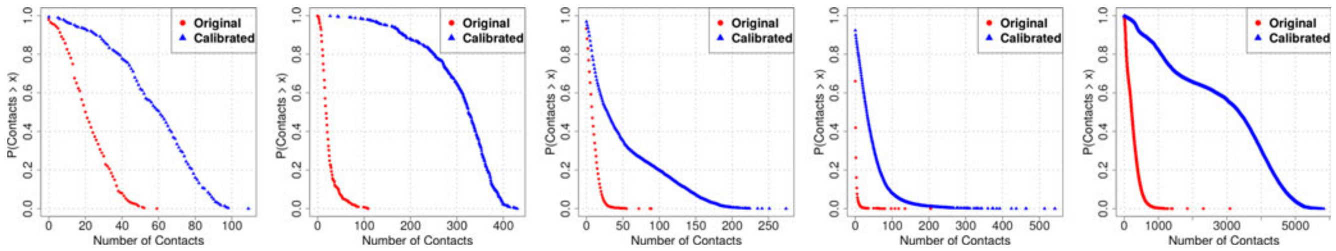
The link stability is measured in terms of the lifetime of pairwise links [49]. This metric plays an important role when building communication paths for routing protocols. Here, the link lifetime is consider as the total time a vehicle is in communication range with another one, until the time they move away from each other and are no longer in contact.

Fig. 11 depicts the Complementary Cumulative Distribution Function of all pairwise link duration. It can be seen that links in the original traces last mostly for just 1 second, while in the calibrated traces, many links last for



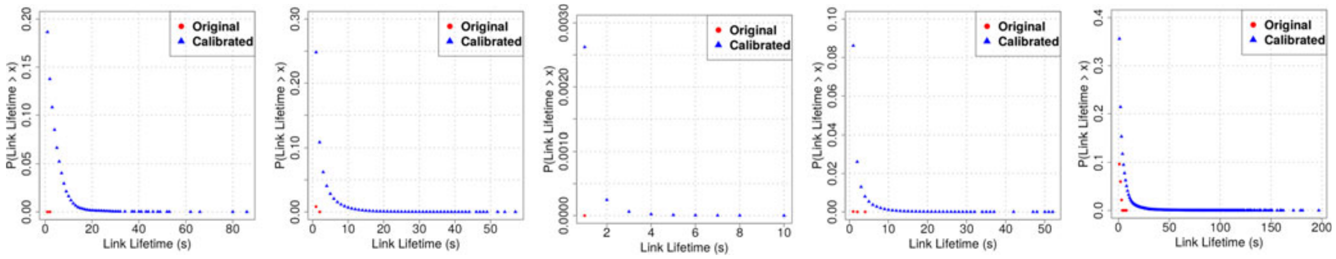
(a) Rome (b) San Francisco (c) Shanghai (d) Beijing (e) Shenzhen

Fig. 9. CCDF of the 2-hop reachability of all vehicles. The calibration method increases the number of vehicles reached in two hops.



(a) Rome (b) San Francisco (c) Shanghai (d) Beijing (e) Shenzhen

Fig. 10. CCDF of the number of contacts for each vehicle. It is possible to see how the contacts increase after the calibration.



(a) Rome (b) San Francisco (c) Shanghai (d) Beijing (e) Shenzhen

Fig. 11. CCDF of the link lifetime for all contacts. The calibration makes links to last for longer periods than in the original traces.

significantly longer periods. This result is due to the calibration method that increases the traces' granularity by inserting new points, thus enabling the contacts between vehicles to have a longer duration. Therefore, in addition to increasing the number of contacts, the calibration method also improves the traces in terms of the stability of contacts.

7.5 Path Length

The path length is the number of hops between two vehicles [49]. The average path length is calculated by averaging the shortest paths between all pairs of vehicles. Table 5 presents the average path length for the communication graphs built from the original and calibrated traces. It can be noted that the average path length is lower for the calibrated traces, due to the fact that more contacts lead to more possible paths, allowing shortest paths between a pair of vehicles.

8 Impact of Calibrated Traces on Vehicular Networking

Having evaluated and discussed how the interactions appear in both original and calibrated traces with respect to connectivity and topology in vehicular networks, we are now interested in understanding the effects of the calibration in realistic vehicular network scenarios. As mentioned above, the focus of the previous analysis was to understand how the different topologies obtained from the traces differ in terms of network connectivity. To this end, we employed a connectivity graph model and disregarded details of the protocol stack. In this section, we suggest a networking application and analyze the results for both the original and the calibrated traces using a vehicular protocol stack that considers issues such as medium access, collision and channel error.

TABLE 5
Average Path Length

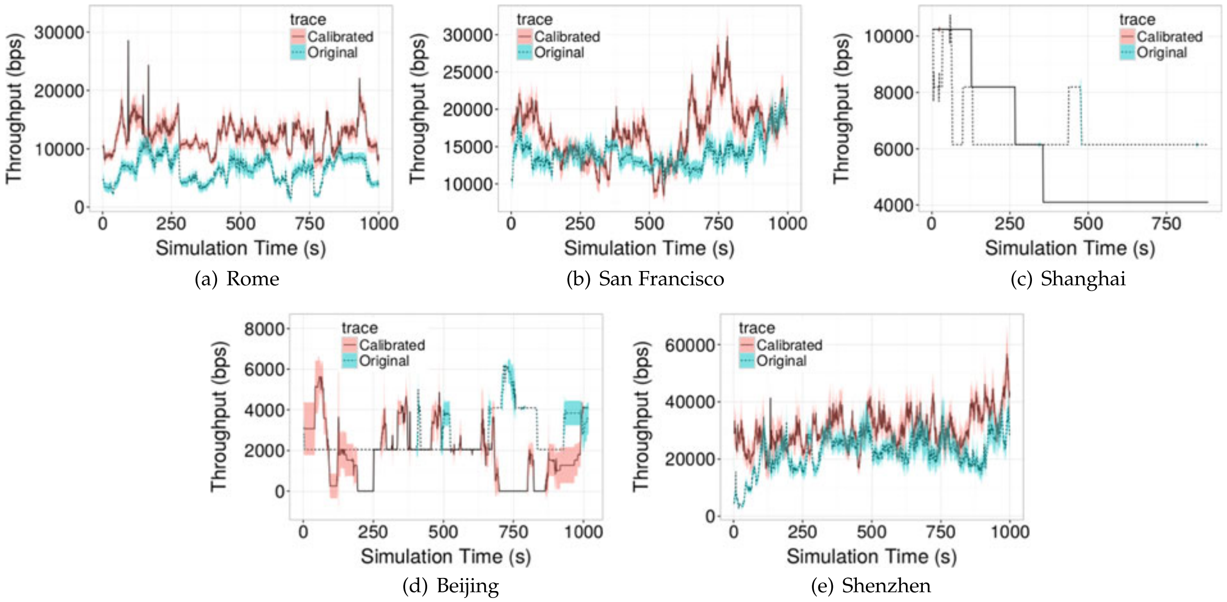
Metric	Trace	Original	Calibrated
Average Path Length	Rome	2.42	1.95
	San Francisco	2.35	1.38
	Shanghai	4.20	2.89
	Beijing	5.52	2.81
	Shenzhen	2.23	1.74

More specifically, we address the problem of multi-hop dissemination in an instantaneous network topology, where packets are routed through the network using multiple hops between the origin and destination vehicles, considering the dynamics of the existing connections over time. For each scenario, all vehicles in the network transmit 64-byte packets at a communication rate of 2,048 kbps to half the vehicles selected as sinks, reflecting an application with probe vehicles (e.g., taxis and patrol cars) acting as mobile sensors for sensing the urban scenario, and sending data to mobile sink vehicles [50].

To simulate the vehicular mobility and the protocol stack, we use the Network Simulator 3 (NS-3),[3] a well-known discrete-event network simulator. Its current version has important modules for the VANET simulation: the Nakagami propagation model, mobility module, and network with support to IEEE 802.11p [51] and IEEE 1609/WAVE [52] standards. Additionally, we use a well-known routing protocol, called AODV [53], for message forwarding. It is important to note that NS-3 "moves" objects by using a linear interpolation, i.e., its "calibration method" uses linear segments between consecutive positions of objects.

The metric evaluated in our simulations was the throughput, which represents the number of packets received by the destination vehicles at every second. This metric allows us to understand how the instantaneous topologies obtained from both calibrated and original traces differ in a vehicular communication scenario. Fig. 12 depicts the variation of the average throughput over time for calibrated and original traces. All results represent the average, considering 95 percent confidence interval from

[3] https://www.nsnam.org/



(a) Rome (b) San Francisco (c) Shanghai

(d) Beijing (e) Shenzhen

Fig. 12. Comparison of the average throughput along the simulation time between the original and calibrated traces.

15 simulation runs. Our evaluation considers a simulation time of 1,000 s starting at 10:00 am in each city.

Fig. 12a shows the average throughput for the traces of Rome. It shows a similar pattern, but with a higher throughput for the calibrated data. In this case, as the average granularity of the original trace is 7 s, we get a slightly different instantaneous communication topology from the two traces. However, the throughput is greater in the calibrated trace because the contacts are longer, resulting in greater network capacity. This confirms the results of the link lifetime in Fig. 11a, where the calibration makes links last for longer periods than in the original traces.

For the San Francisco traces, as depicted in Fig. 12b, we have a significant distinction in the behavior obtained for the two traces. The reason for this difference can be explained by the fact that the original trace is represented by the interpolation between long gaps, since the average granularity of the original data is 60 s. The results around 250 s are similar because the intensity of mobility of the vehicles is reduced in that period. At this moment, we see another importance of the calibration, because when we perform a linear interpolating between distant points, the path made by the vehicle during the simulation may be quite different from the actual one.

Fig. 12c depicts the results for the calibrated and original traces of Shanghai. We can see that the average throughput changes discreetly over time for the two traces. This occurs because vehicles have more intense movement at the beginning of the simulation. However, when we use a mobility visualization tool,[4] many vehicles remain static during the simulation, compromising the routing of packets. These results confirm the plot observed in Fig. 11c, where the link lifetime is short and the probability of having a link lifetime for a long period is extremely low.

As expected, the size of the gaps affects the instantaneous topology of vehicular networks. As a matter of fact, the average throughput of the Beijing dataset, shown in Fig. 12d, indicates that the gaps in the original trace, with average granularity of 177 s, directly influence the topology, and, consequently, the performance of the protocol. Obviously, the interpolation method used to construct the mobility in the simulator causes non-realistic results, when we have larger gaps.

Fig. 12e depicts the results of the average throughput for the calibrated and original traces of Shenzhen. We can see that the throughput using the calibrated trace is higher during the simulation. This reflects the influence of the global connectivity and link lifetime discussed in Section 7. Although the average granularity of the original trace is 60 s, at certain times, the throughput presents similar results. This occurs because most vehicles travel through a set of major highways with a straight shape. For these cases, interpolation does not compromise as much as in scenarios with curvilinear trajectories. Furthermore, for all traces presented in this section, we can see that Shenzhen exhibits the highest average throughput. Clearly, this is related to the link lifetime, as shown in Fig. 11e.

The results discussed in this section show that large gaps in real vehicular mobility traces lead to unrealistic topologies, when not calibrated or calibrated using a linear interpolation, affecting the performance evaluation of routing protocols. Indeed, the interpolation method used by simulators, when applied to traces with large gaps, introduces significant bias, particularly when there is no road map associated with the vehicles' movements. The method introduced in this work improves the quality of the traces, leading to more realistic scenarios, and, consequently, increasing the reliability of the evaluation results.

9 Conclusion

This work shows that existing real vehicular mobility traces present gaps that lead to network topologies differing from reality, and, consequently, to an unreliable performance evaluation. To tackle this problem, we have proposed and validated a solution to find and fill gaps by adopting a cluster-

[4] https://www.nsnam.org/wiki/NetAnim

based reference system and a calibration method. The results revealed that our approach is able to accurately fill the gaps. Moreover, we have observed that the network topologies built from the calibrated traces differ significantly from the original ones. To address this, we have presented the evaluation results that compare the communication graph of the original and the calibrated traces for five real-world traces. Our results provide a clear distinction between the communication graphs from the original and calibrated traces.

The literature indicates that the Cologne trace constitutes the most complete vehicular mobility trace. Despite having a high granularity, the Cologne trace is a synthetic trace and has a duration of 24 hours. On the other hand, the application of the calibration method to real vehicular mobility traces improves their quality, leading to more trustworthy simulation results. To contribute to the research community, we made the calibrated traces publicly available for the five different cities.

As future work, there are some interesting issues to investigate. We plan to fine-tune the calibration solution to avoid adding calibrated points outside roads caused by GPS errors in the traces. It is important to evaluate other clustering algorithms, as well as other strategies for building the reference system. We aim to evaluate other state-of-the-art protocols for vehicular networks considering aspects of communication, and evaluate the impact of our proposal in the simulations of these protocols.

Acknowledgments

The authors would like to thank the Minas Gerais Research Support Foundation (FAPEMIG), the National Council for Scientific and Technological Development (CNPq), and the Coordination for the Improvement of Higher Education Personnel (CAPES) for grants and financial support.

References

[1] A. Grzybek, M. Seredynski, G. Danoy, and P. Bouvry, "Aspects and trends in realistic VANET simulations," in *Proc. IEEE Int. Symp. World Wireless Mobile Multimedia Netw.*, Jun. 2012, pp. 1–6.

[2] S. Joerer, F. Dressler, and C. Sommer, "Comparing apples and oranges?: Trends in IVC simulations," in *Proc. 9th ACM Int. Workshop Veh. Inter-Netw. Syst. Appl.*, 2012, pp. 27–32.

[3] A. Kesting, M. Treiber, and D. Helbing, "Connectivity statistics of store-and-forward intervehicle communication," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 172–181, Mar. 2010.

[4] J. Harri, F. Filali, and C. Bonnet, "Mobility models for vehicular ad hoc networks: A survey and taxonomy," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 4, pp. 19–41, Oct.–Dec. 2009.

[5] M. Fiore and J. Harri, "The networking shape of vehicular mobility," in *Proc. 9th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2008, pp. 261–272.

[6] M. Fiore, J. Harri, F. Filali, and C. Bonnet, "Understanding vehicular mobility in network simulation," in *Proc. IEEE Int. Conf. Mobile Ad Hoc Sens. Syst.*, 2007, pp. 1–6.

[7] R. Baumann, S. Heimlicher, and M. May, "Towards realistic mobility models for vehicular ad-hoc networks," in *Proc. IEEE Mobile Netw. Veh. Environ.*, 2007, pp. 73–78.

[8] R. Amici, M. Bonola, L. Bracciale, A. Rabuffi, P. Loreti, and G. Bianchi, "Performance assessment of an epidemic protocol in VANET using real traces," *Procedia Comput. Sci.*, vol. 40, pp. 92–99, 2014.

[9] Y. Chen, M. Xu, Y. Gu, P. Li, and X. Cheng, "Understanding topology evolving of VANETs from taxi traces," *Adv. Sci. Technol. Lett.*, vol. 42, pp. 13–17, 2013.

[10] A. Cornejo, C. Newport, S. Gollakota, J. Rao, and T. Giuli, "Prioritized gossip in vehicular networks," *Ad Hoc Netw.*, vol. 11, no. 1, pp. 397–409, 2013.

[11] M. A. Hoque, X. Hong, and B. Dixon, "Efficient multi-hop connectivity analysis in Urban vehicular networks," *Veh. Commun.*, vol. 1, no. 2, pp. 78–90, 2014.

[12] H. Huang, D. Zhang, Y. Zhu, M. Li, and M.-Y. Wu, "A metropolitan taxi mobility model from real GPS traces," *J. Universal Comput. Sci.*, vol. 18, no. 9, pp. 1072–1092, May 2012.

[13] H. Huang, Y. Zhu, X. Li, M. Li, and M. Y. Wu, "META: A mobility model of metropolitan taxis extracted from GPS traces," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2010, pp. 1–6.

[14] C. H. Lee, J. Kwak, and D. Y. Eun, "Characterizing link connectivity for opportunistic mobile networking: Does mobility suffice?" in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 2076–2084.

[15] L. Zhang, M. Ahmadi, J. Pan, and L. Chang, "Metropolitan-scale taxicab mobility modeling," in *Proc. IEEE Global Commun. Conf.*, Dec. 2012, pp. 5404–5409.

[16] X. Hou, Y. Li, D. Jin, D. O. Wu, and S. Chen, "Modeling the impact of mobility on the connectivity of vehicular networks in large-scale Urban environment," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2753–2758, Apr. 2015.

[17] D. Zhao, H. Ma, L. Liu, and X.-Y. Li, "Opportunistic coverage for Urban vehicular sensing," *Comput. Commun.*, vol. 60, pp. 71–85, 2015.

[18] H. Zhu, M. Li, L. Fu, G. Xue, Y. Zhu, and L. M. Ni, "Impact of traffic influxes: Revealing exponential intercontact time in Urban VANETs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 8, pp. 1258–1266, Aug. 2011.

[19] M. Gao, T. Zhu, X. Wan, and Q. Wang, "Analysis of travel time patterns in Urban using taxi GPS data," in *Proc. IEEE Int. Conf. Green Comput. Commun. IEEE Internet Things IEEE Cyber Physical Social Comput.*, Aug. 2013, pp. 512–517.

[20] C. Xia, D. Liang, H. Wang, M. Luo, and W. Lv, "Characterization and modeling in large-scale Urban DTNs," in *Proc. IEEE 37th Conf. Local Comput. Netw.*, Oct. 2012, pp. 352–359.

[21] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, pp. 29:1–29:41, May 2015.

[22] H. Su, K. Zheng, J. Huang, H. Wang, and X. Zhou, "Calibrating trajectory data for spatio-temporal similarity analysis," *VLDB J.*, vol. 24, no. 1, pp. 93–116, Feb. 2015.

[23] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, and A. Rabuffi, "CRAWDAD dataset Roma/Taxi (v. 2014-07-17)," Jul. 2014. [Online]. Available: http://crawdad.org/roma/taxi/20140717

[24] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "CRAWDAD dataset EPFL/mobility (v. 2009-02-24)," Feb. 2009. [Online]. Available: http://crawdad.org/epfl/mobility/20090224

[25] SUVnet, "Shanghai data trace," Feb. 2009. [Online]. Available: http://wirelesslab.sjtu.edu.cn/taxi_trace_data.html

[26] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 316–324.

[27] Y. Chen, M. Xu, Y. Gu, P. Li, L. Shi, and X. Xiao, "Empirical study on spatial and temporal features for vehicular wireless communications," *EURASIP J. Wireless Commun. Netw.*, vol. 2014, no. 1, pp. 1–12, 2014.

[28] Wisemap, "Urban mobility," Mar. 2016. [Online]. Available: www.wisemap.dcc.ufmg.br/urbanmobility

[29] F. A. Silva, C. Celes, A. Boukerche, L. B. Ruiz, and A. A. F. Loureiro, "Filling the gaps of vehicular mobility traces," in *Proc. 18th ACM Int. Conf. Model. Anal. Simul. Wireless Mobile Syst.*, 2015, pp. 47–54.

[30] W.-C. Lee and J. Krumm, "Trajectory preprocessing," in *Computing with Spatial Trajectories*. Berlin, Germany: Springer, 2011, pp. 3–33.

[31] G. Trajcevski, "Uncertainty in spatial trajectories," in *Computing with Spatial Trajectories*. Berlin, Germany: Springer, 2011, pp. 63–107.

[32] C.-S. Yang, S.-P. Kao, F.-B. Lee, and P.-S. Hung, "Twelve different interpolation methods: A case study of surfer 8.0," in *Proc. 20th Int. Congr. Photogrammetry Remote Sens.*, 2004, pp. 778–785.

[33] J. Li and A. D. Heap, "A review of spatial interpolation methods for environmental scientists," Geoscience, Canberra, Australia, *Record 2008/023*, 2008.

[34] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Comput. Netw.*, vol. 64, pp. 296–307, 2014.

[35] S. Liu, C. Liu, Q. Luo, L. Ni, and R. Krishnan, "Calibrating large scale vehicle trajectory data," in *Proc. IEEE 13th Int. Conf. Mobile Data Manage.*, Jul. 2012, pp. 222–231.

[36] S. Uppoor, O. Trullols-Cruces, M. Fiore, and J. Barcelo-Ordinas, "Generation and analysis of a large-scale Urban vehicular mobility dataset," *IEEE Trans. Mobile Comput.*, vol. 13, no. 5, pp. 1061–1075, May 2014.

[37] V. Naumov, R. Baumann, and T. Gross, "An evaluation of inter-vehicle ad hoc networks based on realistic vehicular traces," in *Proc. 7th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2006, pp. 108–119.

[38] C. Celes, R. B. Braga, C. T. De Oliveira, R. M. Andrade, and A. A. Loureiro, "GeoSPIN: An approach for Geocast routing based on spatial information in VANETs," in *Proc. IEEE 78th Veh. Technol. Conf.*, 2013, pp. 1–6.

[39] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser, "A parsimonious model of mobile partitioned networks with clustering," in *Proc. 1st Int. Commun. Syst. Netw. Workshops*, Jan. 2009, pp. 1–10.

[40] H. Huang, D. Zhang, Y. Zhu, M. Li, and M.-Y. Wu, "A metropolitan taxi mobility model from real GPS traces," *J. Universal Comput. Sci.*, vol. 18, no. 9, pp. 1072–1092, May 2012.

[41] Y. Zheng, "T-drive trajectory data sample," Aug. 2011. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=152883

[42] J. Yuan, et al., "T-drive: Driving directions based on taxi trajectories," in *Proc. 18th SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2010, pp. 99–108.

[43] L. Cheng, B. Henty, D. Stancil, F. Bai, and P. Mudalige, "Mobile vehicle-to-vehicle narrow-band channel measurement and characterization of the 5.9 GHz dedicated short range communication (DSRC) frequency band," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 8, pp. 1501–1516, Oct. 2007.

[44] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Sep. 2006.

[45] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[46] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. AAAI Workshop Knowl. Discovery Data Mining*, 1994, pp. 359–370.

[47] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2005, pp. 491–502.

[48] A. Cornejo, C. Newport, S. Gollakota, J. Rao, and T. J. Giuli, "Prioritized gossip in vehicular networks," *Ad Hoc Netw.*, vol. 11, no. 1, pp. 397–409, 2013.

[49] B. Ishibashi and R. Boutaba, "Topology and mobility considerations in mobile ad hoc networks," *Ad Hoc Netw.*, vol. 3, no. 6, pp. 762–776, 2005.

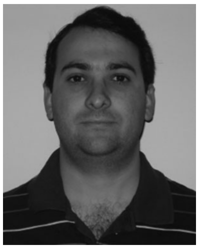
[50] R. Du, C. Chen, B. Yang, N. Lu, X. Guan, and X. Shen, "Effective Urban traffic monitoring by vehicular sensor networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 1, pp. 273–286, Jan. 2015.

[51] D. Jiang and L. Delgrossi, "IEEE 802.11p: Towards an international standard for wireless access in vehicular environments," in *Proc. Veh. Technol. Conf. Spring*, May 2008, pp. 2036–2040.

[52] Y. L. Morgan, "Notes on DSRC and wave standards suite: Its architecture, design, and characteristics," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 4, pp. 504–518, Oct.–Dec. 2010.

[53] C. Perkins, E. Belding-Royer, and S. Das, "RFC 3561 ad hoc on-demand distance vector (AODV) routing," 2003. [Online]. Available: https://www.ietf.org/rfc/rfc3561.txt

Clayson Celes received the bachelor's degree in computer science from the State University of Ceara (UECE), Brazil, in 2010 and the MSc degree in computer science from Federal University of Minas Gerais (UFMG), in 2013. He is currently working toward the PhD degree in computer science at UFMG, Brazil. His research areas are mobile computing, vehicular networks, data analysis, and ubiquitous computing.



Fabrício A. Silva received the BSc and MSc degrees in computer science from the Universidade Federal de Minas Gerais (UFMG), Brazil, in 2004 and 2006, respectively, and the PhD degree in computer science from UFMG, in 2015. During 2014, he held a visiting researcher position at PARADISE Research Laboratory, University of Ottawa. Currently, he is an associate professor with the Universidade Federal de Viçosa, Brazil. His research interests include mobile/ad-hoc networks, distributed systems, and mobile data analysis.

Azzedine Boukerche is a full professor and holds a Canada Research Chair Tier-1 position with the University of Ottawa. He is founding director of the PARADISE Research Laboratory and the DIVA Strategic Research Centre, University of Ottawa. He has received the C. Gotlieb Computer Medal Award, Ontario Distinguished Researcher Award, Premier of Ontario Research Excellence Award, G. S. Glinski Award for Excellence in Research, IEEE Computer Society Golden Core Award, IEEE CS-Meritorious Award, IEEE TCPP Leaderships Award, IEEE ComSoc ASHN Leaderships and Contribution Award, and University of Ottawa Award for Excellence in Research. He serves as an associate editor for several IEEE transactions and ACM journals, and is also a Steering Committee Chair for several IEEE and ACM international conferences. His current research interests include wireless ad hoc and sensor networks, wireless networking and mobile computing, wireless multimedia, QoS service provisioning, performance evaluation and modeling of large-scale distributed and mobile systems, and large scale distributed and parallel discrete event simulation. He has published extensively in these areas and received several best research paper awards for his work. He is a fellow of the Engineering Institute of Canada, the Canadian Academy of Engineering, the American Association for the Advancement of Science, and the IEEE.

Rossana M. C. Andrade received the bachelor's degree in computer science from the State University of Ceará (UECE), Brazil, in 1989, the master's degree from the Federal University of Paraíba (nowadays called Federal University of Campina Grande), Campina Grande, Brazil, in 1992, and the PhD degree from the School of Information Technology and Engineering (SITE), University of Ottawa, Ottawa, Canada, in May 2001. She is an associate professor in the Department of Computer Science, Federal University of Ceará (UFC), Brazil. Her PhD thesis focused on the capture, reuse, and validation of software patterns for mobile systems. She is the founding director of GREat Research Group, UFC. She does research in the areas of computer networks and software engineering.

Antonio A. F. Loureiro received the BSc and MSc degrees in computer science from UFMG, and the PhD degree in computer science from the University of British Columbia, Canada. He is a full professor with UFMG, where he leads the research group on mobile ad hoc networks. He was the recipient of the 2015 IEEE Ad Hoc and Sensor (AHSN) Technical Achievement Award. He is a regular visiting professor and researcher in the PARADISE Research Laboratory, University of Ottawa, and is an international research partner of DIVA Strategic Research Networks. His main research areas include wireless sensor networks, mobile computing, and distributed algorithms. In the last 15 years, he has published regularly in international conferences and journals related to those areas, and has also presented tutorials at international conferences.