Urban Map Inference by Pervasive Vehicular Sensing Systems with Complementary Mobility

ZHIHAN FANG, Rutgers University, USA
GUANG WANG, Rutgers University, USA
XIAOYANG XIE, Rutgers University, USA
FAN ZHANG, SIAT, Chinese Academy of Sciences & Shenzhen Beidou Intelligent Technology Co., Ltd.
DESHENG ZHANG, Rutgers University, USA

Accurate and up-to-date digital road maps are the foundation of many mobile applications, such as navigation and autonomous driving. A manually-created map suffers from the high cost for creation and maintenance due to constant road network updating. Recently, the ubiquity of GPS devices in vehicular systems has led to an unprecedented amount of vehicle sensing data for map inference. Unfortunately, accurate map inference based on vehicle GPS is challenging for two reasons. First, it is challenging to infer complete road structures due to the sensing deviation, sparse coverage, and low sampling rate of GPS of a fleet of vehicles with similar mobility patterns, e.g., taxis. Second, a road map requires various road properties such as road categories, which is challenging to be inferred by just GPS locations of vehicles. In this paper, we design a map inference system called *coMap* by considering multiple fleets of vehicles with Complementary Mobility Features. *coMap* has two key components: a graph-based *map sketching* component, a learning-based *map painting* component. We implement *coMap* with the data from four type-aware vehicular sensing systems in one city, which consists of 18 thousand taxis, 10 thousand private vehicles, 6 thousand trucks, and 14 thousand buses. We conduct a comprehensive evaluation of *coMap* with two state-of-the-art baselines along with ground truth based on OpenStreetMap and a commercial map provider, i.e., Baidu Maps. The results show that (i) for the map sketching, our work improves the performance by 15.9% ; (ii) for the map painting, our work achieves 74.58% of average accuracy on road category classification.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**.

Additional Key Words and Phrases: map sketching, map painting, heterogeneous vehicular fleets, GPS traces

ACM Reference Format:
Zhihan Fang, Guang Wang, Xiaoyang Xie, Fan Zhang, and Desheng Zhang. 2021. Urban Map Inference by Pervasive Vehicular Sensing Systems with Complementary Mobility. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 48 (March 2021), 24 pages. https://doi.org/10.1145/3448076

1 INTRODUCTION

With the urbanization process, the public road length in developing countries has been increasing dramatically, e.g., China has doubled the road length in the past 10 years from 53,913 km in 2007 to 135,500 km in 2017 [4]. This rapid growth of road networks has led to tremendous costs for digital map maintenance. Various mobile

Authors' addresses: Zhihan Fang, Rutgers University, Piscataway, NJ, 08854, USA, zhihan.fang@cs.rutgers.edu; Guang Wang, Rutgers University, USA; Xiaoyang Xie, Rutgers University, USA; Fan Zhang, SIAT, Chinese Academy of Sciences & Shenzhen Beidou Intelligent Technology Co., Ltd. Desheng Zhang, Rutgers University, USA, desheng.zhang@cs.rutgers.edu.

2474-9567/2021/3-ART48 $15.00
https://doi.org/10.1145/3448076

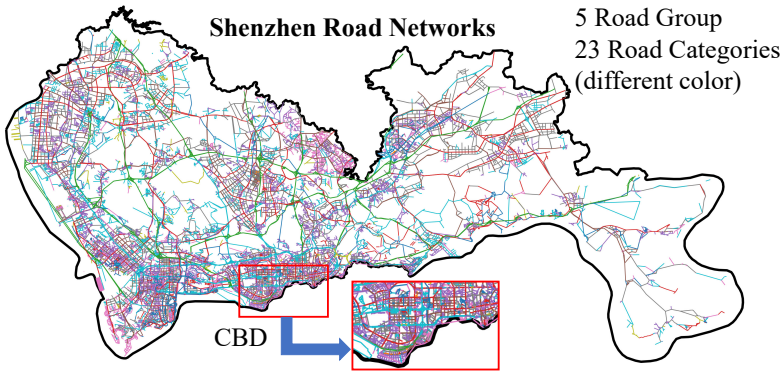


Fig. 1. Road Networks in Shenzhen (Better Visibility In Color)

Highway	motorway (5.8%), truck (4.9%), primary (11.5%)
Road	secondary (10.7%), tertiary (17.2%), road (0.07%), living street (0.1%), residential (13.7%)
Path	steps (0.3%), path (0.9%), lane (0.01%), footway (3.3%), cycleway (2.3%), pedestrian (0.08%)
Link	motorway link (2.3%), truck link (2.6%), primary link (2.8%), tertiary link (0.2%), secondary link (0.6%)
Special	bridleway(0.0001%), service (7.2%), track(0.8%), unclassified (11.3%)

Fig. 2. Road Categories

systems and services, such as navigation, Point of Interest (PoI) recommendation, Autonomous Driving, and Intelligent Transportation System, rely on accurate digital road maps for reliable operations [23] [30]. Current map service providers, e.g., Google Map, HERE, TomTom, and Baidu Maps, update road maps by sending probe fleets combined with satellite images, which requires significant human efforts and high monetary cost for map maintenance. Some crowded sourcing maps, e.g., Waze, required users to actively participate to mainly report real-time traffic condition. The company HERE reports millions of updates to their global database every day [13]. A similar cost is reported by TomTom [1] and OpenstreetMap [39] and Waze. TomTom estimates that 15% of roads change every year [1] including new road opening and existing road upgrades. In this context, an automatic way of creating and updating digital maps is essential because complete manual efforts are too costly in the recent decade and inefficient to satisfy the increasing demand.

Recently, the ubiquity of GPS sensors on vehicles or smart phones has generated massive amounts of GPS data from their host vehicles [10, 14, 28, 39, 43], which provide new opportunities for automatic map inference and updates. However, map inference and updates using vehicle traces is very challenging for two reasons. (1) A practical map requires not only road structures but also road properties, e.g., road categories, to provide practical services (e.g., navigation), which is challenging to infer based on just locations of vehicles. (2) Road structures are complicated in cities such as interchanges and parallel roads. Due to the GPS noises, it is challenging to separate those interchanges and parallel roads based on GPS observations.

Lots of works have been proposed to automatically produce maps based on GPS traces and other context data in different communities, e.g., SIGKDD [14] [33], SIGSPATIAL [28] [17] [10] [40], and SIGMOBILE Community [39] [14] (please see our related work section for details). Most of these works have explored the map inference by constructing road structures with kernel density models [7] [12] [31] or clustering methods [8] [37]. However, these works have not **explicitly** considered **Complementary Mobility Features (CMF)** of different vehicle types (e.g., taxis, private vehicles, buses, and trucks) and often (1) produce a map representation without road properties (e.g., road categories); and (2) fail to differentiate some close located parallel roads or interchanges in complicated situations with GPS with vehicle types only. To the best of our knowledge, few existing studies, if any, have inferred road properties such as road types based on trajectory data only for map inference. One of the major reasons for the above limitations is most existing works are built upon the monotony of vehicle mobility, which does not provide enough diversity to differentiate road structures and road properties. For instance, the number of vehicles and travel speeds on both secondary roads and residential roads are very close

without considering vehicle types, which makes it difficult to differentiate those two types of roads. Motivated by the observation that different types of vehicle fleets (taxis, buses, trucks, private cars) present diverse mobility patterns (number of vehicles on roads, travel speeds, etc) on roads, we explicitly integrate the **CMF** in map inference for both road structures and properties in our system design to improve the diversity for differentiating roads.

In this paper, we design a map inference system called *coMap* to address these two key challenges. The key idea of *coMap* is that we design two components to utilize **CMF** to infer a complete map presentation, i.e., road structures and road properties, with high coverage and quality by explicitly utilizing vehicle types. We implement our system in Shenzhen, a pilot city for the smart city initiative in China. We summarize our key contributions as follows:

- To best of our knowledge, our work is the first effort to explicitly consider type-aware vehicular fleets infer road networks with both structures and property. Our large-scale mobility pattern study enables us to comprehensively compare these vehicular fleets and analyze their complementary mobility patterns for urban road map inference, which is difficult to be obtained by previous studies on a single fleet or type-agnostic data. However, complementary mobility features (CMF) of different types of fleets also bring new challenges to map inference due to the different number of vehicles of each type on one road and their different speeds, which has not been recognized by type-agnostic methods. For the benefit of the IMWUT community, we will share a one week sample of our vehicle data.
- Technically, we design two key components to infer and update a complete road map: (i) a *map sketching* component to integrate CMF with a graph-based fusion model to infer high-quality road structures by integrating speeds, directions, and vehicle locations of multiple fleets; (ii) a *map painting* component to discover road categories by utilizing features of CMF with a learning model.
- We implement and evaluate *coMap* with two-month GPS data in Shenzhen from four type-ware fleets: an 18 thousand taxi fleet, and a 14 thousand bus fleet, and a 6 thousand truck fleet, and a 10 thousand private vehicle fleet. Based on our results, our key findings include: (1) coMap improves the precision by 16.9% from 0.623 to 0.728 compared with the KDE based method and outperforms the trajectory clustering method by 42.7%, and *coMap* performs better in downtown areas with a small radius. (2) coMap achieves the best performance on the road category classification, with a precision of 74.58%. In addition, fleet types and selected features have impacts on classification performance, e.g., the taxi fleet achieves the best performance on the road category classification, and the speed is the most important feature on the road category.
- Moreover, we design and evaluate an application of *coMap* for map updating with data from four fleets to detect and report 98.7km roads, which was not in OpenStreetMap. We verified our results with Google Earth as the ground truth.

2 PRELIMINARY

Vehicular Fleets: As part of a smart city initiative team in Shenzhen, we have access to GPS traces of four city-scale vehicular fleets, i.e., a bus fleet, a taxi fleet, a truck fleet, and a private vehicle (PV) fleet. The details about the data are given in Table 1. The bus and taxi data are collected for control and dispatching purposes. The truck and private data are collected by an insurance company for usage-based insurance where drivers consent to contribute their data for both business purposes with incentives of an insurance discount. Therefore, our *coMap* system is at no extra cost to the vehicular fleets since all data are collected for existing services.
(i) Bus GPS record include *id, longitude, latitude, timestamp, direction, speed* with an interval of 20 seconds. They have prefixed regular pattern due to their operating routes. *(ii) Taxi* GPS record include *id, longitude, latitude, timestamp, occupied or not, direction, speed* with an interval of 30 seconds. Their mobility patterns are more

Table 1. Heterogeneous Vehicular Fleets

Fleet	Vehicles	Daily Size	Daily Record
Taxi	18,288	6 GB	66 million
Bus	14,514	6.5 GB	44 million
Truck	6,002	1.5 GB	32 million
Private Vehicle	10,021	0.4 GB	8 million

randomly. *(iii) Private Vehicle (PV)* GPS record include *id, longitude, latitude, timestamp, direction, speed* with an internal of 10 seconds for pay-as-you-go insurance programs. Their mobility patterns are mainly for commutes and personal trips. *(iv) Truck* GPS records includes *id, longitude, latitude, timestamp, direction, speed* with an interval of 45 seconds. Similar to private vehicles, the truck traces are accessed with onboard GPS recorders installed by an insurance company. Their mobility patterns are focused on highways, industrial areas, and residential areas for delivery.

Road Networks: OpenStreetMap [27] is one of the most active crowd-sourcing map services. To date, there are 72,676 road segments in Shenzhen from OpenStreetMap, of which the total length is 10,711 *km*. Since OpenStreetMap includes detailed branches and paths and differentiates lanes from opposite directions, we found the road length calculated from the OpenStreetMap database is larger than the road reported by governments. The geographic distribution of roads is presented in Figure 1, which shows different road categories with colors. Specifically, the OpenStreetMap labels roads with 24 types in 5 groups based on the functions of the roads. The details of the functions are given in [41]. Figure 2 lists the 24 types of roads and their percentages in Shenzhen, i.e., the road length of each road category in the total road length.

3 MOTIVATION

Our motivate includes two aspects: i) the real-world demand of the automatic map inference is rapidly increasing; ii) the mobility patterns of heterogeneous fleets are complementary.

3.1 Rapid Road Growth

Since road networks build the foundation of economic growth, we have witnessed a rapid road expansion in developing countries, e.g., China and India [3] [4]. We study the road length increase from 2000 in China based on the census data [5]. As in Figure 3, the length of roads in China has been increasing by 185.6% from 1.67 million km in 2000 to 4.77 million km in 2017. In particular, the rapid developing areas have experienced a more significant increase in road expansion, e.g., in Figure 3, the road network has been increasing by 356.8% from 44.4 thousand km in 2000 to 203.2 thousand km in 2016 in Anhui province, a province in China with an area of 140 thousand km^2. The rapid change of road networks has introduced tremendous costs for digital map creations and maintenance [1] [41] [9].

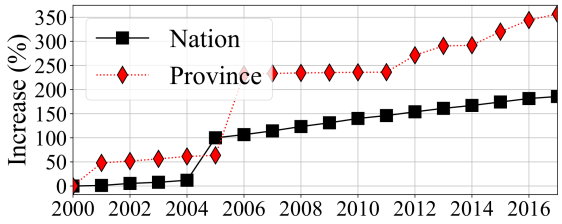


Fig. 3. Road Length Increase

3.2 CMF for Map Inference

We motivate our design philosophy with CMF for map inference (for both structure and property inference) by vehicle distribution differences on roads. We found the GPS distributions in vehicular fleets are not identical on any two roads especially roads with different road categories. *(i) Inter-category difference:* We select two roads in the downtown area in Shenzhen as an example, i.e., a residential road and a primary road, as shown in Figure 4. Even these two roads a primary road *A* and a residential road *B* are closely located, the vehicle distributions on the roads are different. The selected residential road *B* is dominated by private vehicles; whereas there are much more trucks than the other three types of vehicles on the primary road *A*. To generalize to citywide, we found trucks and taxis mostly move on the main road; whereas buses and personal vehicles move on the branch roads (i.e., non-main road). To quantify the difference of mobility patterns on the road categories, Figure 5 presents the GPS distribution on five categories.

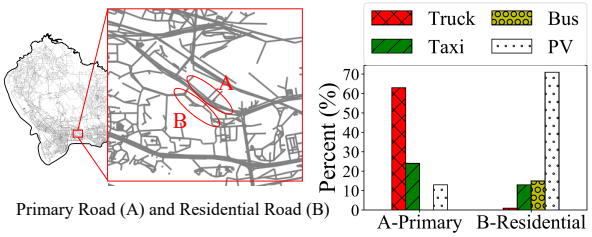


Fig. 4. Road Types

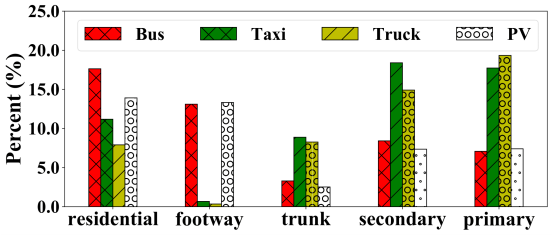


Fig. 5. Percentages on Five Major Road Types

The GPS distribution is calculated by the ratio of the number of GPS observations on a specific road category in the total GPS observations. There are more buses and private vehicles on residential roads and footways. In contrast, we found more taxis and trucks on the trunk roads, primary roads, and secondary roads. *(ii) Intra-category difference:* Even for two roads in the same category, the vehicle distributions are not identical. We quantify the vehicle distribution in terms of the four types of fleets and calculate the similarity of vehicles distribution on roads As a result, we found the average Pearson similarity for roads in the different categories is 0.32, and the average Pearson similarity for roads in the same road category is 0.48. It indicates the vehicle distribution shows a significant difference on different roads, and the difference is larger for roads from different road categories.

The *inter-category* and *intra-category* differences introduced by CMF motivates us (i) to improve road structure inference, i.e., differentiate parallel roads; and (ii) to infer road properties such as road categories. For instance, if a new road is open and the system detects more GPS observations from private vehicles and buses, it has a high possibility that the new road is a residential road instead of a primary road or a trunk. Furthermore, although CMF can make a more dense and diverse vehicle coverage, it also causes challenges to infer road structures due to the different number of vehicles of each type on one road and their different speeds, which has not been recognized by type-agnostic methods. Instead, *coMap* is a type-aware system utilizing CMF to improve the map inference for both road structures and properties.

3.3 CMF for Road Coverage

We define road coverage as percentage of road segments covered by fleets in total road segments. Previous works study the map inference based on a single fleet, e.g., taxi [33], which is potentially restricted by the sparse coverage and mobility pattern. A single fleet may only cover certain roads, e.g., most branch roads or narrow paths can be barely discovered by taxis, which restricts the spatial coverage on the map inference. Consolidating heterogeneous fleets enhances coverage since some of their mobility patterns are complementary. We study the hourly spatial coverage of road segments in four fleets in Figure 6.

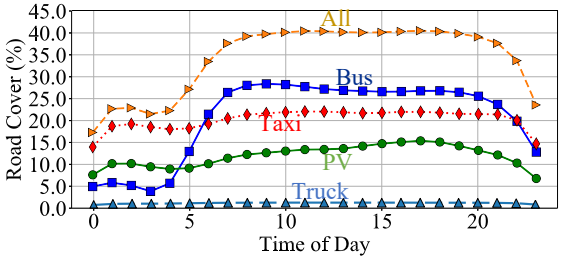


Fig. 6. Spatial Coverage

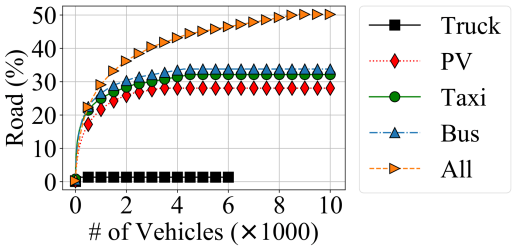


Fig. 7. Road Coverage with Vehicle Increase

We found the bus, taxi, private vehicle, and truck fleets covers 28.4%, 22.2%, 15.9%, and 1.22% of 72 thousand road segments, respectively. More importantly, we found not only fleet sizes but also their mobility patterns make a difference.

As shown in Figure 7, we study the CDF of the road coverage distributions. With the increase of the number of vehicles, the spatial coverage increases accordingly only during the initial increase of the fleet size. During these initial increases, the spatial coverage is limited by fleet sizes. However, when the fleet sizes reach a certain value, i.e., 33.8% for the bus fleet, 28.1% for the private vehicle fleet, 32.2% for the taxi fleet, 1.36% for the truck fleet, the increase of fleet size does not lead to a higher coverage increase. We define the threshold value of the road coverage as ***Ceiling Point***, from which increasing certain types of vehicles cannot change the road coverage significantly. For example, the bus, taxi, private vehicle, and truck fleets have a *Ceiling Point* of 33.8%, 32.2%, 28.1%, 1.36%, respectively.

Combining heterogeneous fleets with different types of vehicles increase the *Ceiling Points*, e.g., in our study, the *Ceiling Point* is 50.1% with 4 fleets and we believe integrating more fleets, e.g., shared bikes, can lead to a higher *Ceiling Point*. When combining heterogeneous fleet data, we achieve 50.1% road coverage, which improves the road coverage from an average of 16.9% (i.e., (28.4%+22.2%+ 15.9%+1.22%)/4) when using a single fleet to 50.1%. The major reason for the ceiling point is that there is an uneven distribution of vehicles on different types of roads. For instance, the road coverage of trucks is low because trucks are mostly distributed on highways and primary roads. The combination of the four types of vehicles covers only 50.1% of the roads and left uncovered roads are mostly restricted by other vehicle types, e.g., pedestrian, bike roads, steps, and bridleways, as illustrated in Fig. 5. Even though we only have data access to the four types of vehicles, the design philosophy and insights, i.e., utilizing CMF, can be applied and extended to more vehicle types to increase the road coverage. For instance, there are currently large-scale bike-sharing systems deployed in cities including Shenzhen. When integrating such datasets, the spatial coverage will increase since bicycle roads can be discovered by the GPS on bicycles. Similarly, other roads such as bridleways can be reported with extra sensing devices, e.g., cellphones, or by combining with satellite image data.

4 COMAP DESIGN

Definition: We use a graph $\mathcal{G} = (V, E, P)$ to present a digital map. In a standard graph, V is the collection of vertices of the graph, E is the collection edges of the graph, P is the collection of edge properties for all edges. In the map inference, a *road point* is defined as a vertex in V, a *road segment* between two road points is defined as an edge in E, a *road property* such as road category is defined as an edge property.

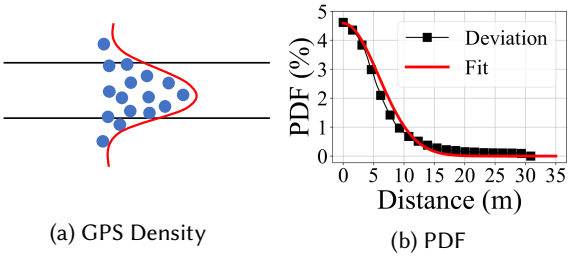
Design Philosophy: Most of the existing studies focused on map inference $\mathcal{G} = (V, E)$, which lack road property P in the inferred map. Moreover, their accuracy on road structure inference, i.e., V and E, is highly impacted by complex road environments such as crossroads and interchanges. To solve the above two challenges, we integrate CMP in map inference with two key components. (i) a *map sketching* component to infer road structures with

high accuracy especially in regions with complex road distribution, i.e., we infer road points V and road segments E for $\mathcal{G}$; (ii) a *map painting* component to label road segments with road properties such as road categories, i.e., we infer road properties P for $\mathcal{G}$. Specifically, in *map sketching*, we design a Type-Aware mean-shift sampling method to iteratively select road points, V. In *map painting*, we explore a type-aware learning method to bright the gap between raw GPS distribution and road properties.

4.1 Map Sketching

The type-aware map sketching is implemented with three steps: (i) Road Point Identification (Construct V): we identify the points by combining prior knowledge with GPS information, which includes density, speed, and direction. To address the challenge of the sparse and biased spatial coverage of a single fleet, we utilize GPS of four heterogeneous type-aware fleets to feed this task. Prior knowledge is the correlation between roads and heterogeneous fleets. (ii) Road Segment Construction (Construct E): we link the road points based on the direction and smoothness of the road structures. (iii) Link and Turn Inference (Optimize E): we complete the links between inferred road segments and in the intersections by inferring missing road segments. We use the area in Figure 4 (Ground Truth map) as an example to illustrate how our *coMap* system starts from raw GPS points (in Figure 9) to infer road points V (in Figure 10) and road segments E (in Figure 12).

4.1.1 Road Point Identification. Previous works assume the road central lines are the lines with the highest density [18] [42] [15]. We calculate the distance between the GPS points and their actual roads in four fleets. The GPS density on the road is highest, and the GPS density decreases as the distance between GPS and road increases, e.g., towards the edge of roads. It can be fit by a normal distribution $0.67 \times \mathcal{N}(0, 5.8^2)$ as shown in Figure 8.



(a) GPS Density (b) PDF

Fig. 8. GPS Deviations

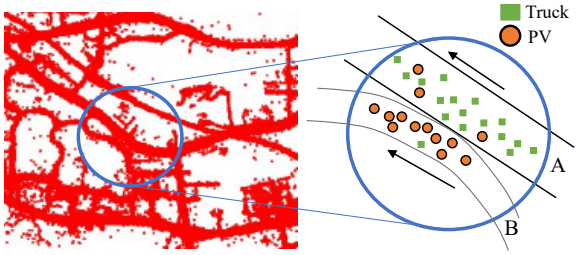


Fig. 9. Left: Raw GPS Points; Right: Vehicle Type Distribution on Underlying Unknown Roads.

Challenges: However, identifying road points with only density of GPS is challenging in certain cases, e.g., sparse GPS data, overpasses, opposite lanes, or parallel roads. To illustrate the challenge and motivate CMF, we use an example with two parallel roads in Figure 9. The left side of Figure 9 shows the GPS point distributions from vehicular fleets; the right side of the figure shows different vehicles on underlying roads. There are a primary road A and a residential road B to be inferred. Because the two roads are closely located, the GPS noise from vehicles on the road A affects GPS distribution on road B, and vice versa.

Type-Aware Sampling with CMF: The traditional methods for estimating density only consider the number of GPS points in the surrounding areas [7] [31], which do not consider the different mobility patterns of complementary fleets. Motivated by the vehicle distribution difference of CMF in our motivation section, we conduct Type-Aware mean-shift sampling strategy for road point identification in *coMap*. The general idea is that we can assign dynamic weights to points based on different vehicle types. The weights are determined by the possibility of co-existence of different types of vehicles on the same roads based on existing observations, e.g., we may assign a higher weight to the point surrounded by similar types as the example in Figure 9. Similarly, other

factors such as heading direction and speed of GPS points are similar on the same road compared with different roads. With this idea, the weights of points are determined by three factors in our work, i.e., heading direction similarity, speed similarity, and vehicle type similarity. As an example in Figure 9, the private vehicles on the central line of road B and the trucks on the central line of road A should be assigned with a higher weight since the surrounding GPS observations from other vehicles have a high similarity on directions, speeds, and vehicle types. We quantify the weights with a weight function in Eq. 1,

$$g(x_i, k) = \sum_{x_j \in \Omega(x_i)} I(Grid(x_j), k) W(x_i, x_j)$$
$$W(x_i, x_j) = W_{type}(x_i, x_j) \cdot W_{heading}(x_i, x_j) \cdot W_{speed}(x_i, x_j) \tag{1}$$

where x_i is a GPS point; k is a grid number; $\Omega(x_i)$ is collection of GPS points in grids at a direction that is perpendicular to the direction of x_i; $Grid(x_j)$ returns the grid number of x_j and $I(Grid(x_j), k) = 1$ when x_j is in grid k and otherwise 0. W_{type} is calculated by Pearson correlation [6] of two types of vehicles on the same roads; $W_{heading}$ is calculated by the $cos(heading(x_i) - heading(x_j))$; W_{speed} is calculated by the reciprocal of the absolute value difference between x_i and x_j, i.e., the higher the difference, the lower the weight. Based on this weight function, we identify the road points with an EM (Expectation-Maximization) algorithm. For a given GPS point x_i, we project nearby points to the heading direction of x_i and calculate the histogram in grids based on Eq. 1. In detail, we draw a line that is perpendicular to the direction of x_i and divide the perpendicular line into grids with equal length and a certain width. Therefore, if there are more points similar to x_i in grid k, we can get a larger $g(x_i, k)$. We select a sample point s_i as the representative of x_i in the peak grid of the histogram of x_i. We set s_i as x_i and iterate this process until all x_is have been covered by sample points, and those sample points are identified road points. Therefore, In the E-step, we search for peak grids based on sample points; in the M-step, we re-select sample points from the peak of the grids. Based on the iteration, the road points will be shifted to sample points with higher weights defined in Eq. 1. As we illustrated in the example area, a sample point with more surrounding vehicles of the same types, more similar directions, and more similar speed is more possible to be located in a road or lane center. In other words, in every iteration, we change the road points from identified point from the last iteration to new points with higher weights (more surrounding vehicles with same types, more similar with same directions, and more similar speed). We stop the iteration when the process converges to a certain set of road points. In this process, the data errors are dropped at each iteration since they are always at the edge of the roads (a lower weight on type, heading, and speed). Moreover, this algorithm only relies on data samples, thus is not restricted by low traffic or observations on roads. In Figure 10, black points represent the identified road points V with weights of directions, speeds, vehicle types, and densities.

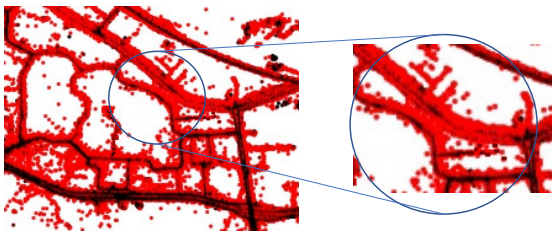


Fig. 10. Road Point Identification (black points are identified as road points and red points are noises).

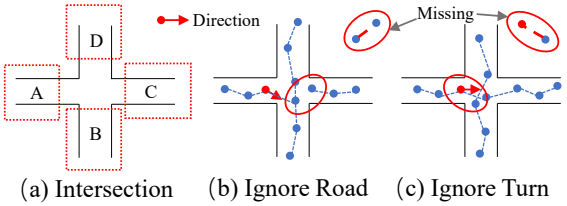


Fig. 11. Complete Links on Intersection

4.1.2 Road Segment Construction. In the second stage, we link the previously obtained road points to road segments.
Challenges: Some turns and important links may be ignored in the road segment construction since the road points are samples. For example, Figure 11 shows an intersection connecting four road segments, i.e., A, B, C, D. Since our road segment construction links road points by the direction of the vehicles, it may lead to two cases in the intersections . In Figure 11 (b), the red arrow is the direction of the vehicle, the road segment construction constructs a right turn based on the direction but ignores the straight road. In Figure 11 (c), it ignores the turn since the direction of the sample point (red arrow) points to the parallel road. Therefore, based on our road segment construction, if the direction of the road point (red arrow) is from A to B, the construction algorithm ignores the road construction connecting A and C. Thus, it leads a missing segment between A and C. As a result of such an algorithm property, there is a high probability for missing segments at road intersections.
Opportunities of CMF: To solve the problem, we first design an identification method to identify areas with missing segments based on the vehicular type and CMF (Identification). Second, we apply a patching algorithm to infer the missing road segments in the identified areas (Patching).

Identification: To infer the missing segments, we first identify the following two items, i.e., disconnected road segments and intersections, since the missing segments happen at intersections of roads and lead to disconnected segments: (i) disconnected segments by finding edges with one end point with zero in-degree yet non-zero out-degree; (ii) intersections, which are given as follows. In order to identify an intersection, we explore one of its key features: there is a high randomness of directions and vehicle distributions of different fleets (CMF). So we identify locations of intersections by searching for locations with high randomness of direction, speed, vehicle distribution compared with surrounding areas. This randomness is calculated by the distribution difference between the location and the surrounding areas, i.e., the entropy of direction, speed, and different vehicle types.

Patching: Based on the identified disconnected segments and intersections, we identify missing segments as follows. In a range of areas centering the identified intersections, we apply a segment matching algorithm [24], which is similar to a map matching algorithm, to map trajectories from heterogeneous fleets to the inferred road segments. For each trajectory crossing two disconnected road segments, the segment vote between the two road segments will increase by one, which is denoted as a vote from real-world trajectories. If the vote for the link between the two disconnected segments is larger than a threshold, we connect the two road segments. Further, we compare the difference between the two road segments in terms of directions and the number of vehicle distributions from complementary fleets. If the difference is small, we label the segment a normal road segment, otherwise, we label the segment as a turn since a normal road segment is smooth while a turn has a certain angle.
Summary: In short, we construct a road structure by our map sketching with (i) Road Point Identification, (ii) Road Segment Construction. Figure 12 presents the connected road segments on the two parallel roads with the three steps. Comparing Figure 9, Figure 10, Figure 12 and the ground truth in Figure 4, we found based on the two steps, the close located parallel roads can be clearly separated and identified in *coMap*.

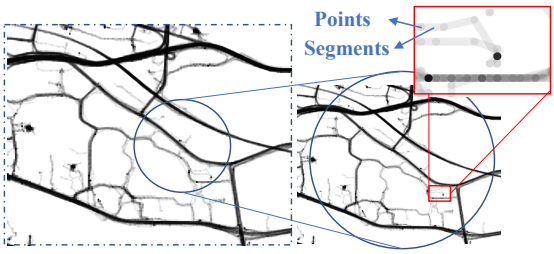


Fig. 12. A Constructing Map (Truth is in Fig. 4)

4.2 Map Painting

Based on road structures identified by the map sketching component, we add road properties in this map painting component. To achieve this goal, we learn the relation between GPS observations and road properties with a learning-based model. We study the road category, which is one of the most important road properties for applications such as navigation systems and autonomous driving. A complete list of road categories is given in Figure 2.

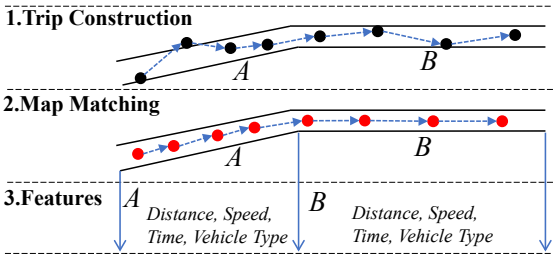


Fig. 13. Preprocessing for Map Painting

Fig. 14. Trajectory and Map Matching

4.2.1 Preprocessing. The preprocessing consists of three procedures, i.e., trip construction, map matching, and feature extraction. We briefly illustrate these three procedures in Figure 13, in which we show some GPS data of a vehicle passing through two connected roads, i.e., road A and road B.

① The *trip construction* groups the GPS locations by driver IDs, and sorts raw GPS into continuous GPS traces by time, and finally segment raw GPS observations into logical trips by a trip segmentation function. ② The *map matching* maps raw GPS points into existing roads, which consists of mapped GPS locations on the roads and road information such as road name, road id, road category, and road speed limit with a Hidden Markov Model [39] with an existing road map on OpenStreetMap [27].

Figure 14 presents the trip trajectories before and after the *map matching* in the taxi fleet. ③ The *feature extraction* function constructs the feature vectors for trips on each road. For instance, based on the trip in Figure 13, we extract four features for road A and B, i.e., (i) traveled distance of a vehicle on the road, (ii) average speed and speed variance on the road, (iii) time duration of the trip, and (iv) vehicle type.

4.2.2 Opportunities of CMF for Road-Category Learning. A road feature is a statistical value of vehicles on the road, e.g., number of taxis. For a road e_j, its road features are presented by a vector with 13 dimensions, a time slot t, and the rest of 12 dimensions are three key features (e.g., distance d, speed s, and number of vehicles n) for each of four vehicular fleets. Specifically, d is the maximum distance of all trips during the time slot on the road; s is the medium speed of all trips during the time slot on the road. We fill zero with no observation for a certain feature. Complementary Mobility Feature (CMF) can benefit map painting for the following reasons: (i) it will be possible to identify road categories with exclusive types of vehicles, such as bus roads only for buses; (ii) the difference of mobility features (such as travel speeds and number of vehicles) in single vehicle fleets are small among certain road categories; whereas the difference of these features among different vehicle types is large and distinctive due to Complementary Mobility Features.

4.2.3 Classification. Based on the road features and the road category label from existing human-created map, which is labeled by human effort, e.g., OpenStreetMap, we train an effective classifier from road features to the road categories. The impact of our learning model on the inference performance is investigated in the evaluation of Section 5.

4.2.4 Inferred Map Matching. The general map matching is based on existing human-created maps. Inferred map matching maps GPS points to the inferred map, which is the output of our map sketching. Similar to the general

map matching, in the first step, we construct trips with the same vehicle ID and continuous GPS points. In the second step, we apply a Hidden Markov Model [36] to match the GPS points to the inferred map structures. In the final step, we extract the same road features for inferred roads based on trajectories on the roads. We use these road features as input in the classifier to predict the road categories for the inferred roads.

5 EVALUATION

5.1 Implementation Details

Due to the data-driven nature of coMap, we introduce how to obtain and manage our fleet data as follows. We employ a high-performance cluster with Spark for data processing. The details are given as follows: (i) 12 HP machines with 2 Tesla K80c each; (ii) 10 Dell machines with 4 Tesla K80c each; (iii) 4 Xeon E5-2650 with a half TB memory each; (iv) A series of 800GB SSD and 15TB of spinning-disk spaces; (v) 2 PB additional disk space. We download the map from OpenStreetMap with a rectangular area covering Shenzhen and then filter the map by the city boundary. We extract the road networks from map with osm2pgsql (i.e., a command-line based program that converts OpenStreetMap data to postGIS-enabled PostgreSQL databases) and use PostgreSQL to store the existing road networks. Since the maximum distance a GPS observation and its matched road is 38 meters as shown in Figure 8, we apply a Hidden Markov Model map matching algorithm [36] with a 40 meter searching radius.

5.2 Evaluation Methodology

5.2.1 Ground Truth. We utilize two existing road maps, i.e., OpenStreetMap, and a commercial map, i.e., *Baidu Map*, as our ground truth to validate the accuracy of the map inference. OpenStreetMap is a crowd-sourcing data source, and we can access detailed attributes of each road segment such as locations, structures, and road types. However, some parts of the road network lack up-to-date information due to limited crowd-sourcing efforts. Instead, commercial maps have a higher coverage but limited access to certain attributes. Therefore, we evaluate our model on both types in data sources. Figure 15 presents the road network based on OpenStreetMap in the city, in which we use colors for different road categories.

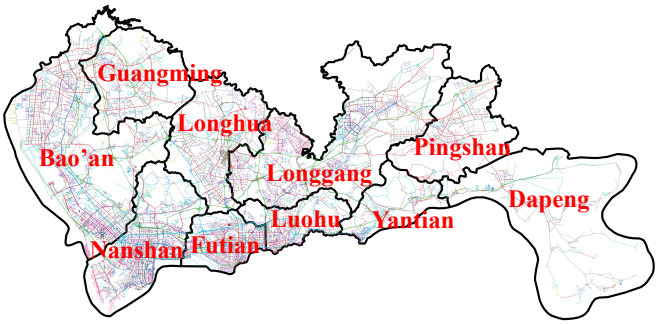


Fig. 15. City Administrative Districts

5.2.2 Setting. The total land area of Shenzhen is 2,050 km^2, including 10 administrative regions. In China, the urbanization process and road construction are closely related to the admin istrative regions. In general, urbanization and road construction start from the central business area of the city and then expands to the surrounding regions. For example, Longgang district near the border of the city has a higher rate of infrastructure constructions compared with the downtown Futian district. Therefore, we divide road networks in the city into 10 regions based on its administrative boundaries. We use all trajectory data from all fleets in the evaluation. First, we

apply a cross-validation of map inference on the region level by extracting prior knowledge, and then apply to the test region, we use the existing road network in the test region as our ground truth. Second, we simulate the urbanization and road construction process by extracting prior knowledge from the highly urbanized districts, e.g., Nanshan, and infer the road networks in the developing regions, e.g., Longgang. We divide the inferred map into two groups of roads, i.e., roads in the ground truth dataset and new roads. We evaluate the map sketching and map painting accuracy based on the first group of roads.

5.2.3 Metrics. We use precision and recall to evaluate road structure inference and use the accuracy score for road property inference.

(1) Metrics for Map Sketching of Road Structure Inference: we apply the evaluation method in [21] to evaluate the map topology and geometries for structure inference. Given the generated map $\widetilde{\mathcal{G}}$ and the real world map $\bar{\mathcal{G}}$. We randomly select n start points and apply a DFS search on both $\widetilde{\mathcal{G}}$ and $\bar{\mathcal{G}}$. We set a distance interval d and a threshold distance s. When searching $\widetilde{\mathcal{G}}$, we drop a label called *marble* every d distance until s is researched. When searching $\bar{\mathcal{G}}$, we start with the same start point, and we drop a label called *hole* with same d and s. Then we draw a circle on the holes with the radius r. A marble is matched with a hole when it is located in the circle of the hole. The precision and recall are defined based on the matched holes and marbles.

$$\begin{aligned} precision &= \frac{\#\ of\ matched\ marbles}{total\ \#\ of\ marbles} \\ recall &= \frac{\#\ of\ matched\ holes}{total\ \#\ of\ holes} \end{aligned} \tag{2}$$

(2) Metrics for Map Painting of Road Category Inference: we use the precision to evaluate the map painting, which is defined as the ratio between road segments with corrected road category labels and the total number of road segments.

$$accuracy = \frac{corrected\ labels}{total\ \#\ of\ road\ segments} \tag{3}$$

5.2.4 Baselines. We consider baselines for map sketching and painting, respectively.

(1) Baselines for Map Sketching of Road Structure Inference. In general, existing studies can be categorized into two major groups based on their technical contributions: kernel density models [7] [12] [15], trajectory clustering methods [8] [37], and intersection linking [22] [32]. (i) Kernel Density Estimation (KDE) methods [7] [12] [15]; (ii) Trajectory clustering solutions (Traj-C) methods [8] [37]; We select one of the state-of-the-art models in each group as our baseline models and label them by two groups. To be fair, all baselines are built upon four vehicular fleets with all trajectories. If a road is covered by fewer than four types of vehicles, we will fill their features in *coMap* as 0.

- *KDE [13]:* The roads are identified as the central line of Gaussian distribution with the highest GPS density in kernel density estimation models. KDE infers road structures by the density of the GPS distribution.
- *Traj-C [39]:* Trajectory clustering models cluster trajectories with close distance or similar moving patterns, which potentially suffers from the fact that road networks are dense in cities and it is challenging to differentiate clusters with a distance metrics.

(2) Baselines for Map Painting of Road Category Inference. Different from map sketching, few works have explored map painting. We compare our learning-based model with a greedy-based model, which searches the existing roads with the highest similarity in terms of road features.

- *RS (Road Search) [39]:* One of the alternative methods to label road properties is to search the most similar roads in terms of GPS distributions. Therefore, we implement a road search model with proposed features and compare its performance with our *coMap*.

- *coMap-:* One of the key designs of *coMap* is based on the complementary mobility features of multiple vehicular fleets. To justify the performance of our type-ware design, we implement *coMap-* as a type-agnostic version based on general features such as total traffic and average speed without differentiating vehicle types in different fleets.

5.2.5 Impact of Factors. We investigate the impact of four factors on *coMap*. (i) *Impact of Spatial Dimension on Map Sketching*. Since road networks are nonuniformly distributed in cities, e.g., dense road deployment in the central business district but sparse road deployment in suburban areas, the spatial difference influences the performance of road inference. (ii) *Impact of Learning Models on Map Painting*. We adopt different learning models in the training process of *map painting* to compare their performance difference. (iii) *Impact of Fleets on Map Painting*. Due to the difference in fleet sizes and mobility patterns, we study the impact of different fleets on the performance of *coMap*. (iv) *Impact of Selected Features on Map Painting*. The selected features, e.g., speed and traffic, have different weights on road property inference. We evaluate their impacts by dropping a specific feature.

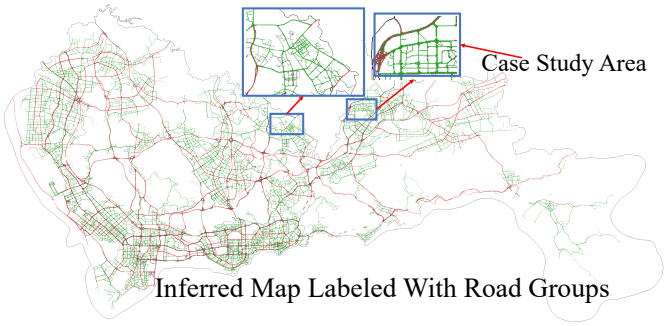
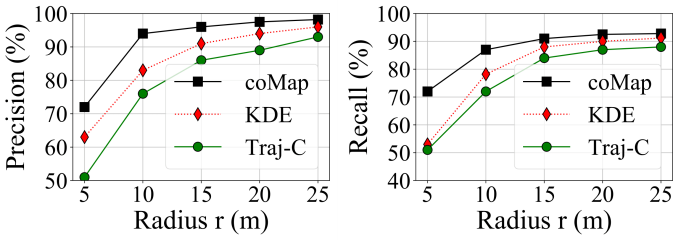


Fig. 16. Inferred Map

5.3 Evaluation Results

5.3.1 Qualitative Output Map. We present our qualitative result in Figure 16, in which we visualize the inferred road networks with cross-validation in administrative regions. The map is colored by inferred road category groups. We found less coverage compared with OpenStreetMap, which is caused by the ceiling point effect. However, we found some new roads, e.g., the case study area, on which we give a detailed case study in our map update evaluation.

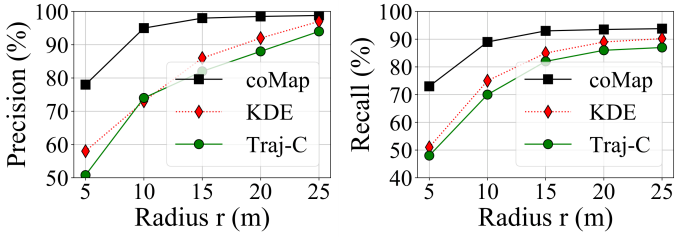
5.3.2 Map Sketching. We use cross-validation in the administrative regions to validate the quality of the inferred maps by comparing with both OpenStreetMap, and a commercial map, i.e., *Baidu Map*.



(a) Precision (b) Recall

Fig. 17. Performance of Map Sketching (OSM)

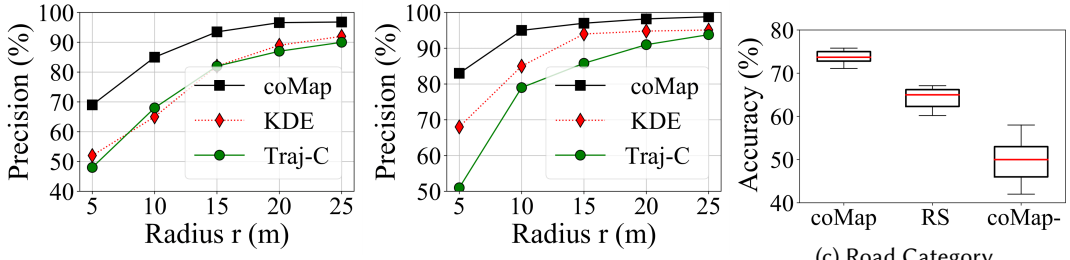
(1) Overall Performance using OpenStreetMap as Ground Truth. We compare *coMap* with two baseline methods in terms of the precision and recall in Figure 17. In the two figures, *coMap* shows better performance on the two metrics. We found with the increase of radius r (which is the threshold to match the marbles and holes), the precision and recall increase accordingly. When the radius is small, i.e., high accuracy, the difference of precision and recall between *coMap* and the baselines is large. The reason is that the *coMap* performs better on differentiating roads closely located or intersected. With the highest accuracy (5-meter radius), *coMap* improves the precision by 16.9% from 0.623 to 0.728 compared with the KDE based method and outperforms the trajectory clustering method by 42.7%. When the radius is larger than certain values, e.g., 25 meters, the spatial granularity is coarser than the local road difference and covers the missing roads, which leads to a close performance of three models.



(a) Precision (b) Recall

Fig. 18. Performance of Map Sketching (Baidu)

(2) Overall Performance using Commercial Map as Ground Truth. We collected data from one of the biggest map providers in China [2] to evaluate map inference. We report the results in Figure 18. Surprisingly, we found a higher precision and recall in *coMap* when we use commercial maps as ground truth. The reason is that compared with OpenStreetMap, the commercial map has a more complete road network coverage. Therefore, it reduces errors caused by incomplete map coverage in the ground truth data. On the other hand, we found a lower precision and recall in the baseline methods. The possible reason is that there are more branch roads in the commercial map, which are difficult to be differentiated by the baseline methods.



(a) Downtown (b) Suburban (c) Road Category

Fig. 19. Impact of Spatial Dimension

(3) Impact of Spatial Dimension. We further compare the performance under the impact of different spatial areas in Figure 19. The road density is much higher in downtown areas compared with suburban areas. It is challenging to infer roads from dense and complicated road structures in downtown areas. Therefore, the inference performance is low in downtown areas compared with suburban areas. Our model achieves a better road presentation in both downtown and suburban areas. Especially, *coMap* outperforms baseline models significantly in downtown areas with a high-quality map, i.e., 5 meters.

(4) Performance in Areas with Complex Road Structures. In particular, to investigate the performance in areas with complex road environment, we evaluate *coMap* in areas with dense road distributions. We divide the city into $0.5km \times 0.5km$ grids, which result in 180×92 grids in Shenzhen as shown in Fig. 20. We select 5% of grids with the

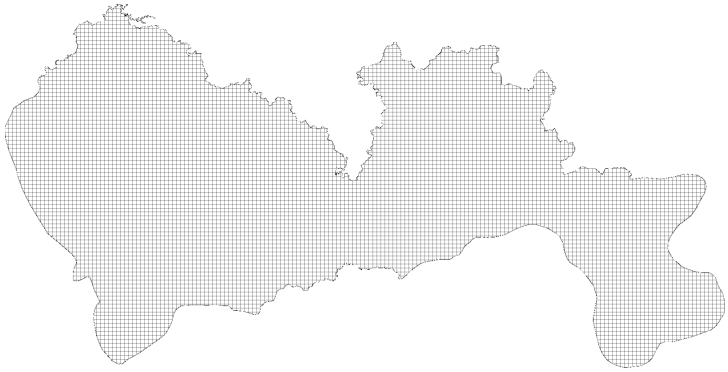
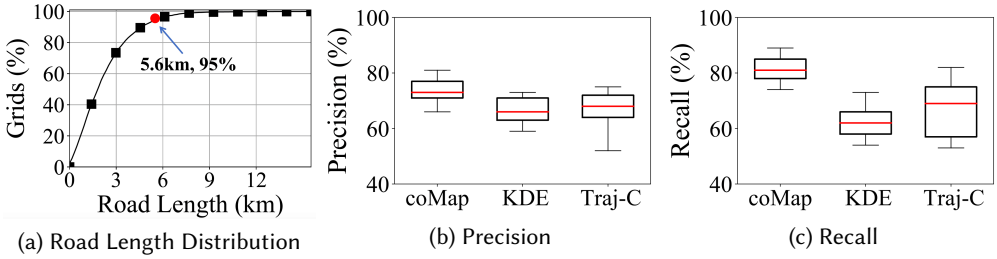


Fig. 20. Grid Regions

densest road distribution and compare the performance of *coMap* with baseline models with a 15-meter radius. We found those grids are located in either city centers or transportation hubs with complex road environment. The road length distribution is is given in Fig. 21a. We report the evaluation results in Fig. 21b and Fig. 21c, where *coMap* achieves the best results, i.e., 78% precision and 81% recall, compared with the baseline models. One of the major reasons is that we found a higher entropy, i.e., a more diverse distribution of vehicle types, in those grids. In other words, the impact of CMF is higher in the same areas where the map inference is more challenging.



(a) Road Length Distribution (b) Precision (c) Recall

Fig. 21. Performance in Regions with Complex Road Structures

(5) Impact of Cities and Fleets. We validate the model generality by implementing *coMap* in another city Hefei city, which is the capital city of a mid-area province in China. We use trajectory data from two types of vehicles, i.e., personal cars and trucks, and Fig. 22a and Fig. 22b present their mobility patterns. Different from the large metropolitan city Shenzhen (one of four tier-1 cities in China), Hefei is a tier-2 city with a central downtown and several small surrounding towns, which are not covered by truck data as shown in the heat maps as shown in the figures due to low traffic of trucks. We implement *coMap* based on these two datasets, and the evaluation results are given in Fig. 22c and Fig. 22d. Even with two types of vehicles, the performance gain of *coMap* is obvious as shown in the results. Based on the analysis, we found the performance of *coMap* keeps stable for different areas mainly for two reasons: (i) the areas with a smaller CMF gain (i.e., lower diversity of vehicle type distribution) mostly have clear road structures and lower road density, e.g., suburban areas. (ii) the areas with a larger CMP gain (i.e., higher diversity of vehicle type distribution) mostly have complex road structures and higher road

density, e.g., transportation hubs. Therefore, for both Hefei and Shenzhen, the performance of *coMap* are better compared with two baseline models.

(a) Personal Cars in Hefei (b) Trucks in Hefei (c) Precision (d) Recall

Fig. 22. Performance of Map Sketching (Baidu)

5.3.3 Map Painting. We first evaluate the performance and then investigate how three factors impact the *map painting* component, i.e., the impact of selection of learning model, mobility features and different vehicular fleets.

(1) Overall Performance. We compare *coMap* with two baseline methods in Figure 19c. The results show our model achieves higher performance than the two baseline models. Compared with *coMap-*, we found the different distribution of vehicles on the roads, which are important features to differentiate various road properties. The greedy-based search model finds the most similar road in terms of selected features. However, even though we utilize four large-scale vehicular fleets, the involved vehicles are sample observations of all traffic on roads. A single road with high similarity of features is biased due to the sparse observations in the *RS* method. More importantly, *coMap* is built on a learning model with the prediction time complexity of $O(1)$, and while the time cost of the search algorithm is $O(n)$. *(2) Impact of Learning Models.* To justify SVM in the road type classification,

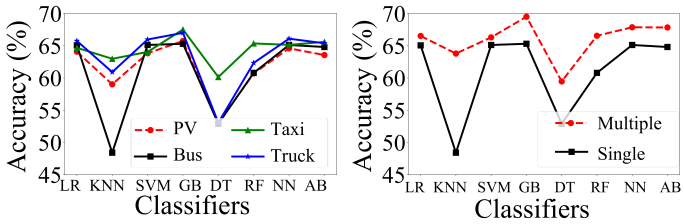
Table 2. Performance Comparison of Classifiers

CNN	Logistic Regression	Nearest Neighbors
62.69%	71.11%	67.73%
SVM	Gradient Boosting	Decision Tree
74.58%	72.31%	58.34%
Random Forest	Neural Network	AdaBoost
67.34%	72.39%	73.81%

we extract common features on the roads, i.e., speeds, distances, and volumes, and then evaluate the performance of *map painting* on the most widely used classifiers. The performances are listed in Table 2, in which SVM achieves the best performance on the road type classification.

(3) Impact of Fleets. We study the impact of fleets in the Figure 23. Figure 23a presents the results with a specific type of vehicular fleets; Figure 23b compares the single fleet with heterogeneous type-aware fleets with different classifiers. In the four vehicular fleets, the taxi fleet achieves the best performance on the road category

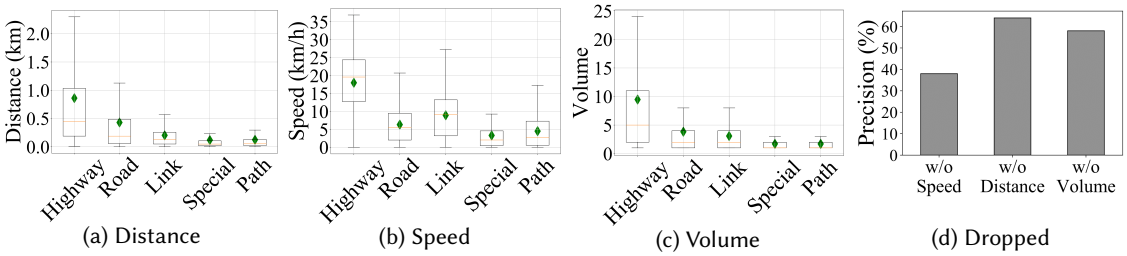
classification; while other three fleets present similar performance. The possible reason is that most taxi drivers are experienced drivers. As a result, their driving patterns follow the speed limits and diverse distributions on different road categories. Instead, even though buses have the highest spatial coverage and the largest ceiling point (as in Figure 7), the variance of their mobility is small, e.g., buses run with similar speeds on different types of roads. Comparing single fleets with multiple fleets, we found that multiple fleets show better performance on all the road category classifications.



(a) Single Type Fleet (b) Multi Type Fleet

Fig. 23. Impact of Vehicular Fleets

(4) Impact of Selected Features. We further investigate the impact of features on the classification, i.e., speeds, distances, and volumes in Figure 24. First, we study the distribution of features on different road groups for distances in Figure 24a, speeds in Figure 24b and traffic volumes in Figure 24c. The road distance and traffic volume have a similar distribution on different road categories; whereas the speed distribution is different. Thus, the speed is the most important feature on the road category. We study the performance difference after dropping a feature, and the result is given in Figure 24d, in which we test the performance after dropping a feature.



(a) Distance (b) Speed (c) Volume (d) Dropped

Fig. 24. Impact of Selected Features

6 APPLICATION: MAP UPDATES

We show a concrete application of our map inference for real-world map updates and evaluate it with Google Earth. We use Google Earth for evaluation of map update since it reflects the most recent changes on roads, which may be missing on either OpenStreetMap or even commercial maps.

6.1 Application Background

6.1.1 New Road Detection: We detect new roads and report them to the existing map service providers by combining our map inference with a map matching algorithm in Figure 25.

Given (i) a streaming GPS trace from our multiple fleets, (ii) an existing map (e.g., OpenStreetMap), and (iii) our inferred map, we first apply the map matching algorithm to map the GPS trace on both maps. Second, we

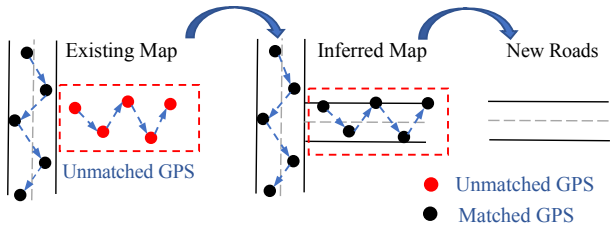


Fig. 25. New Road Detection

filter out the GPS points that have been matched in our inferred map but unmatched in the existing map, i.e., the red unmatched GPS points in Figure 25. Third, we use these unmatched GPS points to identify the roads that were matched to our inferred map but not matched to the existing map. We report these roads to the existing map. To reduce the false positive caused by noise, we apply a voting strategy in which the identified roads with at least n vehicles matched will be reported to the existing map. We empirically set $n = 3$ in the implementation. We label the identified roads with enough votes as the updated roads. We update identified new roads with road structures, road properties, and found time in our map.

6.1.2 Road Closure Detection: Roads could be temporarily closed (e.g., caused by construction or rain floods) or permanently closed (e.g., due to urban planning). Intuitively, a road can be marked if no trip is matched to this road for a long period of time. However, in practice, a trip can be matched to a closed road because of GPS noises, which leads to a closed road very challenging to be identified.

To solve the problem, we conduct a two-stage detection: (i) we label roads with a very small number of trips as candidates; (ii) we detect closed roads from these candidates with a statistical test. We compare the number of trips on a candidate road with its historical trip distributions. Specifically, we construct a 24-dimension vector, and each element in the vector is the number of matched trips in an hour. If the number of trips falls more than 2 standard deviation below the mean, i.e., 97.8% confidence, we label a *closed* in the hour. If more than 12 dimensions are labeled with *closed* in the vector, we label the day a *closed*. If a candidate road is closed for three consecutive days, we label the road as a temporarily closed road. We label a temporarily closed road back as an open road if the same test fails. A temporally closed road becomes a permanently closed road if the road remains closed with a long-time period, e.g., 2 months, and the road will be deleted on the active map and can be recovered by the new road detection function.

6.2 Application Evaluation

We evaluate our map update in the area where our map disagrees with OpenStreetMap. We compare the detected new roads with the latest satellite images on Google Earth to verify the correctness of our detection. We evaluate our map update in two ways, i.e., a case study for road open and external data source for road close.

6.2.1 Results on New Road Detection: In our map update, we found unmatched roads, which are not in the OpenStreetMap dataset. We use the existing map as prior knowledge for map inference, and report the missing roads or disagreements to the existing map service providers. With four heterogeneous data, we found 98.2 km unmatched roads, and we gave one example in Figure 26, which presents four roads in our inferred map, but not in the existing OpenStreetMap. The roads are detected by the *map update* component, which reports the missing roads to existing map services. Different from the previous map inference works, which are independent from the existing maps, our *coMap* constructs a complementary loop with existing maps, in which we extract prior knowledge from the existing maps, e.g., vehicle distribution on roads, and then send positive feedback to the existing maps for the map update.

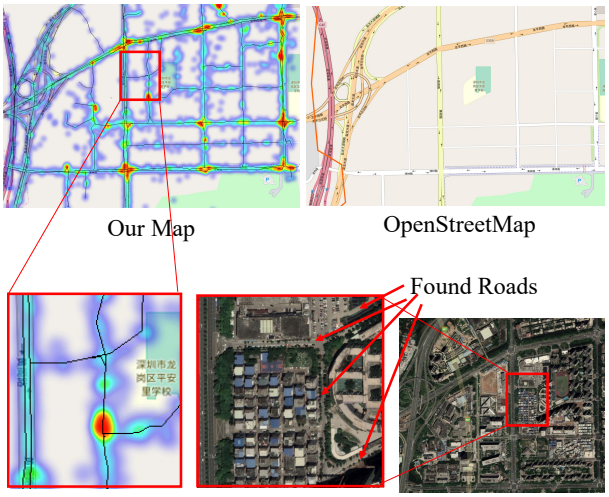


Fig. 26. New Road Captured By Multiple Fleets

6.2.2 Results on Road Closure Detection: We detected 18 temporary closed roads and 1 permanent close road with 2-month mobility data in the four fleets. We construct a ground truth data set based on natural language processing and extraction from tweets, i.e., official Weibo account (the Chinese equivalent of Twitter) of road transport police, and online news. We found all detected roads can be validated by the ground truths, i.e., 100% precision and 91% recall.

7 DISCUSSION

Lessons Learned: We summarize three key lessons based on our type-aware modeling, which has the potential to advance our understanding in this direction.

(i) Type-aware Modeling: Utilizing complementary mobility features of heterogeneous fleets to both separate close road segments and infer road category. We utilize a fundamental vehicular mobility nature where the density of different fleet types are different on different roads to design the road structure inference component in coMap. As shown in Figure 17, 18, and 19, coMap has a better capability to separate close road segments for better road structure inference based on our type-aware fleet density analyses, compared to the state-of-the-art models, which are mostly, if not all, type-agnostic. As shown in Figure 19c and Table 2, when combining mobility features from heterogeneous type-aware fleets in terms of travel distance, speeds, and density, we infer road properties that cannot be obtained by type-agnostic models.

(ii) Coverage Limitation of Homogeneous Fleets. Surprisingly, more vehicles do not lead to higher road coverage due to ceiling point effects, as shown in Figure 7. Increasing a homogeneous fleet size improves its road coverage until the coverage reaches its ceiling point, which is inherently determined by mobility patterns of a fleet. However, the mobility patterns of heterogeneous type-aware fleets are usually complementary to each other. We can increase the ceiling point of road coverage of individual fleets by integrating heterogeneous fleets, which potentially improve the quality of the inferred map. On the other hand, it indicates a single-fleet solution cannot infer a map with high spatial coverage.

(iii) Impacts of Classifiers, Fleet Types, Features. We explored a few of classifiers and features and identified SVM and vehicle speeds are the best method and feature for the road category inference, respectively, as in Table 2 and Figure 24. We also found that personal vehicle, truck, and bus fleets alone achieved similar performances but

was worse than the taxi fleet as in Figure 23. These insights have the potential to guide future real-world map inference based on vehicular systems.

Impacts on Real-world Maps: We detected and reported 98.7km of missing roads to an existing map service provider in Shenzhen, which have been included in the newer version of the map. Given the cumulative GPS data and continuous road construction in Shenzhen, we believe more roads can be automatically detected or updated based on our coMap.

Vehicular Data Requirement: A major limitation of *coMap* is to require historical GPS records from heterogeneous type-aware fleets of diverse types. We believe this may not be difficult to achieve in the real-world setting for the following reasons. Most map service providers have access to GPS records for a large number of volunteer users, who report vehicle types through the map applications in either implicit or explicit ways. The recent efforts on connected vehicles [34] [11] [35], together with the more stable wireless connection [16], make GPS record collection and map update more efficient. We envision the data accessibility will not be a major obstacle for the implementation and application of *coMap* in the future.

More Road Properties: In this work, we only infer one road feature, i.e., road categories, in our case study. However, we believe that with similar approaches on road categories inference, we can infer other road properties based on the diversity of heterogeneous fleets, e.g., the road speed limit for different vehicle types, the shape of intersections, the number of lanes, and the number of branches on the primary roads.

Generality: Even with the above limitations, the design philosophy of *coMap* (i.e., utilizing the diversity among vehicular fleets and complementary patterns of their mobility to improve the performance on urban services) can be generalized to other scenarios. For instance, we can customize a navigation service for a specific vehicular fleet with their road category preferences. Please note that our *coMap* is not designed to replace existing commercial digital maps. Instead, we envision our coMap functions as a center for heterogeneous data systems and can create beneficial cycles among *map service providers*, *city infrastructures*, and *end client users*. The implementation and evaluation of *coMap* on different cities (e.g., Shenzhen and Hefei), different areas of the same city (e.g., complex road areas, downtown, and suburban areas), various combination of vehicle types, indicates the design philosophy of *coMap* with *CMF* has the potential to be generalized to other cities and scenarios.

Privacy Protections and Data Consent: While map inference has great social and commercial benefits, e.g., to reduce human efforts, it relies on GPS traces for map creation and maintenance. Therefore, we discuss privacy protection in our study. (i) First, we never expose any personal information in the application, e.g., the inferred map, our research output, is independent from personal traces. (ii) Second, when the data is collected by the service providers, all the data have been anonymized or hashed by the service providers when the data was sent to the server. (iii) Third, even given the full access of heterogeneous type-aware fleet data, we only process and analyze data directly related to our project by generating new intermediate data, and we drop all other information for minimal exposure of sensitive information, and such privacy protection can be improved in the future [25]. The commercial vehicle drivers are employees of the transportation agency we are working with, so their consents were obtained. For private vehicle drivers, their consents were obtained when they sign up for their usage-based insurance.

8 RELATED WORK

We divide the existing works into 4 categories based on two features, i.e., Single/Multiple Fleets and Independent /Dependent (i.e., using external information or not) in Table 3.

Independent Map Inference: For independent map inference, researchers use GPS data without external knowledge. Given the relative ease of access to single vehicular fleets by individual researchers, most works have

Table 3. Survey on Map Inference

Categories	Heterogeneous Fleets	
	Single	Multiple
Independent	[10] [39] [14] [19] [28]	[43] [29] [44]
Dependent	[13] [38] [45]	*coMap*

been focused on map inference from GPS data in single vehicular fleets. For instance, Cao *et al.* combine map generation problems with route planning problems from GPS traces of non-specified vehicles [10]. Wang *et al.* build a map update system based on data collected by a small-scale navigation app and a large-scale taxi GPS trace dataset [39]. He *et al.* designed a two-stage algorithm for map inference in dense areas [28]. Chen *et al.* infer a road map with a privacy-preserving model based on GPS data in two cities. Different vehicular fleets show diverse mobility patterns [43]. Some work has been using multiple fleets to infer traffic speeds, which can be used to infer road segment features [44]. However, independent map matching does not consider external information (e.g., vehicle types) other than physical GPS data, which limits its performance in practice.

Biagioni *et al.* propose a hybrid pipeline using KDE with an adaptive thresholding scheme to obtain an initial road network graph, followed by geometry and topology refinement and map-matching-based pruning to further improve accuracy [7]. Chen *et al.* propose a supervised learning framework that leverages prior knowledge on real-world road networks to learn the shape of different junctions, and integrate this with a cluster-based algorithm [12]. Stanojevic *et al.* develop a novel model in which map construction is framed as a network alignment problem. The derived optimization problem is then mapped into a hybrid algorithm combining k-means clustering and graph spanners [38]. Zheng *et al.* propose a revisited version of the trace merging method, applying a novel clustering algorithm that uses a partial curve matching method based on Frechet distance to measure the partial similarity between any trajectory and a previously created link [46]. While these methods improve on earlier schemes, none utilize the long-term connectivity between observations in the same trajectory when constructing the road network graph.

Dependent Map Inference: For dependent map inference, a map inference system uses both internal information (i.e., GPS data) and external information (e.g., commercial maps) for high-quality results. The inferred map from GPS traces is calibrated by the external data source to improve the map quality and inference efficiency. The most popular approaches are still based on single vehicular fleets. For example, Chen *et al.* design a framework to create up-to-date maps with rich knowledge from GPS trajectory collections and the existing maps [13]. Stanojevic *et al.* fuse the road networks inferred from 400 vehicles with existing online road map [38]. Compared with previous works, *coMap* is the first work to infer road maps from multiple vehicular fleets and further improve map quality with external data sources, i.e., OpenStreetMap [27].

On the other hand, existing works can be categorized into several groups based on their techniques. Map sketching (road structure inference) has been widely studied in previous works, such as kernel density models [7] [12] [15], trajectory clustering methods [8] [37], and intersection linking [22] [32]. However, most existing works inferring the road structures are built upon single vehicular fleets [10] [32] [20] [26], e.g., taxi, which is restricted by the bias of fleet mobility patterns, or without differentiating vehicle types, which loses diversity of mobility patterns. We design a graph-based model utilizing CMF as contextual information in map sketching to identify road points, intersections, and links. Map painting (road property inference), to the best of our knowledge, has not been systematically investigated by previous map inference studies. Based on the difference of vehicle distribution on roads, we estimate road properties with CMF and a learning-based model.

9 CONCLUSION

In this paper, relying on GPS trajectories from heterogeneous type-aware fleets, we design, implement, and evaluate a map inference system called *coMap* to infer a complete map representation at city scale. *coMap* system has three key components: *map sketching*, *map painting*, and *map update*. We conduct a comprehensive evaluation of *coMap* based on real-world data and results show (i) for the map sketching, our work improves the performance by 15.9% compared with the state-of-the-art methods; (ii) for the map painting, our work achieves 74.58% average accuracy on road type classification; (iii) more importantly, we detect and report 98.2 km of missing roads in *OpenStreetMap* with our map update component. We will also release our vehicle data used in this paper for researchers interested in this direction to build further applications.

ACKNOWLEDGMENTS

The authors would like to thank anonymous reviewers for their valuable comments and suggestions. This work is partially supported by NSF 1849238, NSF 1932223, NSF 1952096, NSF 2003874.

REFERENCES

[1] [n.d.]. 'https://www.tomtommaps.com/'. [Online; Retrieved December 8, 2018].
[2] [n.d.]. 'https://map.baidu.com/'. [Online; Retrieved December 8, 2018].
[3] [n.d.]. Government of Indian, Ministry of Statistics Programme Implementation. 'http://www.mospi.gov.in/'. [Online; Retrieved December 14, 2018].
[4] [n.d.]. Ministry of Transport of China. (n.d.). Total length of public highways in China from 2007 to 2017 (in kilometers). In Statista - The Statistics Portal. 'https://www.statista.com/statistics/276050/total-length-of-chinas-freeways/'. [Online; Retrieved December 8, 2018].
[5] 2018. National Bureau of Statistics of China. 'http://www.stats.gov.cn/tjsj/ndsj/'. [Online; Retrieved December 14, 2018].
[6] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer, 1–4.
[7] James Biagioni and Jakob Eriksson. 2012. Map inference in the face of noise and disparity. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. ACM, 79–88.
[8] Rene Bruntrup, Stefan Edelkamp, Shahid Jabbar, and Bjorn Scholz. 2005. Incremental map generation with GPS traces. In *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*. IEEE, 574–579.
[9] Chu Cao, Zhidan Liu, Mo Li, Wenqiang Wang, and Zheng Qin. 2018. Walkway discovery from large scale crowdsensing. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 13–24.
[10] Lili Cao and John Krumm. 2009. From GPS traces to a routable road map. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 3–12.
[11] Bin Bin Chen and Mun Choon Chan. 2009. MobTorrent: A framework for mobile Internet access from vehicles. In *INFOCOM 2009, IEEE*. IEEE, 1404–1412.
[12] Chen Chen and Yinhang Cheng. 2008. Roads digital map generation with multi-track GPS data. In *Education Technology and Training, 2008. and 2008 International Workshop on Geoscience and Remote Sensing. ETT and GRS 2008. International Workshop on*, Vol. 1. IEEE, 508–511.
[13] Chen Chen, Cewu Lu, Qixing Huang, Qiang Yang, Dimitrios Gunopulos, and Leonidas J. Guibas. 2016. City-Scale Map Creation and Updating using GPS Collections. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1465–1474.
[14] Xi Chen, Xiaopei Wu, Xiang-Yang Li, Yuan He, and Yunhao Liu. 2014. Privacy-preserving high-quality map generation with participatory sensing. In *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2310–2318.
[15] Jonathan J Davies, Alastair R Beresford, and Andy Hopper. 2006. Scalable, distributed, real-time map generation. *IEEE Pervasive Computing* 5, 4 (2006), 47–54.
[16] Pralhad Deshpande, Anand Kashyap, Chul Sung, and Samir R Das. 2009. Predictive methods for improved vehicular WiFi access. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*. ACM, 263–276.
[17] Ole Henry Dørum. 2017. Deriving double-digitized road network geometry from probe data. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 15.
[18] Stefan Edelkamp and Stefan Schrödl. 2003. Route planning and map inference with global positioning traces. In *Computer science in perspective*. Springer, 128–151.

[19] Stefan Edelkamp and Stefan Schrödl. 2003. Route planning and map inference with global positioning traces. *Lecture Notes in Computer Science* (2003), 128–151.

[20] Alexandros Efentakis, Sotiris Brakatsoulas, Nikos Grivas, Giorgos Lamprianidis, Kostas Patroumpas, and Dieter Pfoser. 2013. Towards a flexible and scalable fleet management service. In *Proceedings of the Sixth ACM SIGSPATIAL International Workshop on Computational Transportation Science*. ACM, 79.

[21] Jakob Eriksson. [n.d.]. INFERRING ROAD MAPS FROM GPS TRACES: SURVEY AND COMPARATIVE EVALUATION 2 James Biagioni* 3 Ph. D. Student 4 Department of Computer Science 5. ([n. d.]).

[22] Alireza Fathi and John Krumm. 2010. Detecting road intersections from GPS traces. In *International Conference on Geographic Information Science*. Springer, 56–69.

[23] Shilpa Garg, Pushpendra Singh, Parameswaran Ramanathan, and Rijurekha Sen. 2014. VividhaVahana: smartphone based vehicle classification and its applications in developing region. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ICST (Institute for Computer Sciences, Social-Informatics and ..., 364–373.

[24] Chong Yang Goh, Justin Dauwels, Nikola Mitrovic, Muhammad Tayyab Asif, Ali Oran, and Patrick Jaillet. 2012. Online map-matching based on hidden markov model for real-time traffic sensing applications. In *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*. IEEE, 776–781.

[25] Michaela Götz, Suman Nath, and Johannes Gehrke. 2012. Maskit: Privately releasing user context streams for personalized mobile applications. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 289–300.

[26] Tao Guo, Kazuaki Iwamura, and Masashi Koga. 2007. Towards high accuracy road maps generation from massive GPS traces data. In *Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007. IEEE International*. IEEE, 667–670.

[27] Mordechai (Muki) Haklay and Patrick Weber. 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing* 7, 4 (2008), 12–18.

[28] Songtao He, Favyen Bastani, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, and Sam Madden. 2018. RoadRunner: improving the precision of road network inference from GPS trajectories. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 3–12.

[29] Christian Heipke. 2010. Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing* 65, 6 (2010), 550–557.

[30] Shaohan Hu, Lu Su, Shen Li, Shiguang Wang, Chenji Pan, Siyu Gu, Md Tanvir Al Amin, Hengchang Liu, Suman Nath, Romit Roy Choudhury, et al. 2015. Experiences with eNav: A low-power vehicular navigation system. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 433–444.

[31] Avdhut Joshi and Michael R James. 2015. Generation of accurate lane-level maps from coarse prior maps and lidar. *IEEE Intelligent Transportation Systems Magazine* 7, 1 (2015), 19–29.

[32] Sophia Karagiorgou and Dieter Pfoser. 2012. On vehicle tracking data-based road network generation. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. ACM, 89–98.

[33] Xuemei Liu, James Biagioni, Jakob Eriksson, Yin Wang, George Forman, and Yanmin Zhu. 2012. Mining large-scale, sparse GPS traces for map inference: comparison of approaches. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 669–677.

[34] Pin Lv, Xudong Wang, Xiuhui Xue, and Ming Xu. 2015. SWIMMING: seamless and efficient WiFi-based internet access from moving vehicles. *IEEE Transactions on Mobile Computing* 14, 5 (2015), 1085–1097.

[35] Ratul Mahajan and Aruna Balasubramanian. 2013. Interactive WiFi connectivity for moving vehicles. US Patent 8,457,546.

[36] Paul Newson and John Krumm. 2009. Hidden Markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*. ACM, 336–343.

[37] Stefan Schroedl, Kiri Wagstaff, Seth Rogers, Pat Langley, and Christopher Wilson. 2004. Mining GPS traces for map refinement. *Data mining and knowledge Discovery* 9, 1 (2004), 59–87.

[38] Rade Stanojevic, Sofiane Abbar, Saravanan Thirumuruganathan, Gianmarco De Francisci Morales, Sanjay Chawla, Fethi Filali, and Ahid Aleimat. 2018. Road Network Fusion for Incremental Map Updates. *arXiv preprint arXiv:1802.02351* (2018), 91–109.

[39] Yin Wang, Xuemei Liu, Hong Wei, George Forman, Chao Chen, and Yanmin Zhu. 2013. CrowdAtlas: self-updating maps for cloud and personal use. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. 469–470.

[40] Yin Wang, Hong Wei, and George Forman. 2013. Mining large-scale gps streams for connectivity refinement of road maps. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 448–451.

[41] OpenStreetMap Wiki. 2018. Key:highway — OpenStreetMap Wiki,. http://wiki.openstreetmap.org/w/index.php?title=Key:highway&oldid=1687287 [Online; accessed 12-December-2018].

[42] Stewart Worrall and Eduardo Nebot. 2007. Automated process for generating digitised maps through GPS data compression. In *Australasian Conference on Robotics and Automation*, Vol. 6. Brisbane: ACRA.

[43] Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, and Tian He. 2014. Exploring human mobility with multi-source data at extremely large metropolitan scales. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 201–212.

[44] Desheng Zhang, Juanjuan Zhao, Fan Zhang, and Tian He. 2015. UrbanCPS: A Cyber-physical System Based on Multi-source Big Infrastructure Data for Heterogeneous Model Integration. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems* (Seattle, Washington) *(ICCPS '15)*. ACM, New York, NY, USA, 238–247. https://doi.org/10.1145/2735960.2735985

[45] Lijuan Zhang, Frank Thiemann, and Monika Sester. 2010. Integration of GPS traces with road map. In *Proceedings of the Third International Workshop on Computational Transportation Science*. ACM, 17–22.

[46] Yu Zheng, Tong Liu, Yilun Wang, Yanmin Zhu, Yanchi Liu, and Eric Chang. 2014. Diagnosing New York city's noises with ubiquitous data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 715–725.