



Delft University of Technology

Map-matching for cycling travel data in urban area

Gao, Ting; Daamen, Winnie; Krishnakumari, Panchamy; Hoogendoorn, Serge

DOI

[10.1049/itr2.12567](https://doi.org/10.1049/itr2.12567)

Publication date

2024

Document Version

Final published version

Published in

IET Intelligent Transport Systems

Citation (APA)

Gao, T., Daamen, W., Krishnakumari, P., & Hoogendoorn, S. (2024). Map-matching for cycling travel data in urban area. *IET Intelligent Transport Systems*, 18(11), 2178-2203. <https://doi.org/10.1049/itr2.12567>

Important note

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

Copyright


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

Takedown policy

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

ORIGINAL RESEARCH

Map-matching for cycling travel data in urban area

Ting Gao  | Winnie Daamen | Panchamy Krishnakumari | Serge Hoogendoorn

Department of Transport & Planning, Delft
University of Technology, Delft, The Netherlands

Correspondence

Ting Gao, Department of Transport, and Planning,
Delft University of Technology, 4.21 Stevinweg 1,
2628 CN Delft, The Netherlands.
Email: t.gao-1@tudelft.nl

Funding information

EUROPEAN COMMISSION - Directorate-
General for Communications Networks, Content
and Technology under Horizon Europe research
and innovation programme, Grant/Award Number:
101093051

Abstract

To promote urban sustainability, many cities are adopting bicycle-friendly policies, leveraging GPS trajectories as a vital data source. However, the inherent errors in GPS data necessitate a critical preprocessing step known as map-matching. Due to GPS device malfunction, road network ambiguity for cyclists, and inaccuracies in publicly accessible streetmaps, existing map-matching methods face challenges in accurately selecting the best-mapped route. In urban settings, these challenges are exacerbated by high buildings, which tend to attenuate GPS accuracy, and by the increased complexity of the road network. To resolve this issue, this work introduces a map-matching method tailored for cycling travel data in urban areas. The approach introduces two main innovations: a reliable classification of road availability for cyclists, with a particular focus on the main road network, and an extended multi-objective map-matching scoring system. This system integrates penalty, geometric, topology, and temporal scores to optimize the selection of mapped road segments, collectively forming a complete route. Rotterdam, the second-largest city in the Netherlands, is selected as the case study city, and real-world data is used for method implementation and evaluation. Hundred trajectories were manually labelled to assess the model performance and its sensitivity to parameter settings, GPS sampling interval, and travel time. The method is able to unveil variations in cyclist travel behavior, providing municipalities with insights to optimize cycling infrastructure and improve traffic management, such as by identifying high-traffic areas for targeted infrastructure upgrades and optimizing traffic light settings based on cyclist waiting times.

1 | INTRODUCTION

Bicycles are becoming increasingly popular among people as a sustainable and convenient way of transportation, particularly in countries such as the Netherlands, Sweden, and Denmark [1]. Besides serving as an eco-friendly alternative for short to medium-distance car journeys, bicycles provide citizens with faster first or last-mile connectivity to public transport compared to walking.

As the country with the highest number of bicycles per capita, the Netherlands maintains its dedication to evidence-based traffic policy-making, leveraging extensive data collection. Notably, between 2020 and 2022, the Dutch government launched the Talking Bike Program to collect bicycle GPS trajectory data throughout the entire country [2, 3]. Urban planners invest efforts in understanding bicycle traffic patterns not only to

enhance cycling infrastructures but also to improve traffic performance from a multi-modal perspective [4, 5]. Since bicycles serve as a vital link in connecting public transport systems and significantly impact road traffic conditions, maintaining a multi-modal view is essential when processing bicycle data. Regarding map-matching, it is unreasonable to exclusively consider official cycleways, as this approach overlooks potential interactions with other traffic modes.

The increasing availability of GPS data has underscored the importance of map-matching as an integral process for adapting such data to diverse traffic applications. Map-matching is pivotal as it aligns GPS data points with specific road segments, necessitating a detailed map with comprehensive information about the road network. In our study, we choose OpenStreetMap (OSM) as the mapping resource for analysis and implementation. OSM stands as the prominent map for many end-users,

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *IET Intelligent Transport Systems* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

including industrial actors and researchers, providing a collaborative, freely accessible, and extensive map of the world [6]. This platform encompasses an array of geographical information crucial for map-matching, encompassing features such as longitude, latitude, and road types. We opt for BBBike¹ for OSM data extraction. BBBike is a widely used tool in mapping and cycling-related projects, providing high-quality and up-to-date OSM data.

The quality of map-matching is significantly influenced by multiple factors. GPS drift and building disturbances, often resulting in deviations in mapped roads, introduce uncertainty in accurately determining traveled paths. Lower sampling rates necessitate the inference of intermediate roads, further increasing uncertainties about mapped road precision. OSM quality also significantly impacts map-matching results: inaccurately labeled roads or erroneous geographical coordinates can significantly affect the accuracy of mapped roads and thus the resulting route. These problems become more pronounced in urban areas, as high buildings can aggravate GPS errors [7], and the dense road network presents more possible roads to map the GPS coordinates as well as origin-destination route choices.

While numerous algorithms cater to map-matching car GPS data, as extensively discussed in [8–10], there are notably fewer algorithms designed for bicycles. Specifically, it is possible to directly apply car-focused map-matching methods to bicycles, but their performance is suboptimal [11]. Further improvements are needed due to the dissimilarities between car and bicycle travel patterns: cars adhere to specific speed limits and road restrictions, whereas there are few speed controls for cycling, and cyclists possess the flexibility to traverse various routes. Existing bicycle-specific algorithms, such as those found in [12, 13], define the bicycle network based solely on the information indicating whether a road is accessible for bicycles in OSM. However, this approach may be insufficient if OSM is inaccurate or if other infrastructure more in favour of the cyclists is nearby.

In this study, we introduce an innovative map-matching algorithm tailored for bicycles. Our main contributions are as follows: (1) Improved road availability classification for bicycles: We meticulously assess road availability for cyclists based on route choice studies, OSM street descriptions, and real-world observations, with particular emphasis on the main car road network. (2) Extended trajectory matching method: Based on the refined network, we have developed an extended map-matching method that improves upon existing approaches by penalizing unrealistic speed and expanding the notion of travel cost along different road segments. (3) Real-world implementation: We conducted experiments using real-world data from Rotterdam. We manually labeled 100 trajectories to evaluate model performance, demonstrating a significant improvement compared to the baseline model. Regarding sensibility, our method also proved to be robust against parameter changes and GPS data quality. Additionally, we analyzed user behavior based on the mapped results. (4) Accessible code: Our contribution

includes codes featuring functions for parallel processing on laptops or supercomputers and visualization tools to display the map-matching results, the code implementation is accessible on GitLab².

The following sections are organized as follows: Section 2 provides a synthetic review of current map-matching methods and outlines the unsolved challenges. In Section 3, a detailed introduction to our method is presented. Section 4 applies our method to real-world data and evaluates the results. Finally, Section 5 draws conclusions and opens the discussion for future research.

2 | BACKGROUND

Map-matching methods have been extensively studied in recent years and are broadly categorized into three types: geometric, topological, and advanced algorithms [14]. Geometric methods match GPS records to the nearest network elements (nodes or links) based on distance, and their accuracy heavily relies on the precision of GPS data. In contrast, topological methods analyze sequences of GPS records and the network's connectivity. Advanced methods integrate both geometric and topological approaches, often using advanced techniques like Hidden Markov Models (HMMs) [8, 15].

The HMMs model the road segment to be mapped in the network as a state, with the state probability indicating the likelihood of observing the provided GPS record under the condition that the bicycle is on the corresponding road segment (state). The transition probability represents the likelihood of moving between different road segments (states). Designing these state and transition probabilities is crucial. The geometric method mentioned earlier can provide the geometric score as the state probability, while the topological method can offer the topological score as the transition probability. However, this remains insufficient, as the roads chosen by cyclists among different roads also depend on road types and the surrounding environment.

With the development of machine learning, data-driven methods have emerged as prominent options. A deep reinforcement learning framework for map-matching cellular data is proposed in [16]. In [17], data sparsity and noise are addressed through deep learning-based data augmentation. However, this method requires ground truth data, which are not always available for bicycles due to privacy concerns. Similarly, a method based on representation learning is proposed in [18], where high-frequency trajectories are needed to enhance the expressive capability of representations. Transfer learning is implemented in [19], where labelled data are still needed and generated data are based on a feasible network. While these methods have proven efficient for car traffic, they face limitations when applied to non-labelled bicycle data. Data generation is also challenging because cyclists have a wide range of choices over road types, and rule violations are common [20], making it hard to define a feasible network.

¹ <https://download.bbbike.org/osm/bbbike/Rotterdam/>

² <https://gitlab.tudelft.nl/T.Gao-1/mm4b.git>

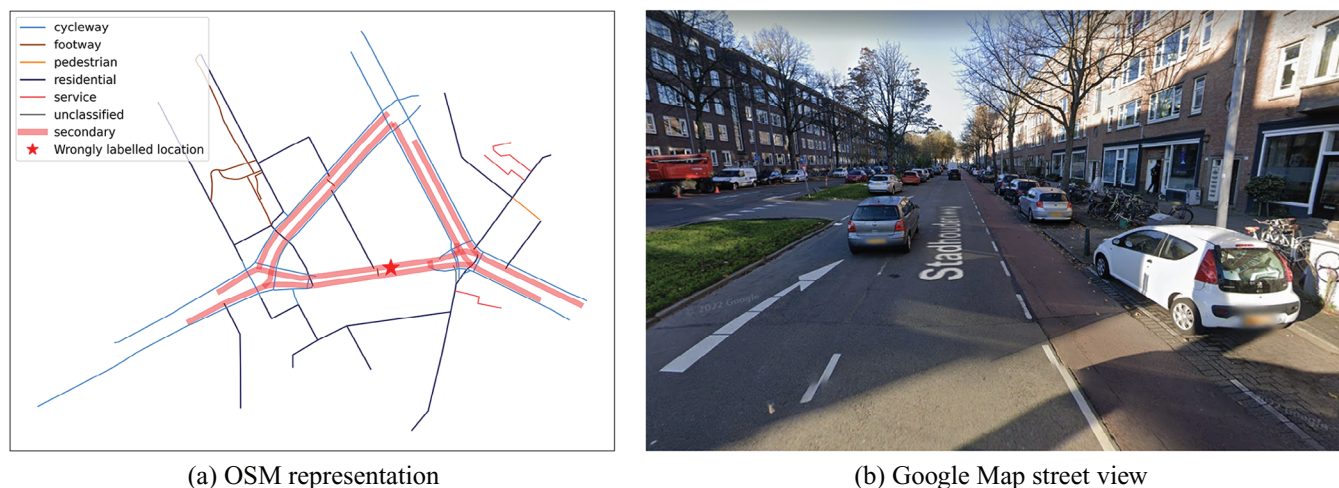


FIGURE 1 At the location marked by the star (latitude = 51.93156, longitude = 4.465156) on Stadhoudersweg street, 3039CD Rotterdam, (a) OpenStreetMap (OSM) displays only secondary roads, despite (b) the presence of nearby cycleways.

To the best of our knowledge, since 2020, no publications specifically designed for bicycle map-matching include open-source code and do not require ground truth training data. A study detailed in [11] compared six prominent advanced map-matching algorithms [8, 12, 21–24] using active travel data, comprising 88% bike/e-bike data and 12% walk/run data. The evaluation matrices included six ground-truth dependent and six ground-truth independent measures. Based on the findings in [11], pgMapMatch [8] adopted the HMM and emerged as the top-performing method. Therefore, we take this method as a starting point for our research. However, since pgMapMatch is primarily designed for car traffic and ignores the real-world scenarios for bicycles and utilizes speed limitations which are much less strict for cyclists, we need to incorporate additional functionalities to tailor it to bicycle traffic.

When dealing with real-world bicycle data, the necessity for a more precise bicycle network is crucial due to the limitations of GPS data quality. This limitation can impede map-matching algorithms from consistently selecting the most suitable road types for cyclists. The precision of the bicycle network is pivotal for deriving various traffic variables, such as traffic flow, which relies on precise matching on the respective roads. Publications specifically tailored for bicycles take the bicycle network into consideration. In [12], bicycle travel data are only mapped to roads reserved for cycling. Meanwhile, [13] derives the bicycle network from OpenStreetMap, excluding road types like motorways, footways, steps, and paths unless an additional tag specifies permission for cycling. While these methods do distinguish between networks for cars and bicycles, they overlook the challenges encountered in real-life scenarios, as outlined below.

One of the primary challenges stems from the ambiguity of the road network for cyclists. In many countries, cyclists share roads with either pedestrians (on sidewalks) or cars (on the road). In the Netherlands and Denmark, even though cyclists have designated protected cycling routes, they sometimes navigate through roads shared with other modes of transportation

[25]. This mixing of cycling routes with other road types makes it more challenging to determine the route a cyclist has taken, as more options exist.

However, since the goal of map-matching is to replicate roads taken in real-life as closely as possible, the approach of simply removing road types originally designed for other modalities, as suggested by [12, 13], is not a practical solution. Moreover, from the perspective of multi-modal transport, these roads may host interactions among various transportation modes and cyclists may frequently interact with pedestrians or other vehicles on these shared roads.

The second challenge stems from the inherent inaccuracy of OSM data. The extraction of precise geographic information is a complex task, and OSM may not always provide reliable representations of real-world conditions [26]. For example, Figure 1 illustrates a case where OSM data exhibits inconsistencies with reality. This observation underscores the importance of not immediately removing roads from the dataset, as some roads in OSM are labeled solely as car roads (e.g. secondary roads) but are also utilized as bicycle routes. Erroneous labeling presents specific challenges for bicycle map-matching due to the diverse types of roads available in OpenStreetMap. Complications also arise from varying regulations regarding infrastructure usage by bicycles across different countries.

In summary, while numerous map-matching methods have been designed for car traffic, there is still a gap in addressing the specific challenges posed by bicycle traffic in the real world: (1) The type of roads and surrounding environment influence cyclists' route choices, which geometric and topological methods fail to account for. (2) Most algorithms are designed for car traffic, whereas bicycle traffic has its own specificities—fewer speed limitations, a wider variety of usable roads—resulting in an ambiguous road network for cyclists. (3) Finally, the OSM dataset is noisy, containing roads usable by cyclists but not labeled as such. These challenges persist and have motivated our research, forming the foundation of our approach.

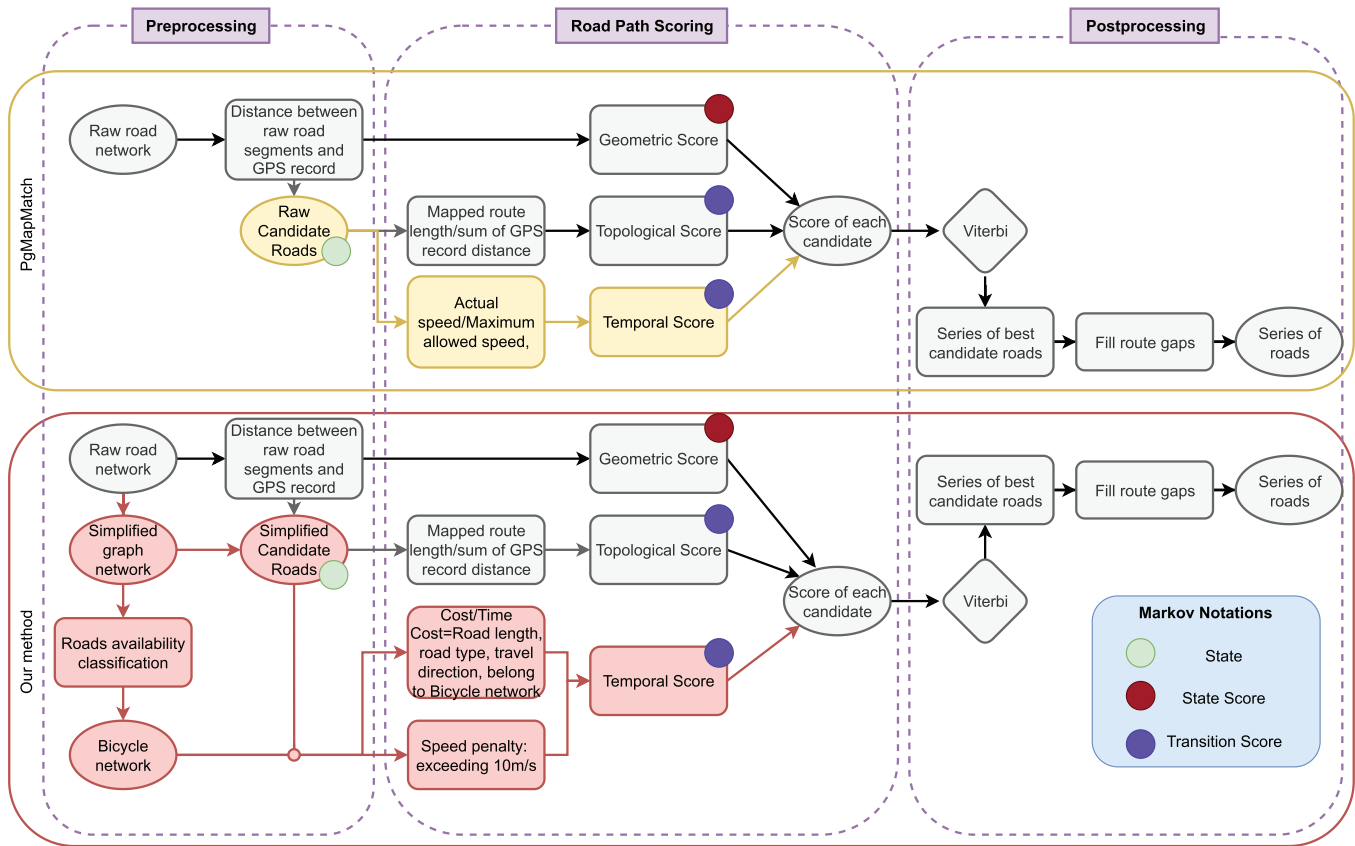


FIGURE 2 Comparison of matching GPS travel data for one route between the pgMapMatch method and our method. Each method is divided into three stages: preprocessing, road path scoring, and postprocessing. The common parts are denoted in grey, while the differences are highlighted in yellow and red. Actions are represented by squares, and input/output are denoted by circles.

3 | NOVEL MAP-MATCHING METHOD FOR BICYCLES

In this section, we introduce our innovative approach to map-matching cycling GPS data. We begin with an overview in Section 3.1 that underscores the novelty of our method which extends the pgMapMatch model to cater specifically to bicycle traffic. Subsequently, we delve into two crucial preprocesses. The first step, outlined in Section 3.2, focuses on creating a simplified graph representation of the road network. The second step, detailed in Section 3.3, automates the classification of road availabilities for cyclists, particularly focusing on the urban main road network to ensure accurate mapping of trajectories to the most probable roads. Finally, we elaborate on our state scoring system and extended transition scoring system in Section 3.4, where each road candidate is assigned a score that integrates geometric, topological, and temporal aspects.

3.1 | Extended map-matching method overview

The pgMapMatch method [8] has been recognized as the optimal map-matching model for car traffic [11]. We extend this method to better suit bicycle traffic. An overview of these two

methods is presented in Figure 2. As depicted in Figure 2, each method comprises three stages: preprocessing, road path scoring, and postprocessing. The entire process falls under the category of HMM, with each stage corresponding to defining the state, determining state score and transition score, and finding the optimal path.

3.1.1 | Preprocessing stage

In this stage, the bicycle network is identified, and a set of candidate roads for each GPS data point is selected. These roads represent the states in Markov theories.

In pgMapMatch, the raw road network is used as the bicycle network, and roads within a 50-meter radius for each GPS data record are chosen as candidates. The 50-meter radius is generally larger than GPS accuracy, ensuring the capture of almost all true road segments. In our method, we start by extracting a simplified graph network from the raw road network to reduce the computational load. This process will be detailed in the following Section 3.2. Based on the simplified network, we first classify the availability of various roads based on the OSM road type descriptions and discrepancies between OSM road types and real-world observations. Subsequently, we refine our classification for the main road network, which serves as the backbone

of the urban traffic network. This refinement is driven by the observation that cyclists often avoid main car roads when well-established parallel cycleways are available nearby, prioritizing safety and comfort. A detailed description of this process is in Section 3.3.

It is worth noting that, in our method, the simplified graph is built by merging redundant nodes while maintaining the geometry shape. Therefore, when determining the distance between a GPS record and a specific simplified road (edge), we use the shortest distance of all original roads (edges) that collectively form the given road (edge).

3.1.2 | Path scoring stage

In this stage, each candidate road from the simplified network receives a state score to evaluate its likelihood of being part of the route, and each pair of candidate roads from adjacent GPS records is assigned a transition score to assess the connection possibility. In our approach tailored for bicycle traffic, we have implemented several crucial adaptations.

Regarding the state score, we adopt the same method as pgMapMatch, only considering the geometric component, which is the distance between GPS data points and road segments.

As to the transition score, the pgMapMatch method considers two aspects: (1) a topological component based on the length of mapped road segments and the distance between adjacent GPS points, favoring shorter routes in the geographical context, and (2) a temporal component favoring smaller speeds and penalizing significantly for speeds exceeding the maximum allowed speed. However, the latter may not correspond effectively to bicycle traffic, as road speed limits might not be as relevant, and there are few control devices (such as speed cameras) for bicycles. In our method, in addition to the topological component, we use a temporal speed score to penalize unfeasible speeds for all trips, especially those exceeding 10 m/s. Such velocities are often unusual in urban cycling environments and are likely the result of GPS drift, where the recorded location deviates from the actual position. Additionally, we extend the temporal component to a time-averaged cost, where the cost combines factors such as road length, the direction of travel (whether or not to go against the flow of traffic), and the pre-defined road availability class. The score of each component is detailed in Section 3.4.

3.1.3 | Postprocessing stage

In this stage, we employ a strategy similar to the pgMapMatch method. By leveraging the pre-defined state score and transition score, we apply the Viterbi algorithm [27] to maximize the accumulated score along the trajectory. This process takes into account conditions such as U-turns and maximum skips. The maximum skip parameter enables the algorithm to skip over a specified number of consecutive GPS points with an associated skip cost. In practical term, if a current GPS point fails to find a suitable match within the defined maximum skip distance on the

road network, the algorithm proceeds to the subsequent point in the sequence in an attempt to find a match. The best candidate route is then selected, and any gaps in the route are filled.

3.2 | Graph extraction

To enhance the clarity of our presentation, we represent the road network with a graph. The graph representing the whole network is denoted as $G_f(V, E)$, where V represents the set of nodes, and E the set of edges. In this context, $e_i = (u, v) \in E$ signifies the i -th road segment from source node u to target node v . The notations used in this article can be found in Table A1 of Appendix A.

In OSM, each road consists of several straight segments to approximate the road shape. Since the original network is large and computationally intensive when mapping trajectories to road segments, we aim to create a more compact graph. We acknowledge that packages such as networkx³ and osmnx⁴ provide functions to simplify graphs. Nonetheless, we decided to develop our own package in Algorithm B1, Appendix B to have the flexibility for specific map-matching demands such as maintaining oneway and road type information. The length of the new edge is the sum of its merged edges. In Figure 3, we present the graph representation of a portion of OSM and its simplified result.

3.3 | Refined road classification method for cyclist availability

Achieving accurate map-matching to desired road segments poses significant challenges due to the intricate nature of roads and inaccuracies in OSM data describing these roads. For cyclists' trajectory mapping, it is crucial to evaluate the availability and safety of different roads, as some are easily available and safe for cyclists, while others may pose significant dangers or are unavailable for cyclists. It is insufficient to classify road segments based solely on road types due to the discrepancies between the road network and OSM data: roads might be missing or wrongly labelled in OSM data. In this section, we pay special attention to these discrepancies for urban main roads and the nearby roads available to cyclists. Indeed, they serve as critical traffic arteries and analyzing their accurate traffic state is essential for effective traffic management.

In the following, we start by discussing common preferences regarding cycling infrastructure. Then, we use Rotterdam as a case study to identify observed inaccuracies within OSM using our visual inspection tool for road types. Based on these observations, we propose a pipeline classifying main car roads.

Understanding cyclist preferences and behaviors is crucial for accurate map-matching, especially in cases where ground truth data is missing. Although cycling behavior varies among

³ <https://gitlab.tudelft.nl/T.Gao-1/mm4b.githttps://networkx.org/documentation/stable/>

⁴ <https://osmnx.readthedocs.io/en/stable/>

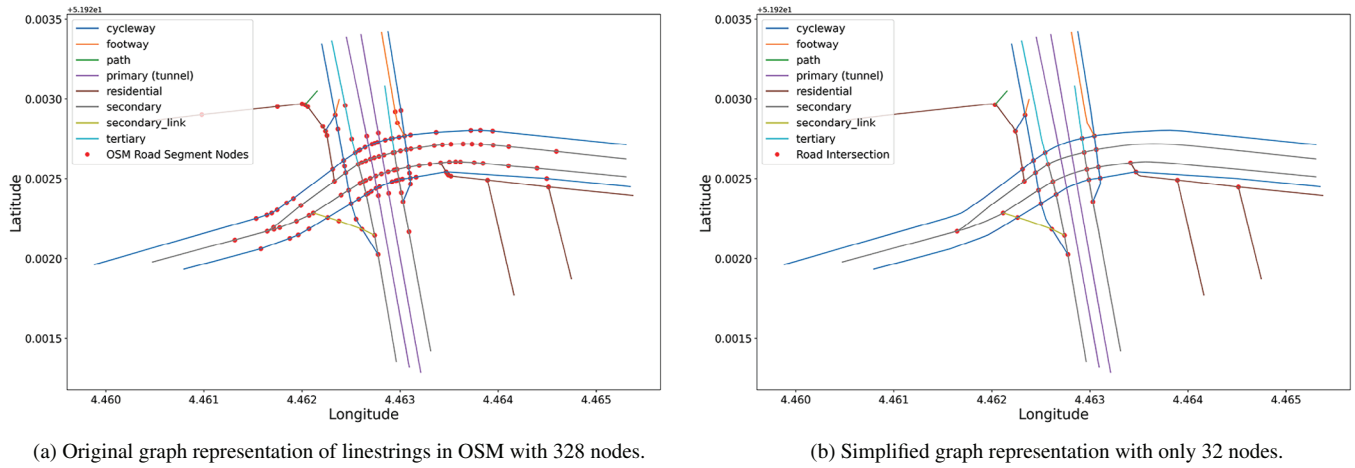


FIGURE 3 Visualization of road segments in OpenStreetMap, with different colors corresponding to different types of roads. In this scenario, primary roads do not intersect with other roads because they are tunnels.

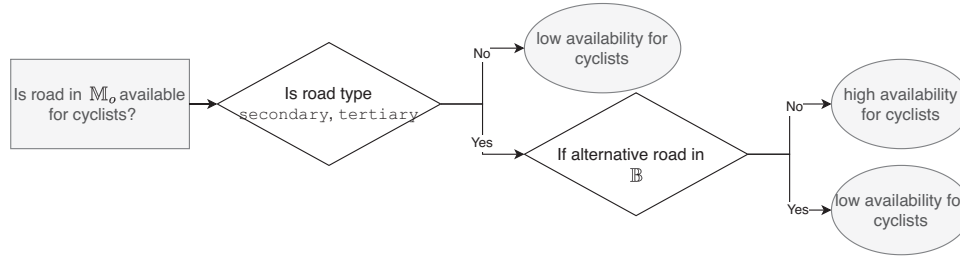


FIGURE 4 Pipeline of analyzing the availability of main car roads for cyclists. $M_0 = \{\text{primary, primary link, secondary, secondary link, tertiary}\}$, $B = \{\text{cycleway, footway, residential}\}$.

individuals, some common preferences is consistently observed across different cyclists:

In scenarios where dedicated cycle paths run alongside main roads, cyclists prioritize segregated cycling infrastructure to enhance safety and comfort [29–32]. Cycling on main carriageways is generally prohibited and considered highly hazardous [33]. Additionally, cyclists' route choices are strongly influenced by the trip length [31, 34, 35].

Cyclists' route choices reflect a balance between safety, comfort, and trip length. Based on Google satellite images in Rotterdam urban area, we observe that dedicated cycleways are always provided parallel to main car roads. In these cases, the difference in trip length between using cycleways and car roads is minimal, and the high risk associated with car roads create a strong preference for cycleways. For pedestrian-dedicated roads, the choice is influenced more by comfort and trip length than by safety considerations. Cyclists may use shortcuts through these roads when they offer significant time savings, although these routes are generally less preferred from a comfort perspective.

In OSM, we denote the hierarchized main car road network as M_0 , consisting of **motorway**, **motorway link**,

primary, **primary link**, **secondary**, **secondary link**, **tertiary** roads. By visualizing different OSM road types in the map, we derive the following observations:

Observation 1 Parallel cycleways alongside **motorway**, **motorway link**, **primary**, **primary link**, **secondary link** are consistently depicted in OSM.

Observation 2 If cycleways run parallel to **secondary** and **tertiary** roads in OSM, they are labeled as **cycleway**, **footway**, or **residential**. Otherwise, they are merged with the main road representation. For simplification, we denote the set of **cycleway**, **footway**, and **residential** as Alternative Bikeable Road Set B .

The cycling infrastructure preferences show a significant priority difference between main car roads and adjacent cycleways. However, OSM inaccuracies often blur the distinction between parallel cycleways and main roads. Therefore, to correctly evaluate the probability of cyclists using different roads, it is necessary to distinguish between actual cycleways and main car roads in the OSM network. The workflow is depicted in Figure 4. For **secondary** and **tertiary** roads, if we can identify alternative roads within set B for cyclists, we confirm that they only represent car roads and are of low availability for cyclists.

Based on this examination on main car road network, we complete a road classification that accounts for OSM inaccuracies and street descriptions. This classification provides a more

TABLE 1 Classification of roads availability considering street descriptions in [28], real-world situations in the Rotterdam urban area, and revised main car road network. **secondary-2** and **tertiary-2** denote roads identified as having low availability, while **secondary-1** and **tertiary-1** represent the remaining roads.

OSM road type	Class	Explanation
Cycleway	0	Dedicated cycling infrastructure in urban areas.
Residential	1	Residential areas are frequently visited by our user group. Many real-life cycleways are wrongly labelled (overlapped) with residential in OSM.
Tertiary-1	1	The next most important roads in a country's system. Many real-life cycleways are wrongly labelled (overlapped) with tertiary in OSM.
Service	2	Roads leading to or within areas such as industrial estates, campsites, business parks, car parks, and alleys. They are easy to be used by cyclists.
Path	2	Generic paths intended for all non-motorized vehicles including bicycles.
Living street	2	Residential streets with generally lower speed limits and reduced traffic volume, creating a safer environment for cyclists and pedestrians.
Unclassified	3	Unclassified roads.
Footway	3	Roads mainly or exclusively designed for pedestrians, typically in park, zoom, and garden. They often feature elements such as uneven surfaces and narrow width, which are not ideal for cyclists to use.
Pedestrian	3	Roads mainly or exclusively for pedestrians, typically in train station and shopping areas. Cyclists are generally required to walk their bikes in train station and shopping areas.
Steps	3	Roads with stairs where cyclists must carry their bicycles.
Bridleway	3	Roads designed for horse riders, they are often unpaved or have rough surfaces, such as gravel or dirt, which can be challenging and uncomfortable for cyclists.
Platform	3	Platforms at bus stops or train stations, they are not designed to accommodate bicycles and may lack features for safe cycling.
Busway	3	Bus lanes next to the road. The infrastructure such as line width and traffic control is tailored for buses instead of bicycles.
Construction	3	Roads under construction. Construction zones often have uneven or unstable surfaces that are hard for cyclists to use.
Services	3	Service stations along highways. They are often located at highway interchanges with limited access for cyclists.
Secondary-1	3	secondary roads in OSM that overlap with real-life cycleways, this identification may be overly optimistic.
Motorway	4	Restricted access highways.
Motorway link	4	Link roads between motorways.
Secondary link	4	Link roads between secondary roads, typically depicted with parallel cycleways in OSM.
Primary	4	Major roads in a country's road system, typically depicted with parallel cycleways in OSM.
Primary link	4	Roads linking primary roads, typically depicted with parallel cycleways in OSM.
Tertiary-2	4	tertiary roads with parallel cycleways in OSM.
Secondary-2	4	secondary roads with parallel cycleways in OSM.

accurate assessment of the route choice's probability based on road type, comfort, and safety consideration for cyclists. As shown in Table 1, we group all the roads into five classes. Class 0 includes only **cycleway**, corresponding to dedicated cycling infrastructure along main roads. Class 1 comprises cycleways that are incorrectly labeled (e.g. **residential** and part of **tertiary** roads). Class 2 encompasses roads such as **service** that are not primarily designed for cycling but are still available to cyclists. Class 3 consists of roads that cyclists are generally not permitted to use but where violations occur occasionally (e.g. **pedestrian**) [20]. Class 4 includes roads used by cyclists only in extreme situations, where usage is highly dangerous (e.g. **primary**) [33].

It is worth noting that, although our observations and road classification table are based on Rotterdam, they are flexible and can be adapted to accommodate specific local infrastruc-

ture conditions and usage patterns. In the Section 3.3.1, we detail the process of finding alternative roads for **secondary** and **tertiary** roads in Figure 4, and address specific cases that necessitate precise handling of inaccuracies in Section 3.3.2.

3.3.1 | Framework of revising main car road availability

We classify an OSM main road as having low bicycle availability when alternative bikeable roads are nearby. An intuitive case is given in Figure 5. Figure 5a provides an illustration of the road network in a selected area in Rotterdam, the Netherlands, which showcases a typical intersection scenario. Figure 5b categorizes these roads based on our previous definition. Considering the expanded road in Figure 5c, it is easy to find available alternative

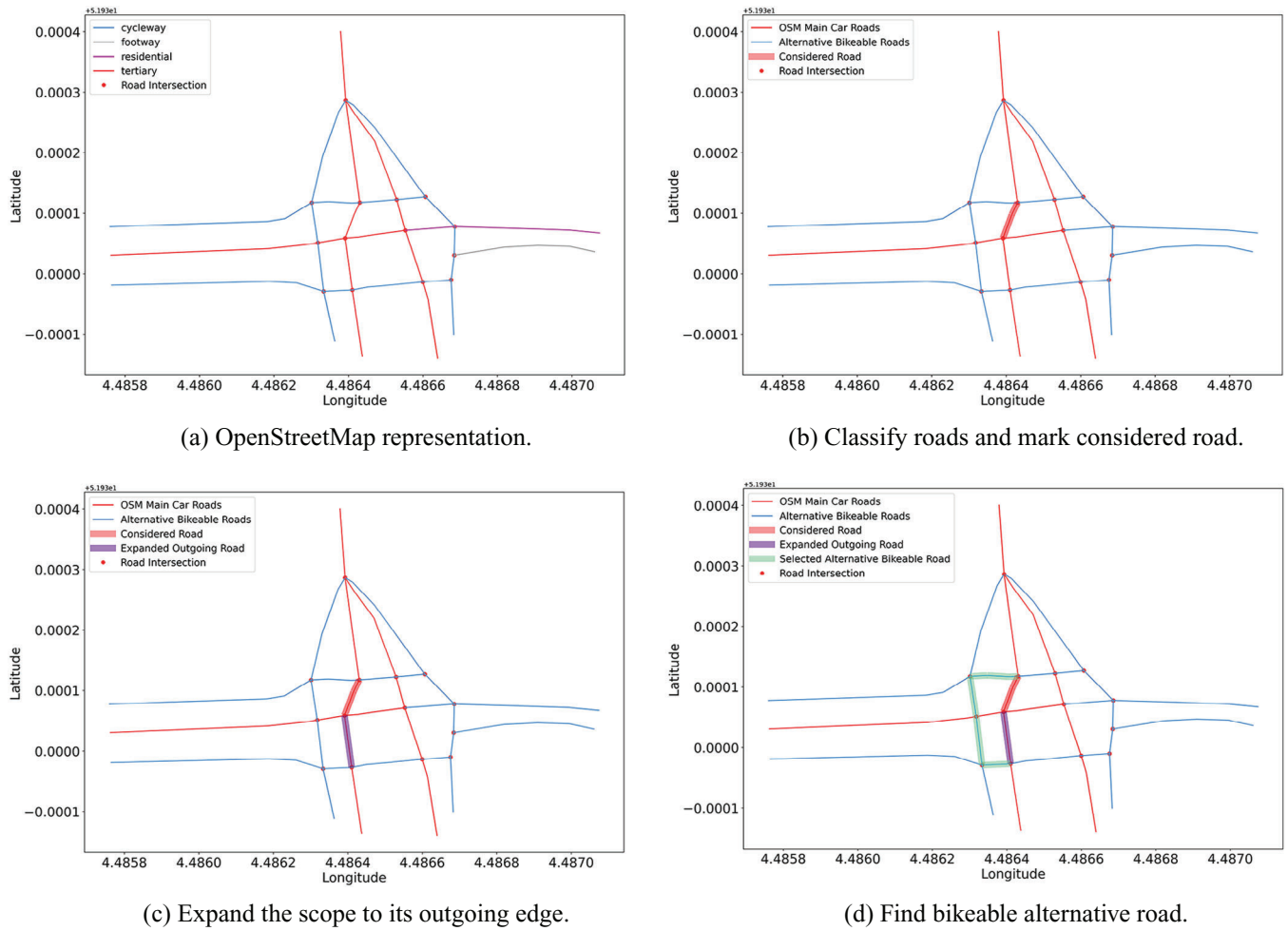


FIGURE 5 Among the **secondary** and **tertiary** roads we evaluate availability for cyclists. First classify all the roads into OSM main car roads and alternative bikeable roads. Later, we expand outgoing roads of the considered road to find practical alternative bikeable routes.

bikeable roads nearby in Figure 5d, making the considered road of low availability for cyclists.

There are two main observations from Figure 5: the first one is that the alternative roads and OSM main car roads do not necessarily geometrically run in parallel but intersect with each other, forming acute angles. Therefore, the process of revising OSM main car roads should be rooted in topological connectivity rather than geometric parallelism.

The second observation is that some OSM main car roads, while not directly connected to bikeable roads in the local view, can still be considered with low availability for cyclists. This is because a narrow local perspective may not suffice to detect potential nearby bikeable road alternatives. Thus, to expand our observation area and make a more informed determination, we consider both the incoming and outgoing edges within a specified distance. An example of expanding outgoing scope is given in Appendix C Algorithm C1.

With the expanded OSM main car roads, it is now possible to identify whether alternative roads sharing the same start node and end node exist. The condition for bikeable roads to serve as alternatives is based on two key metrics: detour ratio θ_{uv} and detour distance δ_{uv} . Denoting the bicycle distance as $B(u, v)$

(the shortest distance achievable via bikeable roads between nodes u and v) and OSM main car road distance as $\mathcal{U}(u, v)$ (the shortest distance achievable via OSM main car road network), the detour ratio θ_{uv} and detour distance δ_{uv} are expressed as follows:

$$\theta_{uv} = \frac{B(u, v)}{\mathcal{U}(u, v)},$$

$$\delta_{uv} = |\mathcal{U}(u, v) - B(u, v)|.$$

The use of the detour ratio θ_{uv} and detour distance δ_{uv} is depicted in Figure 6, where two different rules are presented, each applying to different scenarios. Satisfying any of them confirms the presence of alternative bikeable roads, rendering the existing OSM main car roads unavailable for cyclists: (1) The first rule applies to shortcut situations by only constraining the detour ratio. This is intuitive, as if the bikeable road is comparable or even shorter than the original road, cyclists tend to take the bikeable road, as shown in Figure 6a. (2) The second rule addresses short OSM main car roads where the detour ratio might be substantial. As shown in Figure 6b, we adopt a

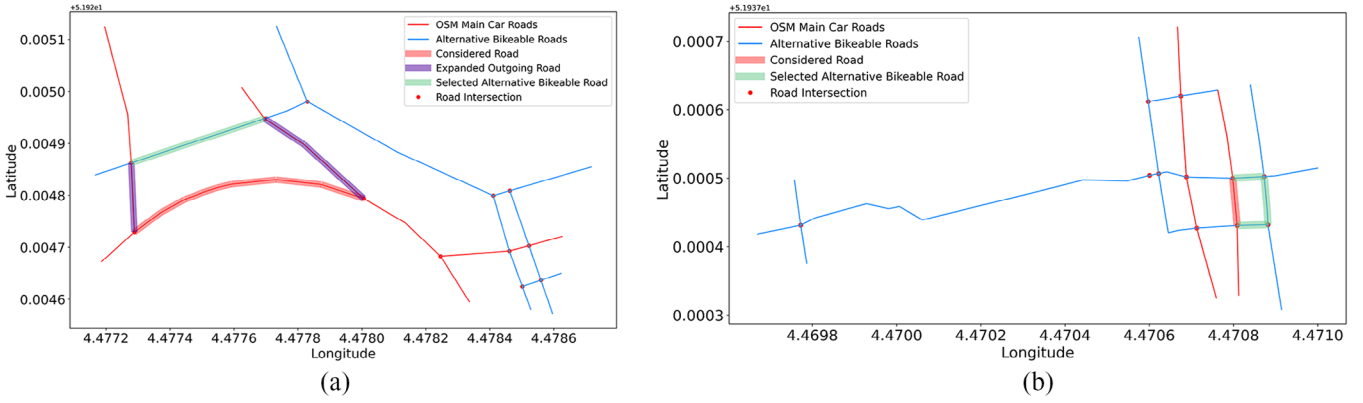


FIGURE 6 Rules are applied in different situations, and the thresholds are determined empirically based on applied dataset (Talking Bike). (a) shortcut: $U_{uv} = 94.5$ m, $B_{uv} = 60.2$ m, $\theta_{uv} = 0.64$, $\delta_{uv} = 34.3$ m. Applied rule: $\theta_{u,v} \leq 1.2$. (b) short considered roads: $U_{uv} = 22.2$ m, $B_{uv} = 33.3$ m, $\theta_{uv} = 1.50$, $\delta_{uv} = 11$ m. Applied rule: $\theta_{u,v} \leq 2$ & $\delta_{u,v} \leq 20$ m.

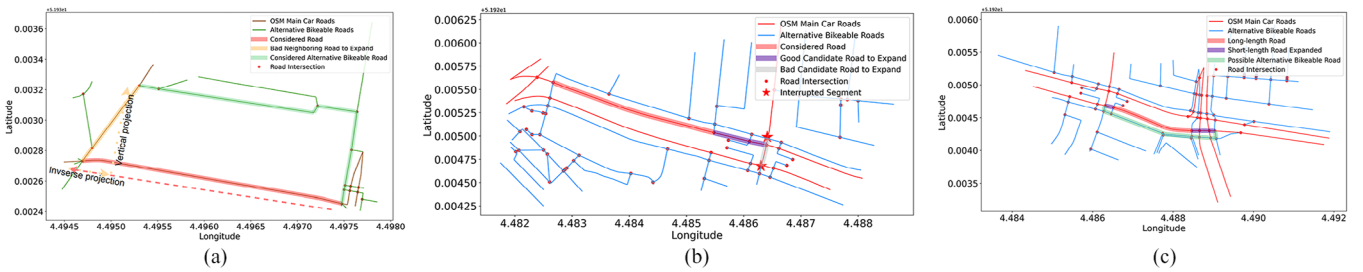


FIGURE 7 Special cases dealing with inaccuracies. (a) Bad expanded neighboring roads (denoted in yellow) form part of the loop. (b) Bad expanded neighboring roads (denoted in grey) disrupt the bicycle network between the two red star markers. (c) Expanding a long-length edge could help the labeling of tiny-length edge for which the rules are hard to satisfy.

more flexible constraint of the detour ratio θ_{uv} while enforcing a constraint on the detour distance δ_{uv} .

The empirical thresholds utilized for these two rules are as provided in the figures. In this study, we developed a visualization tool to display the identified OSM main car roads with low availability for cyclists, allowing users to easily zoom in and out to compare with real-world conditions. The thresholds were determined to ensure that the accurately identified main car roads meet a satisfactory level. However, the selection of diverse hyper-parameters can vary according to user preferences. Looser constraints typically lead to a more coarse-grained screening, resulting in more OSM main car roads being identified as having low availability for cyclists. Additionally, some roads that are integrated with real-life cycleways might be wrongly categorized as having low availability. According to observations with our dataset, the revised network is robust to changes in hyperparameters. The algorithm is given in Appendix Algorithm C3.

3.3.2 | Special cases

Up to this point, we have developed a framework to identify OSM main car roads with low availability for cyclists. Yet, this

framework is not sufficient. The complexity of real-world road situations could lead to undesired expansions, and the selected roads might disrupt the connectivity of Bikeable Roads due to inherent inaccuracies in OSM. These situations are addressed with the first two cases in this section. Additionally, we present a third case to enhance the revising process for very short road segments.

The first case involves confusion of expanded roads. When expanding the observation scope to include neighboring incoming and outgoing roads, the existence of an alternative bicycle road depends largely on the bikeable distance $B_{u,v}$. When $B_{u,v}$ is small, all the rules described above are easily satisfied and the corresponding OSM main car road is regarded to have low availability for cyclists. However, in the case shown in Figure 7a, where the extended neighboring roads form part of a loop, the bicycle distance $B_{u,v}$ naturally decreases. This is not the expansion that we want. Therefore, we impose restrictions on the vertical and inverse projection of neighboring roads to prevent them from forming a loop. This particular constraint is addressed in Algorithm C1.

The second case arises from the incomplete data within OpenStreetMap. As illustrated in Figure 7b, the roads under consideration disrupt the continuity of the bicycle network. Retaining these roads for map-matching is essential since

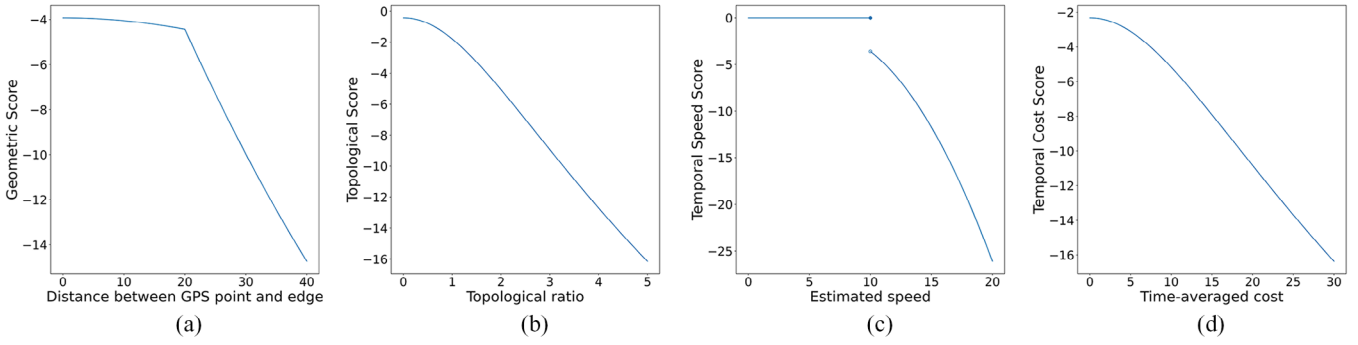


FIGURE 8 The function of (a) geometric score, (b) topological score, (c) temporal speed score, and (d) temporal cost score.

omitting them would force cyclists to take detours to reach the opposite side of the road. To tackle this problem, we evaluate all pairs of candidate nodes forming the outgoing/incoming candidate edges. If significant detours are required through Bikeable Roads to connect one node pair, the edge connecting these nodes should not be classified as having low availability. An extreme case involves two nodes isolated in Bikeable Roads despite being connected in the full road network, and in such instances, these roads should not be classified as low availability as well. The algorithm is given in Appendix Algorithm C2.

The third case pertains to very short uncertain roads, posing challenges in creating rules that universally apply without affecting the assessment of other roads. As shown in Figure 7c, given that most of these short roads come into play when expanding the observation scope of longer or medium-length roads, to which rules defined before apply effectively, it is reasonable to label them as having low availability if alternative bikeable roads are identified for the collective segment involving these short roads.

3.4 | Path scoring function

Given a series of candidate road segments, we use the geometric score to represent the local likelihood of observing the GPS record for each road segment, and the topological score and temporal score to measure the transition likelihood between road segments.

3.4.1 | Geometric score

The function of the geometric score based on the distance between GPS points and road segments (in meters) is depicted in Figure 8a. The pgMapMatch method utilizes a normal distribution with a sharp drop in probability between 0 and 20 m. In our approach, we have implemented a sharp drop for distances greater than 20 m. The design is attributed to the fact that, in many cases, the GPS error for the applied dataset could reach 20 m, as supported by the findings in [36].

3.4.2 | Topological score

The topological score considers the likelihood of road transitions from a spatial perspective and favors short distances. Figure 8b presents the evolution of topological score over topological ratio. The topological ratio is the ratio of estimated road network distance over straight-line distance between GPS points. Similar to the pgMapMatch method, we use a t-distribution in our approach. However, we increase the scale parameter to create a wider distribution. This accounts for the larger sampling interval in the applied dataset and accommodates the potential for cycling network distances to be much longer than the straight-line distance between GPS records, resulting in a higher topological ratio.

3.4.3 | Temporal score

The temporal score measures the likelihood of road transitions from a temporal perspective. As introduced in Section 3.1, our temporal score differs significantly from the pgMapMatch method. In our approach, it is the sum of two parts: speed score and time-averaged cost score, as depicted in Figures 8c and 8d. The speed score only penalizes estimated speeds over 10 m/s, which probably results from GPS drift errors.

For the time-averaged cost, the cost consists of a mapping cost and a traveling cost. Denoting e_0 and e_n two road candidates for two GPS points at time t_s and t_d , with $\{e_1, e_2, \dots, e_{n-1}\}$ being the intermediate road segment set found by the shortest path algorithm.

The mapping cost c_m imposes different penalties according to the classification detailed in Table 1. As discussed in Section 3.3, each class groups roads based on varying levels of availability for cyclists. This classification considers common route choice, OSM street descriptions, and real-world road network observations. Higher classes represent roads that are less available and less preferred by cyclists. Therefore, we apply progressively higher cost coefficients for each class. We denote class 0-4 as $C_0 - C_4$ and $\alpha = [\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4] \in \mathbb{R}_+^5$ the mapping penalty for each class. In the following equation, we set

TABLE 2 Comparison of the data provided by Ring-Ring and Tracefy.

Supplier ID	Nb of recorded trip	Trip duration (min)			Sampling interval (s)			Peak hours
		Mean	Median	75% quantile	Mean	Median	75% quantile	
Ring-Ring	1,032	7.0	5.2	8.7	12.7	4.0	11.0	5h-7h & 16h
Tracefy	100,202	6.2	4.7	8.1	30.0	14.0	30.0	18h-22h

$\alpha = [0, 1, 2, 3, 5]$:

$$c_m = \sum_{i \in \{0, n\}} \alpha_0 \mathbb{1}_{C_0}(e_i) + \alpha_1 \mathbb{1}_{C_1}(e_i) + \alpha_2 \mathbb{1}_{C_2}(e_i) + \alpha_3 \mathbb{1}_{C_3}(e_i) + \alpha_4 \mathbb{1}_{C_4}(e_i).$$

The traveling cost c_t is defined based on the road length, with additional coefficients incorporated for reverse travel direction and roads classification detailed in Table 1:

$$c_t = \sum_{i=0}^n f_i l_i (1 + \beta_d \mathbb{1}_{dir}(e_i)) (1 + \beta_0 \mathbb{1}_{C_0}(e_i) + \beta_1 \mathbb{1}_{C_1}(e_i) + \beta_2 \mathbb{1}_{C_2}(e_i) + \beta_3 \mathbb{1}_{C_3}(e_i) + \beta_4 \mathbb{1}_{C_4}(e_i)),$$

where $f_i \in [0, 1]$ represents the fraction to travel, l_i is the road length, $\mathbb{1}_{dir}(e_i)$ is a function indicating whether the travel is against the designed direction. The time-averaged cost is thus: $c = (c_m + c_t) \setminus (t_a - t_s)$. We use $\beta = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4] \in \mathbb{R}_+^5$ to denote extra travelling cost for each class. In our experiment, we set $\beta_d = 0.2$ and $\beta = [0, 0.1, 0.2, 0.3, 0.5]$. In both cost functions, our coefficients reflect the availability levels of different roads. The sensitivity of model performance to α and β is further discussed in Section 4.2.3.

4 | EVALUATION AND ANALYSIS

The objective of map-matching is to reconstruct the traveled roads from erroneous GPS data, providing insights into user behaviors. In this section, we first introduce our real-world dataset in Section 4.1, then evaluate our network and overall method in Section 4.2. Finally, we examine our mapped results to understand user behaviors in Section 4.3.

4.1 | Dataset description

The dataset utilized to validate our algorithm is sourced from the Talking Bike Program in the Netherlands, spanning a duration of two years from October 1st, 2020. The primary objective of this program is to collect policy-relevant data and gain insights into cycling travel behavior [2]. The Talking Bikes program has amassed an extensive GPS cycling dataset, capturing over one million bike rides annually across various locations throughout the Netherlands. The dataset is collected anonymously

and in line with General Data Protection Regulation (GDPR) [37] in Europe.

In this study, the urban area of Rotterdam is selected as the research area. Rotterdam is a Dutch port city and is home to approximately 0.6 million inhabitants. The municipality of Rotterdam is dedicated to transforming into a cycle-friendly city [38]. The OSM of our study area and the scatter plot of the Talking Bike Data can be found in Appendix D, Figure D1.

We have been very diligent in our data cleaning process, aiming to retain data integrity and minimize objective bias. Specifically, to address duplicate timestamps within a route, we adopted a strategy aligned with [39], keeping only the initial record. Additionally, trips consisting of two or fewer data points were excluded as they are not suitable for map-matching. This process resulted in the identification of 101,256 trips to be matched within the urban area. In the candidate road selection process, if there are only two or fewer data points for which we can find at least one road segment within 50 meters, these trips are removed. This results in 101,234 trips for the mapping process. For a detailed statistical distribution of the dataset, refer to Appendix D, Figure D2.

The dataset is collected by two suppliers: Siemens Mobility (Tracefy) and Ring-Ring. Siemens Mobility uses GPS trackers from Tracefy, collecting real-time trajectory data from shared bikes, constituting more than 98% of the dataset. Ring-Ring is an application designed to enhance cyclist safety by sending notifications to cyclists when approaching intersections. The geographical distribution of these two datasets is given in Figure E4, Appendix D.

Table 2 presents key numbers for these two datasets, and a detailed distribution is available in Figure E1, E2, and E3 of Appendix D. We could see that the Ring-Ring dataset exhibits higher frequency and longer trip durations. Additionally, the Ring-Ring dataset captures both morning and afternoon peak hours, whereas the Tracefy dataset predominantly reflects the afternoon-evening peak hour. The differences between these two datasets arise from the distinct user groups from which data are collected. Ring-Ring collects data from phone application users, while Tracefy shared bikes are mostly used by food deliverers. This disparity contributes to the dense distribution in residential areas within the Tracefy dataset. Moreover, the tendency of people to order food in the afternoon results in an afternoon-evening peak hour observed in Tracefy dataset.

Additionally, the high values for the mean sampling time and the 3rd quartile, relative to the median sampling time, indicate that many trips contain long segments with sparse GPS information. In urban environments, characterized by a dense road

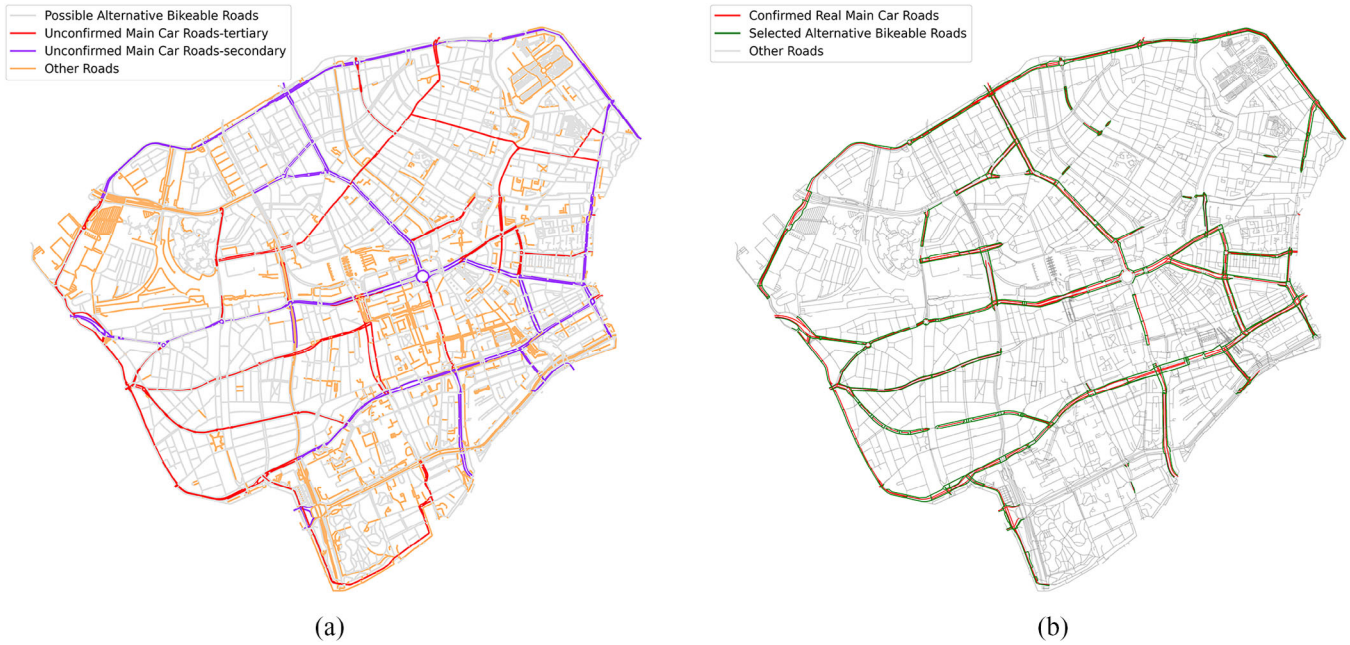


FIGURE 9 (a) The unconfirmed OSM main car roads (**tertiary** in red and **secondary** in purple), possible alternative bikeable roads (grey), and the other roads (orange) in our study area. (b) Confirmed OSM main car with low availability for cyclists roads and their alternative roads. Generally the former is surrounded by the latter.

network, this lack of detailed data makes the map-matching task more challenging.

Regarding the road network dataset, we employed the OSM data provided by BBBike. The raw OSM network contains 26,011 nodes and 61,110 edges (considering reverse edges, as is typical in real life). After removing redundant nodes, these numbers were reduced to 8,452 nodes and 25,696 edges. This indicates a remarkable compression of more than 67.5% of nodes and 57.9% of edges. This significant reduction contributes to lower computing demands, thereby enhancing the overall performance of our methodology. This efficiency is achieved while retaining all essential information, including road length, road type, and other vital parameters.

4.2 | Evaluation

In this section, we conduct a comprehensive evaluation of our method. We start by visualizing our identified OSM main car roads with low availability for cyclists and their alternative routes found in Section 3.3. Figure 9a outlines all unconfirmed main car roads (labelled as **secondary**, **tertiary** in OSM, **Observation 2**) and their possible alternative bikeable road types. Figure 9b presents the confirmed main car roads with low availability and their corresponding alternative roads on the map. Comparing these two figures, we can see that a significant percentage of main car roads are identified as having low availability for cyclists. This distinction highlights different levels of availability within the OSM main road network, even when roads have the same labels.

In the following section, we evaluate our method. Establishing ground truth is crucial for validating the accuracy and

effectiveness of our approach. Due to the lack of ground truth, we create a evaluation dataset by randomly selecting 100 trajectories, similar with [11]. These selected trajectories serve as a reference point, or “ground truth,” for our evaluation. We begin by detailing the evaluation dataset and evaluation metrics, followed by a comparison between our method and pgMapMatch. Finally, we perform a sensitivity analysis to assess model performance under various conditions, including parameter variations and GPS data quality. At this stage, we also validate our refined road availability classification for OSM main car roads (detailed in Section 3.3).

4.2.1 | Evaluation dataset and metrics

During manual labelling, we establish the ground truth trajectories by examining the real-world environment and adhering to the following principles:

1. If a dedicated parallel cycleway nearby is available, cyclists will use the cycleway.
2. When the time interval is large between two sample points, we use Google Maps to assist with routing. In cases the route suggested cause a severe detour in comparison to a more direct trajectory (against the traffic flow of the cycle lane), we assume that cyclists will not follow Google Maps' indications and will go in the reverse direction.
3. In cases where Google Maps propose multiple trajectories and the trip is two-way, we assume that cyclists choose roads they are familiar with. Therefore, we will select the route recommendation that overlaps with the other direction.

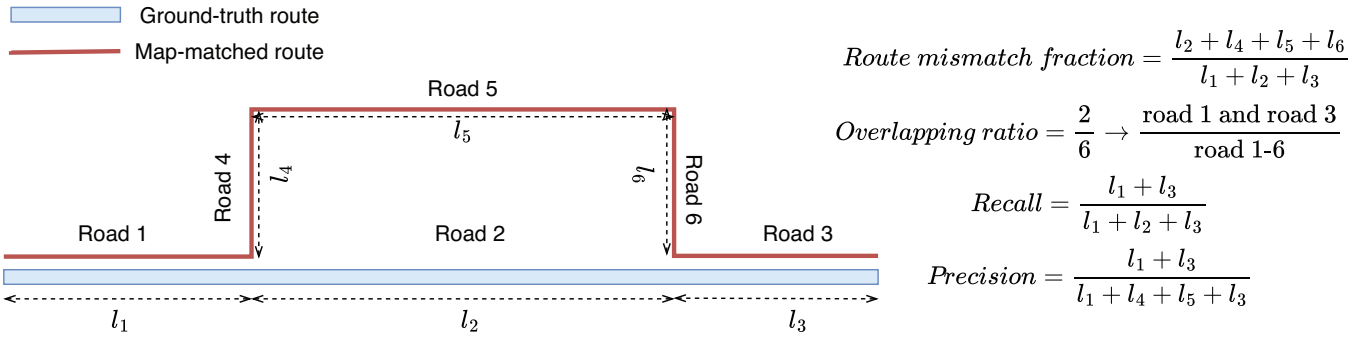


FIGURE 10 Illustrative example of map-matching evaluation metrics.

4. Cyclists typically avoid crossing shopping centers where they need to walk their bikes, unless there are multiple GPS points near the shopping center and the pedestrian roads provide a direct shortcut.
5. Cyclists typically do not cycle in parks and gardens unless there are GPS records very close to these roads and far from other roads.
6. Cyclists typically do not like crossing roads, particularly boulevards with high traffic volumes.

Due to the complexity of manual labeling, only 100 trajectories are selected. The selection and labeling are performed by a single individual. To ensure that our evaluation dataset is representative, we conduct the Kolmogorov–Smirnov test during sampling. We ensure that the p-values for the distributions of sampling time, travel time, and GPS records per trip are all higher than 0.05, indicating that the sampled trajectories are not statistically different from the complete dataset. The p-values are given in Table F1 and the comparison of the sampled dataset and complete one is provided Figure F1, F2, F3, and F4 in Appendix F.

Given the mapped route and ground-truth route, we employ four measures to compare: route mismatch fraction [40], overlapping ratio [41], recall [42], and precision [42]. For each metric, their definitions are as follows and an example is given in Figure 10:

- Route mismatch fraction is the total length of roads that have been erroneously included (either incorrectly added or incorrectly subtracted) by the map-matching algorithm, divided by the length of the ground-truth route.
- The overlapping ratio is the ratio of the number of common roads between the ground-truth and map-matched routes to the total number of unique roads present in both routes.
- Recall is the ratio of the total length of the common roads between the ground-truth route and the map-matched route to the total length of the ground-truth route.
- Precision is the ratio of the total length of the common roads between the ground-truth route and the map-matched route to the total length of the map-matching route.

By definition, higher recall, precision, overlapping ratio, and smaller route mismatch fraction indicate a better map-matched result.

TABLE 3 Performance comparison between our method and pgMapMatch. Our method achieves better performance than pgMapMatch across all metrics.

Method	Route mismatch fraction	Overlapping ratio	Recall	Precision
Our method	0.19	0.794	0.895	0.909
pgMapMatch	0.616	0.523	0.682	0.694
Improvement	69.2%	51.8%	31.2%	31.0%

4.2.2 | Comparison with pgMapMatch

In this subsection, we compare the mapped result of our method with the pgMapMatch method. For a fair comparison, we utilize our simplified network, ensuring that all network properties are maintained without influencing accuracy. We remind that, in pgMapMatch, the temporal component considers the estimated speed over the maximum allowed speed and favors smaller speeds. Given the absence of a real maximum speed limit for cyclists, we adjust the cost function to be road length. We also harmonize the temporal score, geometrical score, and topological score with the same function adopted in our method. These adjustments ensure a fair and accurate comparison, considering the characteristics of the dataset.

From Table 3, we observe an important improvement of our method compared with pgMapMatch. This improvement is primarily due to our comprehensive evaluation of road availability. Our method distinguishes between different levels of availability for cyclists, which is reflected in our cost functions. In contrast, pgMapMatch does not account for road types and primarily focuses on optimizing GPS recorded distances and shortest paths. In Figure 11, we draw the heatmap difference in edge counts between the pgMapMatch method and our approach. Upon comparison with Figure 9b, it becomes evident that the pgMapMatch method tends to match on main car roads with low availability. In contrast, our method intentionally incorporate additional traveling cost to these roads, effectively preventing excessive mapping.

In the following, we present two real-world scenarios in Figure 12 to intuitively show the difference in performance between our method and pgMapMatch. From the satellite image, cycleways are observed alongside the main car roads. In this context, cyclists tend to cycle on the cycleways instead of the

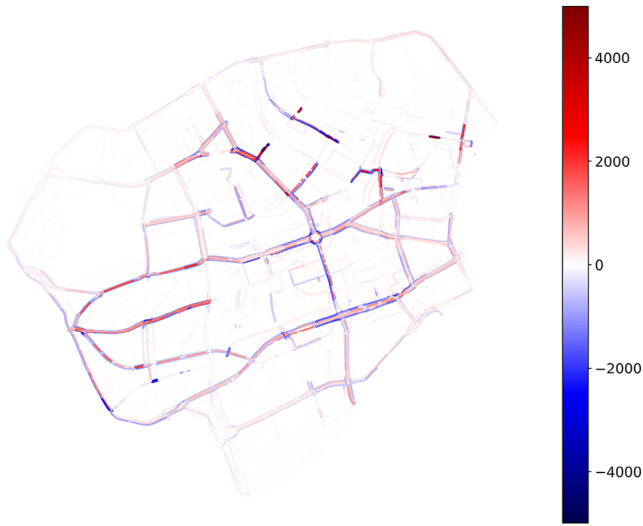


FIGURE 11 The heatmap illustrates the differences in bike counting between the mapping results of pgMapMatch and our method. The red color indicates roads that are more frequently mapped in pgMapMatch, while the blue indicates the inverse.

car roads. We can observe that even when the GPS records fall on or are very close to the car roads, our method successfully maps them back to the cycleways. Moreover, the connection roads among the mapped roads also favor cycleways, a result of our designed extra traveling cost for other road types. In

contrast, the result of the pgMapMatch method shows cyclists cycling on the main car roads, which is unrealistic.

4.2.3 | Sensitivity analysis

In this section, we focus on assessing model performance under different conditions, including variations in parameters and data quality. When varying parameter, we also compare the performance of our method with different classifications of OSM main road availability.

Parameter variations and different main car road availability classification. During the evaluation stage of main car road availability, we developed a visualization script with zoom-in and zoom-out capabilities, which simplifies the visual browsing of large areas. Our observations indicate that the revised network remains relatively stable with variations of ± 0.1 in both the detour ratio θ_{mv} and the detour distance δ_{mv} . This stability suggests that minor adjustments to these parameters have a negligible impact on the overall network. Additionally, the visualization script supports parameter tuning by offering a clear and adjustable view of the network, facilitating the exploration over different sizes of network. Therefore, the sensitivity of model performance to θ_{mv} and δ_{mv} is not discussed here. Instead, this section examines the model's sensitivity to the parameters α and β in the mapping cost c_m and traveling cost c_t .

Given that our classification is based on common route choices, we preserve the relative weight between different

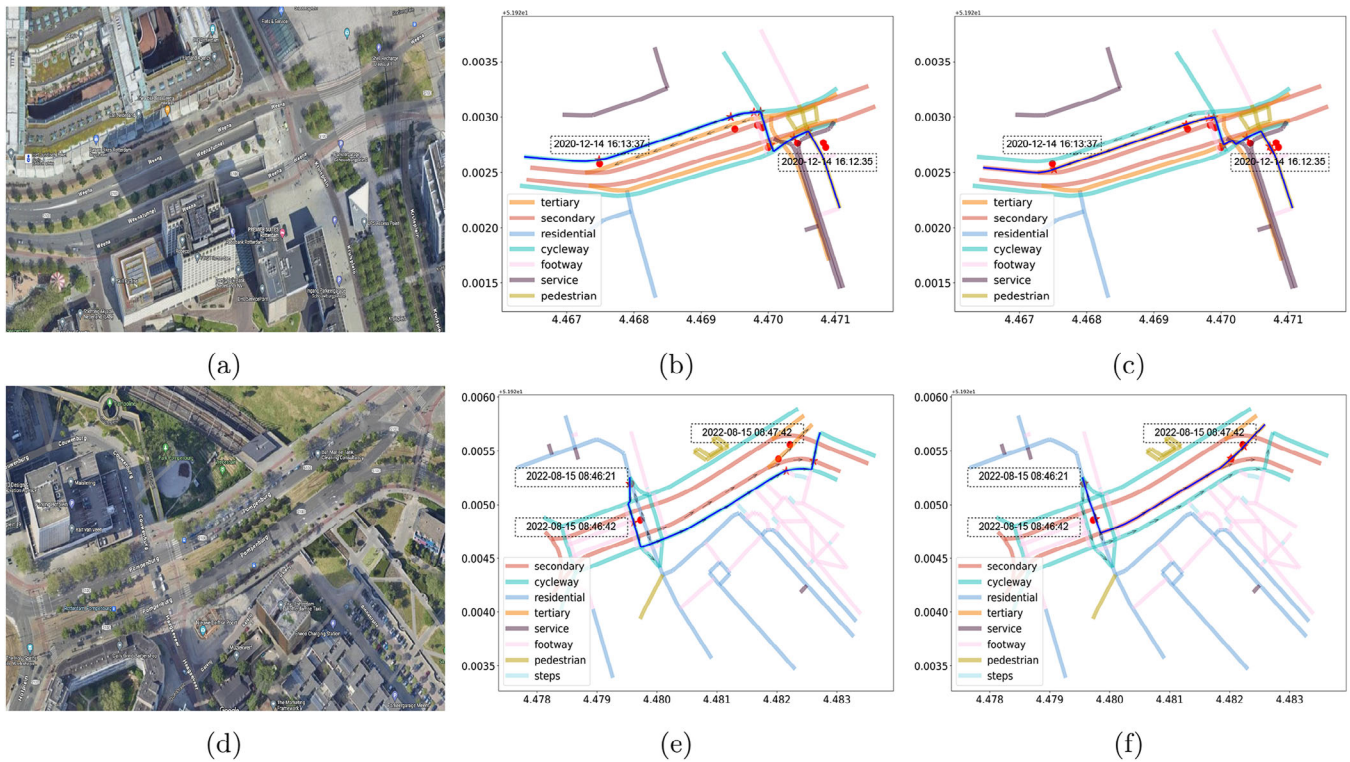


FIGURE 12 Comparison of our method and pgMapMatch using two examples. (a) and (d) show satellite images of the road network where cycleways run alongside the main car roads. (b) and (e) display the mapped results of our method, with red dots representing GPS points, red stars indicating estimated positions, and light blue lines representing the mapped route. (c) and (f) use the same symbols to present the results of pgMapMatch.

TABLE 4 Parameter sensitivity analysis. comparison-1 for different parameter settings. s_α and s_β are scaling factors of α and β , and “—” indicates no change from the original parameter. The figures in bold represent the best values for each metric and for each method.

Method	Scaling factor	Mismatch fraction	Overlapping ratio	Recall	Precision
Our method	—	0.190	0.794	0.895	0.909
	$s_\alpha = 0.2$	0.211	0.783	0.885	0.897
	$s_\alpha = 0.5$	0.203	0.788	0.888	0.902
	$s_\alpha = 2$	0.221	0.771	0.878	0.893
	$s_\alpha = 5$	0.302	0.727	0.832	0.857
	$s_\beta = 0.2$	0.255	0.724	0.861	0.876
	$s_\beta = 0.5$	0.238	0.749	0.87	0.885
	$s_\beta = 2$	0.237	0.773	0.871	0.876
	$s_\beta = 5$	0.258	0.755	0.862	0.875
Comparison 1	—	0.213	0.786	0.883	0.897
	$s_\alpha = 0.2$	0.225	0.772	0.877	0.89
	$s_\alpha = 0.5$	0.216	0.778	0.882	0.896
	$s_\alpha = 2$	0.238	0.77	0.869	0.884
	$s_\alpha = 5$	0.303	0.731	0.832	0.857
	$s_\beta = 0.2$	0.314	0.688	0.831	0.846
	$s_\beta = 0.5$	0.284	0.716	0.846	0.861
	$s_\beta = 2$	0.246	0.772	0.867	0.871
	$s_\beta = 5$	0.277	0.747	0.853	0.865
Comparison 2	—	0.271	0.738	0.855	0.87
	$s_\alpha = 0.2$	0.278	0.741	0.852	0.864
	$s_\alpha = 0.5$	0.26	0.744	0.86	0.875
	$s_\alpha = 2$	0.305	0.714	0.838	0.851
	$s_\alpha = 5$	0.376	0.682	0.795	0.822
	$s_\beta = 0.2$	0.266	0.717	0.855	0.871
	$s_\beta = 0.5$	0.293	0.709	0.843	0.858
	$s_\beta = 2$	0.34	0.7	0.822	0.832
	$s_\beta = 5$	0.428	0.666	0.79	0.793

classes, and use scaling factors s_α , s_β to adjust α and β . For example, when $s_\alpha = 2$, all α values become twice their original values. To examine the effects of the revised main car roads, we introduce two comparison models:

Comparison-1: All **secondary** and **tertiary** roads are not considered to have low availability for cyclists. Specifically, all **tertiary** roads are classified as Class 1, and all **secondary** roads are classified as Class 3 in Table 1.

Comparison-2: All **secondary** and **tertiary** roads are considered to have low availability for cyclists. Specifically, all **tertiary** and **secondary** roads are classified as Class 4 in Table 1.

For both comparison models, the classification of other roads in Table 1 is retained. The performance under different α and β values for these two comparison models, along with our method, is shown in Table 4.

In Table 4, we observe that the performance of all methods significantly decreases when setting $s_\alpha = 5$, indicating that the model pays too much attention to mapped roads at each GPS

point, resulting in insufficient full route optimization. Moreover, when $s_\alpha < 5$, model performance is more sensitive to s_β . This is because the mapped cost applies only to available GPS points, whereas the traveling cost affects the selection of all intermediate roads; thus, variations in s_β have a greater impact than variations in s_α .

Moreover, our method generally outperforms comparison-1, with both methods showing decreased performance when deviating from the original parameter values. Nevertheless, our method remains less sensitive to parameter changes compared to comparison-1, demonstrating its robustness. The difference becomes more significant for $s_\beta = 0.2$, which corresponds to a smaller traveling cost difference among different roads. By differentiating OSM main car roads availabilities, we create a map-matching hierarchy among roads labeled **secondary** and **tertiary** which helps to preserve the traveling cost difference. Additionally, comparison-2 generally performs worse, indicating that treating all OSM main car roads as low availability is ineffective and that further refinement is necessary.

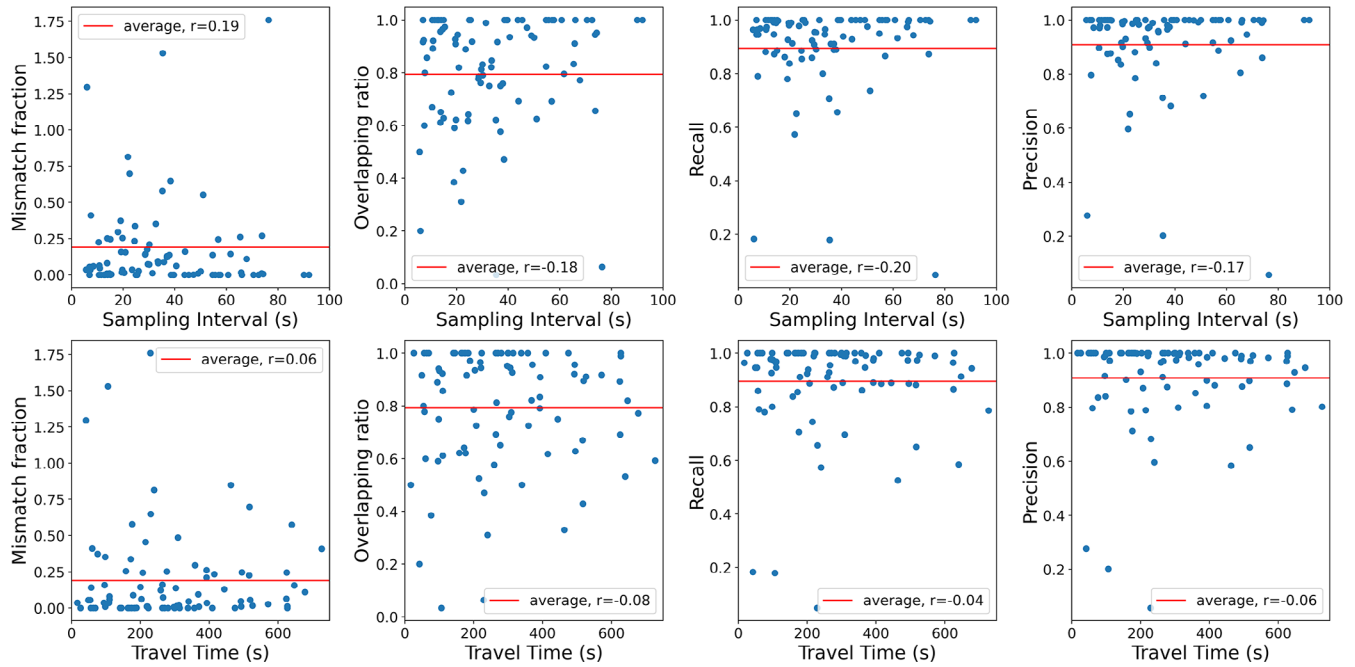


FIGURE 13 Performance metrics (route mismatch fraction, overlapping ratio, recall, and precision) as functions of sampling interval and travel time. For improved visualization, we display only the trips with sampling intervals below 100 seconds and travel times below 750 seconds, as data points beyond these thresholds are too sparse. The Pearson correlation coefficient between the sampling interval/travel time and the evaluation metrics for the entire dataset is denoted as “r” in the figures.

GPS data quality. To evaluate our model’s performance under different data settings, we present Figure 13, illustrating the model performance as a function of GPS sampling interval and travel time. The sampling interval is a crucial aspect of data quality, as it determines how frequently GPS data is collected, and shorter sampling intervals provide more detailed trip information. Excluding a few outlier points, the figure shows no significant difference in model performance for trips with sampling intervals ranging from 0 to 100 s and travel times between 0 and 750 s. This observation is supported by a Pearson correlation coefficient lower than 0.2. Given that a 100-s interval is relatively large, these results indicate that the algorithm is robust to variations in sampling frequency and performs effectively even with datasets that have larger intervals between GPS records.

4.3 | User behavior analysis

The accuracy of map-matching is crucial for aligning identified roads with the actual paths taken by users, forming the basis for extracting meaningful insights into their behavior, preferences, and travel patterns. In this section, we infer user behavior based on our mapped results, focusing on two perspectives: the analysis of the most frequently used road and the examination of differences between two distinct user groups (Tracefy and Ring-Ring).

The heatmap, derived from the edge count in our mapped results, is depicted in Figure 14a. This heatmap directly reveals the most frequently traveled roads for the studied user groups. In contrast, Figure 14b showcases the bicycle network provided

by the Rotterdam traffic department, offering a perspective from a policy design standpoint. It is important to note that the latter network serves as a rough illustration of the real world, given that its coordinates are not in longitude and latitude. The comparison of these two networks allows us to assess whether cyclists are using the main cycling infrastructures as intended by the municipality.

Upon examination, Figure 14a reveals a distinct network hierarchy, where the skeletal structure of frequently used roads is clearly visible. This hierarchy generally aligns with the network map presented in Figure 14b. However, two interesting differences emerge upon closer inspection: (1) Residential roads are utilized more frequently than expected in our dataset, which aligns with findings from [43] on the high usage of low-priority roads. This increased usage may be attributed to their role as shortcuts for cyclists, especially food deliverers. Additionally, frequent food orders from residential areas contribute to the high visitation rates. (2) The city’s outer ring is relatively less visited in our dataset. This could be a result of data filtering for the city borders or the lesser influence of the Talking Bike campaign on individuals residing outside the city center.

To compare different user behaviors in Ring-Ring and Tracefy data, we use the following indices, as described in Section 3.4: speed, time-averaged cost, and distance. The speed represents the estimated speed of cyclists based on the mapped result, the time-averaged cost is the same as introduced in Section 3.4, and the distance refers to the distance between GPS samples and their corresponding mapped roads. Mean values of these indices for the Ring-Ring and Tracefy datasets are presented in Table 5.

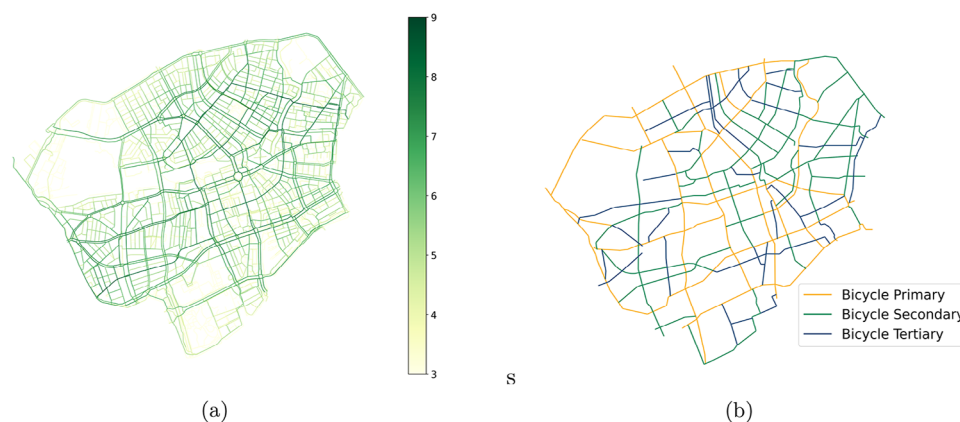


FIGURE 14 (a) Heatmap illustrating edge counts with our method, derived from the mapped results using OSM, the count is plotted with log scale. (b) Policy-privileged roads, representing a theoretical design of the municipality. The difference comes from the fact that people can use roads that the municipality has not identified as cycleways.

TABLE 5 Comparison of the user behaviors in Ring-Ring and Tracefy Dataset.

SupplierID	Mean speed	Mean time-averaged cost	Mean distance
Ring-Ring	4.1 m/s	4.9	5.1 m
Tracefy	5.1 m/s	6.8	7.9 m

In Table 5, it is evident that users in the Tracefy dataset exhibit a higher mean speed compared to those in the Ring-Ring dataset. This aligns with the expectation that food deliverers tend to travel quickly to deliver on time. Additionally, there is a higher mean time-averaged cost in the Tracefy group. Since this additional cost is based on travel direction and road types, it implies a higher frequency of rule violations in the Tracefy group, a common occurrence among food deliverers. Furthermore, we observe a greater distance in the Tracefy group, which can be reasonably attributed to the fact that the Tracefy group travels more frequently in densely built-up areas (as illustrated in Figure E5), which can significantly impact GPS accuracy [7]. In summary, the overall comparison allows us to infer distinct user behaviors and link them to real-world scenarios.

5 | CONCLUSION

Accurate mapping of bicycle travel data plays a crucial role in providing reliable support for traffic management. For example, understanding the bicycle traffic flow is essential for optimizing the overall cycling experience, especially when upgrading cycling facilities within budget constraints. Insights into cycling travel times can also contribute to the design of traffic schemes, prioritizing the cycling mode in urban traffic.

Existing literature on map-matching often relies on networks filtered by road types in OpenStreetMap (OSM) or uses the entire traffic network. The former approach is inadequate because cyclists commonly violate rules for convenience, while the latter is too coarse, as cycleways are typically designed along main car roads in urban areas. Mapping GPS records to main

car roads can introduce bias in the distribution of bicycle trips in real-life scenarios.

In our method, we innovatively address these limitations by:

- **Careful classification of availability for bicycles:** We have conducted a detailed availability evaluation of roads in the OSM data, specifically tailored for bicycles and with particular attention to urban main road networks.
- **Extended map-matching scoring functions:** Our method includes an extended scoring function that accounts for bicycle speed, road types, and road availabilities.

To evaluate the model's performance, we create an evaluation dataset by randomly selecting 100 trajectories and manually labeling them. Compared with the baseline model, our approach demonstrates a more accurate mapping result for bicycle travel data. Additionally, the special attention on main car road availability evaluation has been shown to enhance the model's robustness to parameter changes. Our model has also demonstrated consistent performance across various sampling intervals and travel times. This robustness makes our method highly adaptable for practical applications in real world. Furthermore, we successfully extract user behaviors based on our mapped results, aligning with the user distribution in our datasets.

Although we use Rotterdam as a case study, our method is explainable and grounded in well-established route choice studies, supporting its applicability beyond Rotterdam with minor user changes. However, we acknowledge that infrastructure and usage patterns vary among cities, necessitating adaptations for different urban environments. We provide a detailed classification of road availability. This explanation aids users in adapting our method to other scenarios and cities. Additionally, different user patterns might require adjustments in the cost function definitions to reflect the common usage of certain road types.

Further improvements to our method could focus on enhancing its ability to assess network availability. Our current method struggles with evaluating the availability of circular intersections, as these are often misinterpreted as loop situations. Employing machine learning techniques, particularly

with a sufficient number of network samples, might improve the accuracy of identifying roads with low availability. Given that many real-life travel datasets are collected without ground truth, another avenue for improvement is to propose comprehensive evaluation metrics tailored to this type of data. This would provide a more nuanced and accurate assessment of the performance of map-matching algorithms in scenarios where ground truth is unavailable. Additionally, other road properties, such as grade and pedestrian invasion, could be studied and incorporated into the cost function for future research.

AUTHOR CONTRIBUTIONS

Ting Gao: Conceptualization; data curation; formal analysis; methodology; validation; visualization; writing—original draft; writing—review & editing. **Winnie Daamen:** Resources; supervision; validation; visualization; writing—review & editing. **Panchamy Krishnakumari:** Supervision. **Serge Hoogendoorn:** Funding acquisition; supervision; writing—review & editing.

ACKNOWLEDGEMENTS

This document has been produced by the “EMERALDS” project which has received funding under the Horizon Europe research and innovation programme (GA No. 101093051).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from Rijkswaterstaat (the Dutch Ministry of Transportation). Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the authors with the permission of Rijkswaterstaat (the Dutch Ministry of Transportation).

ORCID

Ting Gao  <https://orcid.org/0009-0003-5677-6545>

REFERENCES

- Euronews: Cycling in europe: Which countries and cities are the most and least bicycle-friendly (2023)
- Mobiliteitsplatform: Siemens en ringring verzamelen fietsdata voor talking bikes (2020)
- Yuan, Y., Wang, K., Duives, D., Hoogendoorn, S., Hoogendoorn-Lanser, S., Lindeman, R.: Bicycle data-driven application framework: a dutch case study on machine learning-based bicycle delay estimation at signalized intersections using nationwide sparse gps data. *Sensors* 23(24), 9664 (2023)
- Geurs, K.T., Paix, L.L., Van Weperen, S.: A multi-modal network approach to model public transport accessibility impacts of bicycle-train integration policies. *Eur. Trans. Res. Rev.* 8(4), 1–15 (2016)
- Meng, L., Somenahalli, S., Berry, S.: Policy implementation of multi-modal (shared) mobility: Review of a supply-demand value proposition canvas. *Trans. Rev.* 40(5), 670–684 (2020)
- Mooney, P., Minghini, M.: A review of openstreetmap data. In: *Mapping and the Citizen Sensor*, pp. 37–59. Ubiquity Press, London (2017)
- Kaplan, E.D., Hegarty, C.: *Understanding GPS/GNSS: Principles and Applications*. Artech House, Boston (2017)
- Millard-Ball, A., Hampshire, R.C., Weinberger, R.R.: Map-matching poor-quality gps data in urban environments: The pgmapmatch package. *Transport. Plan. Technol.* 42(6), 539–553 (2019)
- Hashemi, M., and Karimi, H.A.: A critical review of real-time map-matching algorithms: Current issues and future directions. *Comput. Environ. Urban Syst.* 48, 153–165 (2014)
- Huang, Z., Qiao, S., Han, N., Yuan, C.-a., Song, X., Xiao, Y.: Survey on vehicle map matching techniques. *CAAI Trans. Intell. Technol.* 6(1), 55–71 (2021)
- Berjisan, E., Bigazzi, A.: Evaluation of map-matching algorithms for smartphone-based active travel data. *IET Intel. Transport Syst.* 17(1), 227–242 (2023)
- Schweizer, J., Bernardi, S., Rupi, F.: Map-matching algorithm applied to bicycle global positioning system traces in bologna. *IET Intel. Transport Syst.* 10(4), 244–250 (2016)
- Bergman, C., Oksanen, J.: Conflation of openstreetmap and mobile sports tracking data for automatic bicycle routing. *Trans. GIS* 20(6), 848–868 (2016)
- Wang, J.-H., Gao, Y.: High-sensitivity gps data classification based on signal degradation conditions. *IEEE Trans. Veh. Technol.* 56(2), 566–574 (2007)
- Newson, P., Krumm, J.: Hidden markov map matching through noise and sparseness. In: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 336–343. ACM, New York (2009)
- Shen, Z., Yang, K., Zhao, X., Zou, J., Du, W., Wu, J.: Dmm: A deep reinforcement learning based map matching framework for cellular data. *IEEE Trans. Knowl. Data Eng.* (2024)
- Feng, J., Li, Y., Zhao, K., Xu, Z., Xia, T., Zhang, J., Jin, D.: Deepmm: Deep learning based map matching with data augmentation. *IEEE Trans. Mob. Comput.* 21(7), 2372–2384 (2020)
- Jiang, L., Chen, C.-X., Chen, C.: L2mm: learning to map matching with deep models for low-quality gps trajectory data. *ACM Trans. Knowl. Discovery Data* 17(3), 1–25 (2023)
- Jin, Z., Kim, J., Yeo, H., Choi, S.: Transformer-based map-matching model with limited labeled data using transfer-learning approach. *Transport. Res. Part C: Emerg. Technol.* 140, 103668 (2022)
- O’Hern, S., Stephens, A.N., Young, K.L., Koppel, S.: Personality traits as predictors of cyclist behaviour. *Accid. Anal. Prev.* 145, 105704 (2020)
- Schuessler, N., Axhausen, K.W.: Map-matching of gps traces on high-resolution navigation networks using the multiple hypothesis technique (MHT). *Arbeitsberichte Verkehrs-und Raumplanung* 568, 1–22 (2009)
- Dalumpines, R., Scott, D.M.: Gis-based map-matching: Development and demonstration of a postprocessing map-matching algorithm for transportation research. In: *Advancing Geoinformation Science for a Changing World*, pp. 101–120. Springer, Cham (2011)
- Li, S., Muresan, M., Fu, L.: Cycling in toronto, ontario, canada: Route choice behavior and implications for infrastructure planning. *Transp. Res. Rec.* 2662(1), 41–49 (2017)
- Perrine, K., Khani, A., Ruiz-Juri, N.: Map-matching algorithm for applications in multimodal transportation network modeling. *Transp. Res. Rec.* 2537(1), 62–70 (2015)
- Reggiani, G., Verma, T., Daamen, W., Hoogendoorn, S.: A multi-city study on structural characteristics of bicycle networks. *Environ. Plan. B: Urban Analyt. City Sci.* 50(8), 2017–2037 (2023)
- Zhang, H., Malczewski, J.: Quality evaluation of volunteered geographic information: The case of openstreetmap. In: *Crowdsourcing: Concepts, Methodologies, Tools, and Applications*, pp. 1173–1201. IGI Global, Hershey, PA (2019)
- Forney, G.D.: The viterbi algorithm. *Proc. IEEE* 61(3), 268–278 (1973)
- OpenStreetMap Contributors: <https://wiki.openstreetmap.org/wiki/Key:highway>
- Broach, J., Dill, J., Gliebe, J.: Where do cyclists ride? a route choice model developed with revealed preference gps data. *Transport. Res. Part A: Policy and Practice* 46(10), 1730–1740 (2012)
- Misra, A., Watkins, K.: Modeling cyclist route choice using revealed preference data: an age and gender perspective. *Transp. Res. Rec.* 2672(3), 145–154 (2018)
- Bernardi, S., Puella, L.L.P., Geurs, K.: Modelling route choice of dutch cyclists using smartphone data. *J. Transp. Land Use* 11(1), 883–900 (2018)
- Still, M.L.: Expert cyclist route planning: Hazards, preferences, and information sources. In: *HCI International 2020—Late Breaking Papers: Digital*

- Human Modeling and Ergonomics, Mobility and Intelligent Environments: 22nd HCI International Conference, HCII 2020, pp. 221–235. Springer, Cham (2020)
33. Schepers, P., Twisk, D., Fishman, E., Fyhri, A., Jensen, A.: The dutch road to a high level of cycling safety. *Saf. Sci.* 92, 264–273 (2017)
 34. Menghini, G., Carrasco, N., Schüssler, N., Axhausen, K.W.: Route choice of cyclists in zurich. *Transport. Res. Part A: Policy Pract.* 44(9), 754–765 (2010)
 35. Ton, D., Cats, O., Duives, D., Hoogendoorn, S.: How do people cycle in amsterdam, netherlands?: Estimating cyclists' route choice determinants with gps data from an urban area. *Transp. Res. Rec.* 2662(1), 75–82 (2017)
 36. Mok, E., Retscher, G., Wen, C.: Initial test on the use of gps and sensor data of modern smartphones for vehicle tracking in dense high rise environments. In: 2012 Ubiquitous Positioning, Indoor Navigation, and Location Based Service (UPINLBS), pp. 1–7. IEEE, Piscataway (2012)
 37. GDPR Info: General data protection regulation (gdpr) (2024)
 38. Dutch, B.: Rotterdam takes an important step towards becoming a cycle-friendly city. WordPress (2021)
 39. Zhou, C., Jia, H., Juan, Z., Fu, X., Xiao, G.: A data-driven method for trip ends identification using large-scale smartphone-based gps tracking data. *IEEE Trans. Intell. Transp. Syst.* 18(8), 2096–2110 (2016)
 40. Newson, P., Krumm, J.: Hidden markov map matching through noise and sparseness. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 336–343. ACM, New York (2009)
 41. Yang, C., Gidofalvi, G.: Fast map matching, an algorithm integrating hidden markov model with precomputation. *Int. J. Geograph. Inf. Science* 32(3), 547–570 (2018)
 42. Wei, H., Wang, Y., Forman, G., Zhu, Y.: Map matching: Comparison of approaches using sparse and noisy data. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 444–447. ACM, New York (2013)
 43. Rupi, F., Schweizer, J.: Evaluating cyclist patterns using gps data from smartphones. *IET Intel. Transport Syst.* 12(4), 279–285 (2018)

How to cite this article: Gao, T., Daamen, W., Krishnakumari, P., Hoogendoorn, S.: Map-matching for cycling travel data in urban area. *IET Intell. Transp. Syst.* 1–26 (2024). <https://doi.org/10.1049/itr2.12567>

APPENDIX A: NOTATION

TABLE A1 Notation for road network.

Symbol	Signification
\mathbb{B}	Set of Bikeable Roads (u, v)
\mathbb{M}_o	Set of main car roads in OSM
$G(V, E)$	Directed road network with set of nodes V and set of edges (roads) E
$e_i = (u, v)$	i -th edge from node u to node v
$N^-(v)$	Incoming edges of node v , $N^-(v) = \{(u, v) (u, v) \in E\}$
$N^+(v)$	Outgoing edges of node v , $N^+(v) = \{(v, u) (v, u) \in E\}$
l_i	Length of i -th edge
o_i	Oneway indicator of i -th edge, 0 for two-way road, 1 for one-way road
t_i	Road type of i -th edge, e.g. “primary”

APPENDIX B: SIMPLIFIED GRAPH

To obtain the simplified graph, we start by examining each edge to identify downstream candidate edges that can be merged with it. We employ a recursive function to investigate these candidate edges. The candidate edge can be merged with the original edge

ALGORITHM B1 Simplify graph.

```

Function recursive_programming(parent, edge, d):
    // Return if the edge searching reaches
    // its end or input edge has no
    // downstream edge
    if d[edge] == -1 or len(d[edge]) == 0 then
        | return parent, d
    end
    next_edge ← d[edge][0] // Go to downstream
    // edge
    parent ← parent + [edge] // Update parents
    // (upstream edges)
    if next_edge in parent then
        | return parent, d // Get rid of loops
    end
    parent, d ← recursive_programming(parent,
    // next_edge, d)
    if d[next_edge] ≠ -1 then
        | d[edge] ← d[edge] + d[next_edge]
    end
    d[next_edge] ← -1 // Mark the edge as
    // searching end
    return parent, d

```

```

Function simplify_graph():
    // Read OpenStreetMap linestrings
    raw_df ← transform_osm()
    // d: Dictionary of adjacent linestring
    // segments (intersection excluded)
    // with similar road types and direction
    // information
    // Format d[upstream edge]=[downstream
    // edge]
    d ← adjacent_similar_edges(raw_df) // Update the
    // downstream edges to include all edges
    // that could be merged
    foreach edge in raw_df.index do
        | d ← recursive_programming([], edge, d)
    end
    // Merge linestring segments into one
    // merged_edges ← []
    foreach edge in d.keys() do
        | if d[edge] ≠ -1 then
            | merged_edges.append([edge] + d[edge])
        end
    end
    return merged_edges

```


only when they share the same road type and road direction, and their junction does not connect to the rest of the graph. When a candidate edge is deemed mergeable, it is appended to a list of merged edges, and the recursive function is invoked with the candidate edge as the new starting point. This recursive process continues until no further candidate edges can be merged with the current edge. The length of each edge in the simplified graph is calculated as the sum of the lengths of its constituent merged edges.

APPENDIX C: ALGORITHM

ALGORITHM C1 Enlarge observation scope.

Function Downstream

```

Dfs( $u, e, km, visited, depth, d_1, d_2$ ):
    //  $u$ : Considered node
    //  $e$ : Current edge index
    //  $d_1$ : Dictionary of expanded downstream
    //       nodes {node: browsed distance}
    //  $d_2$ : Dictionary of expanded downstream
    //       nodes {node: browsed edge path}
    //  $\sigma_p$ : Vertical or inverse projection
    //       threshold
    //  $visited$ : List of downstream nodes visited
    //  $km$ : browsed km
    if reached searching depth or searching distance threshold then
        | return  $d_1, d_2$ 
    end
    // Update the browsed path
     $d_2[u] \leftarrow d_2[visited[-1]] + [e]$ 
    if  $len(visited) \geq 2$  then
        |  $a_k \leftarrow$  accumulated projection on the vertical
        |   direction of considered road
        |  $a_i \leftarrow$  accumulated projection on the inversed
        |   direction of considered road
        | if  $a_k \geq \sigma_p$  or  $a_i \geq \sigma_p$  then
        |     | return  $d_1, d_2$ 
        | end
    end
    end
    if  $u$  has incoming or outgoing bikeable roads then
        |  $d_1[u] \leftarrow km$  // Update the browsed
        |   distance
    end
    if  $u$  has no outgoing bikeable and no unbikeable roads then
        |  $d_1[u] \leftarrow -1$  // Mark the it as an end node
    end
    for outgoing bikeable edge  $e_i = (u, v_i)$  with length  $l_i$  do
        |  $d_1, d_2 = \text{Downstream Dfs}(v_i, e_i, km + l_i, visited + [u],$ 
        |    $depth + 1, d_1, d_2)$ 
    end
    return  $d_1, d_2$ 

```

ALGORITHM C2 Clean candidates

```

Input :  $\mathcal{L}$  // All expanded nodes that are
        connected to bikeable roads
         $G_b(V_b, E_b)$  // Directed graph composed
        of only bikeable roads
         $G_f(V_f, E_f)$  // Full graph where all
        roads have two directions, sharing the
        same edge label of OSM
         $\sigma$  // Detour threshold
Output:  $lst$  // Label list of edges that should
        not be categorized as of low
        availability
         $lst = []$  // Initialize output
for  $u \in \mathcal{L}$  do
    for  $v \in \mathcal{L}$  do
        if  $u = v$  then
            | continue
        end
         $s_1, \mathcal{B}^f(u, v) \leftarrow$  shortest path list between  $u$  and  $v$  in
         $G_f$  and its distance
        if  $u$  and  $v$  are disconnected in  $G_b$  then
            | if  $e = (u, v) \in E_f$  then
            |   |  $lst.append(e.label)$ 
            | end
            | continue
        end
         $s_2, \mathcal{B}^b(u, v) \leftarrow$  shortest path list between  $u$  and  $v$  in
         $G_b$  and its distance
        // When a significant detour appears,
        // we consider the bikeable network as
        // interrupted
        if  $\mathcal{B}^b(u, v) / \mathcal{B}^f(u, v) \geq \sigma$  then
            | for  $e \in s_1$  do
            |   |  $lst.append(e.label)$ 
            | end
            | break
        end
    end
end
end

```

ALGORITHM C3 Assess candidate nodes.

```

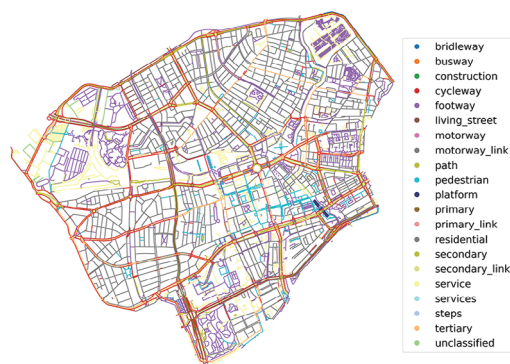
Input :  $e_i$  // Considered edge
 $d_1^+$  // Browsed distance dictionary for downstream nodes
 $d_2^+$  // Browsed path dictionary for downstream edges
 $lst^+$  // List of bad expanded downstream nodes
 $d_1^-$  // Browsed distance dictionary for upstream nodes
 $d_2^-$  // Browsed path dictionary for upstream nodes
 $lst^-$  // List of bad expanded upstream nodes
 $G_b(V_b, E_b)$  // Directed graph only composed of bikeable roads
 $\mathcal{R}$  // A set of rules

Output:  $lst$  // Confirmed edges

for  $u \in d_1^+.keys()$  do
   $km_1 \leftarrow d_1^+[u]$ 
  for  $v \in d_1^-.keys()$  do
     $km_2 \leftarrow d_1^-[v]$ 
    if  $u = v$  then
      | continue
    end
    if  $u$  and  $v$  are disconnected in  $G_b$  then
      | // If the bikeable road network is not connected
      | // We could still keep the expanded upstream and downstream nodes
      | // If they are at the frontier of the road network
      | if  $d_1^+[u] = -1$  or  $d_1^-[v] = -1$  and  $v$  not excluded nodes then
      | |  $lst = lst + d_2^+[u] + d_2^-[v]$ 
      | end
      | continue
    end
     $s_1, B \leftarrow$  shortest path list between  $u$  and  $v$  in  $G_b$  and the distance
    // The sum of extended distance and the length of original road
     $\mathcal{U} \leftarrow d_1^+[u] + d_1^-[v] + l_i$ 
    if  $\mathcal{R}(\mathcal{U}, B)$  is satisfied then
      for  $i$  in  $d_2^+[u]$  do
        if  $i \in lst^+$  then
          | break
        end
         $lst.append(i)$ 
      end
      for  $i$  in  $d_2^-[v]$  do
        if  $i \in lst^-$  then
          | break
        end
         $lst.append(i)$ 
      end
    end
  end
end

```

APPENDIX D: DATASET DESCRIPTION

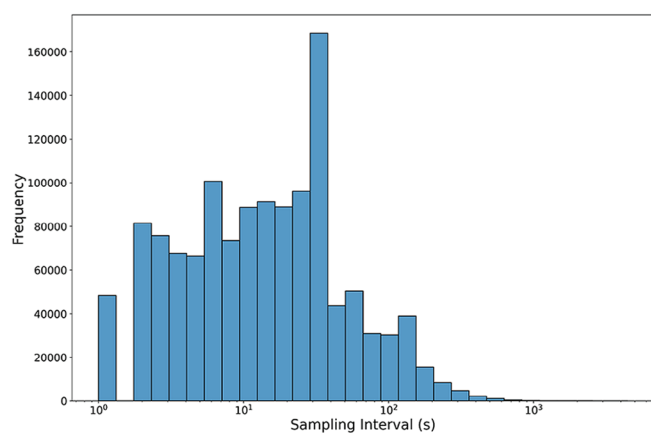


(a) OpenStreetMap of study area in Rotterdam.

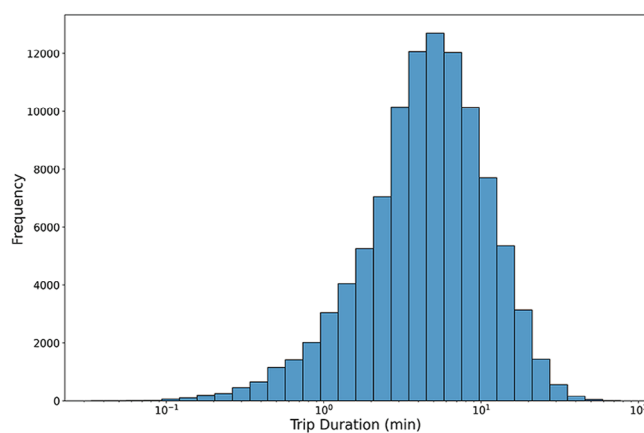


(b) Talking bike GPS data point distribution in Rotterdam. The absence of recorded trips in the North West can be attributed to this zone being a zoo and private gardens.

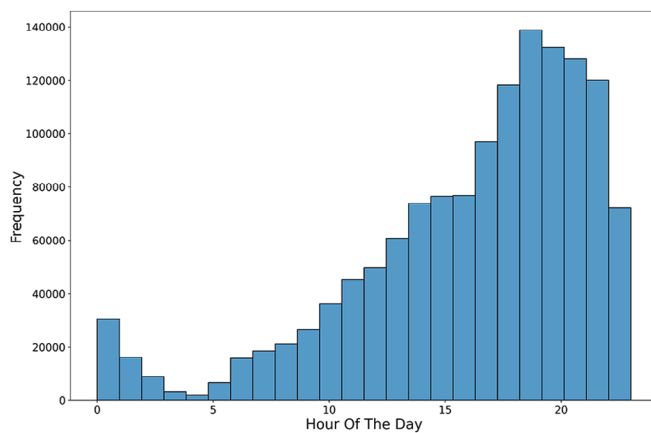
FIGURE D1 Study area.



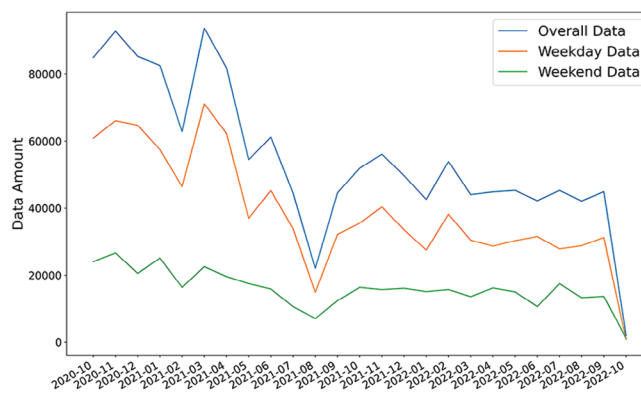
(a) Sampling interval (seconds) distribution.



(b) Trip duration (min) distribution.



(c) Sampling hour distribution.



(d) Evolution of GPS data amount.

FIGURE D2 Statistical property of talking bike data.

APPENDIX E: CROSS-DATASET ANALYSIS

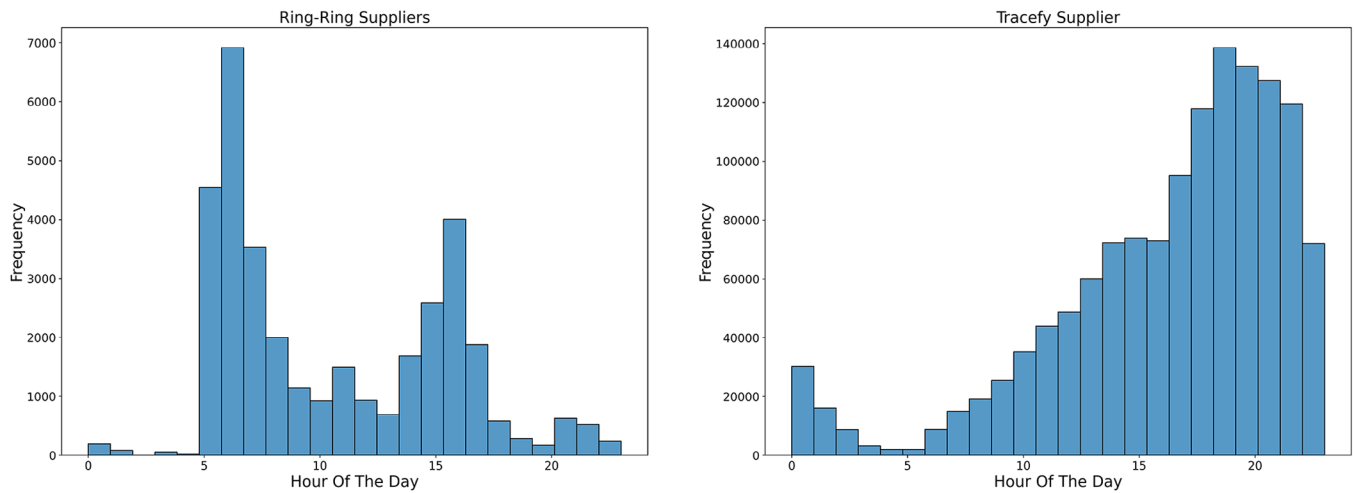


FIGURE E1 Distribution of sampling hour of the day for supplier Ring-Ring and Tracefy.

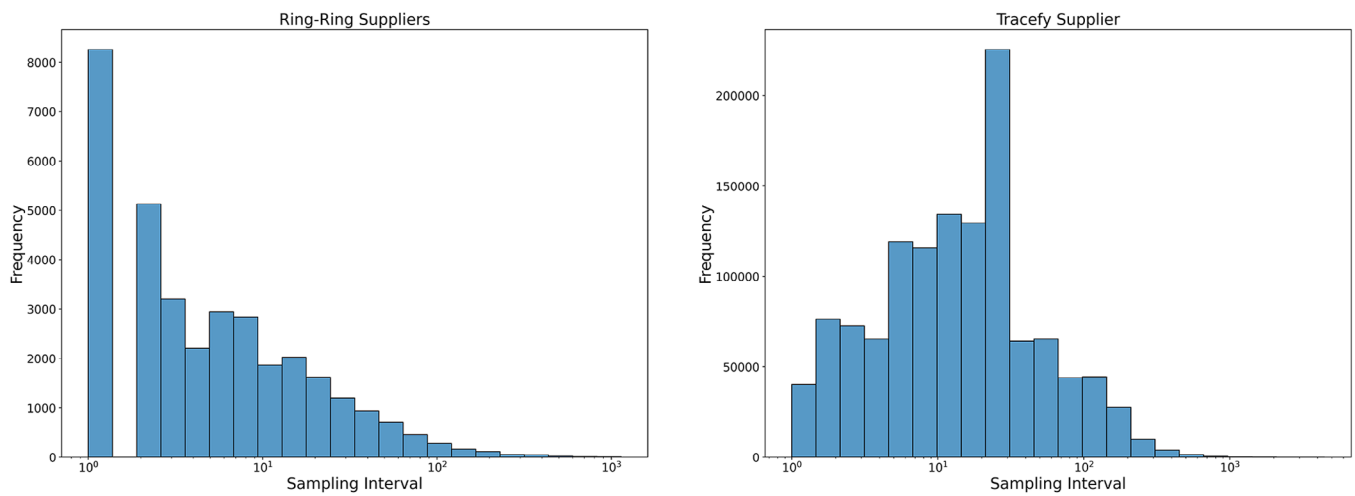


FIGURE E2 Distribution of sampling interval for supplier Ring-Ring and Tracefy.

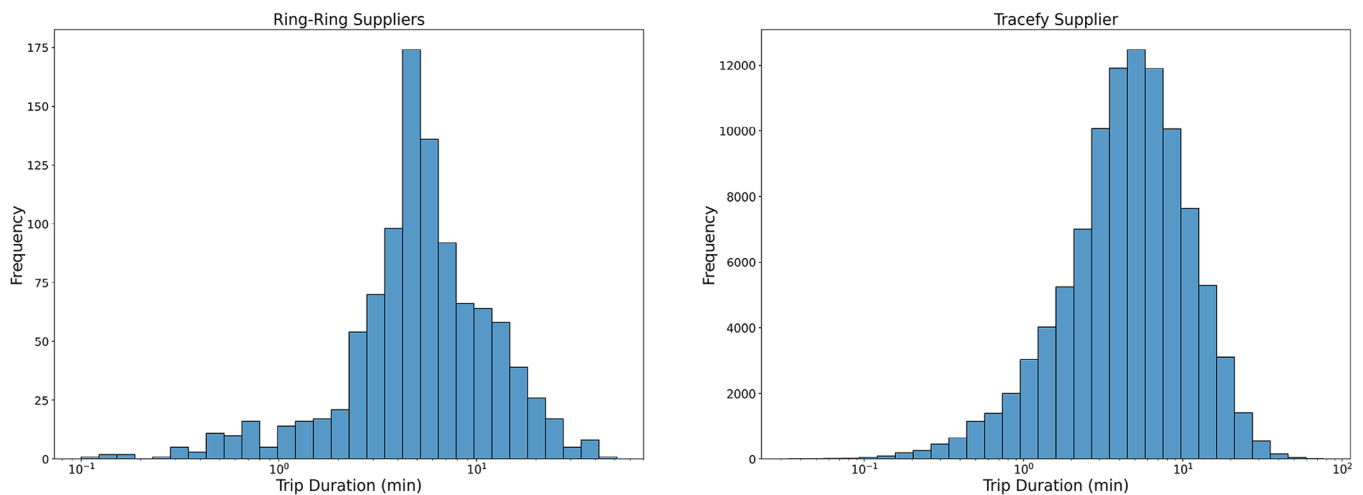


FIGURE E3 Distribution of trip duration for supplier Ring-Ring and Tracefy.

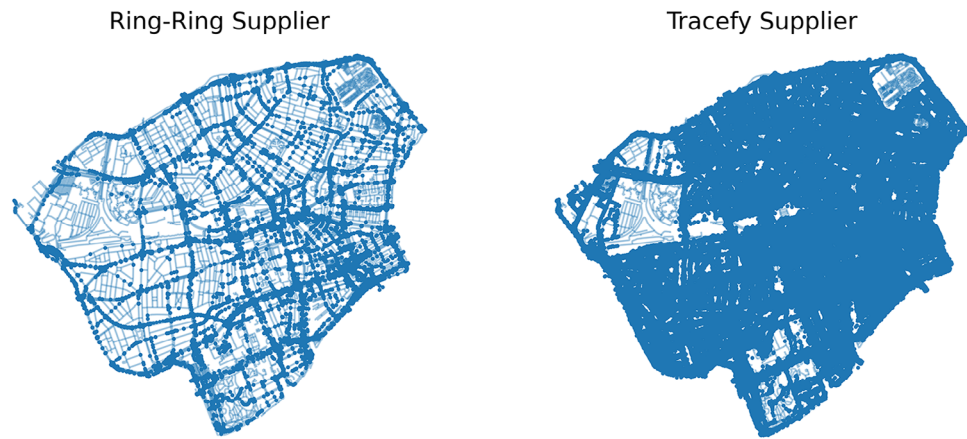


FIGURE E4 Distribution of data samples for different suppliers, Ring-Ring data are more centered in main roads.



FIGURE E5 Heatmap based on bicycle counting amount of mapped data from various suppliers. Due to the large quantity of Tracefy data, it is represented on a logarithmic scale for better visualization.

APPENDIX F: EVALUATION DATASET

TABLE F1 p-value of Kolmogorov–Smirnov test comparing the evaluation dataset and the raw dataset.

	Sample time	Trip-averaged sample time	Travel time	Nb of trip segments
p-value	0.74	0.65	0.94	0.99

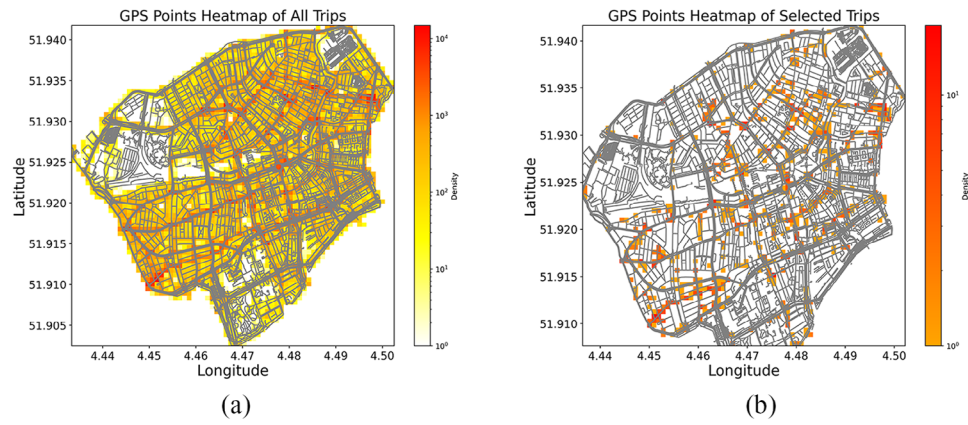


FIGURE F1 GPS point distribution heatmap of (a) all trips and (b) selected trips for evaluation.

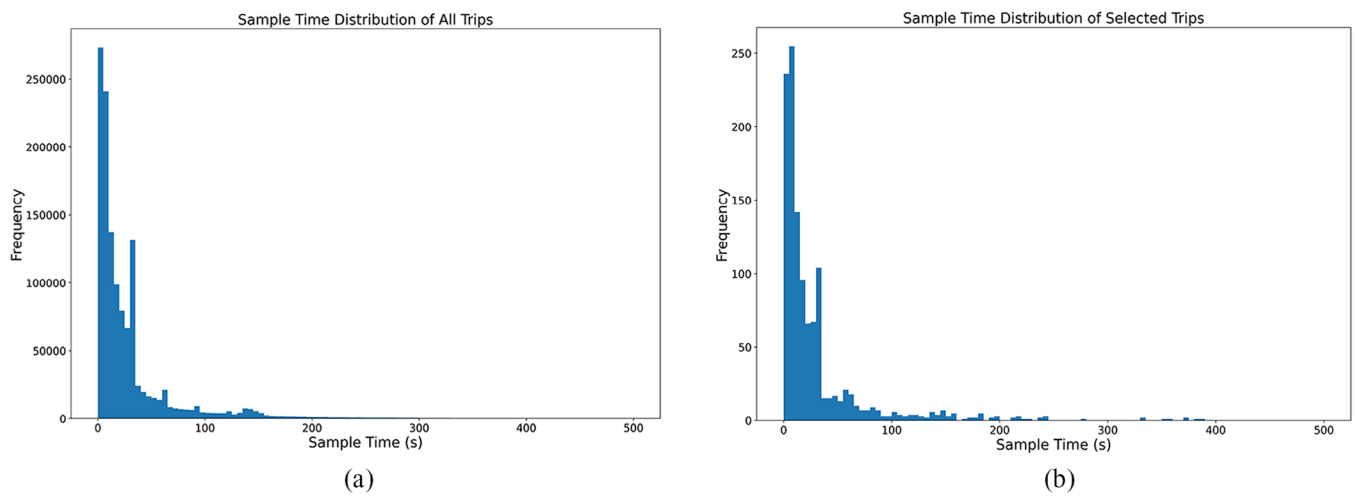


FIGURE F2 Sample time distribution of (a) all trips and (b) selected trips for evaluation.

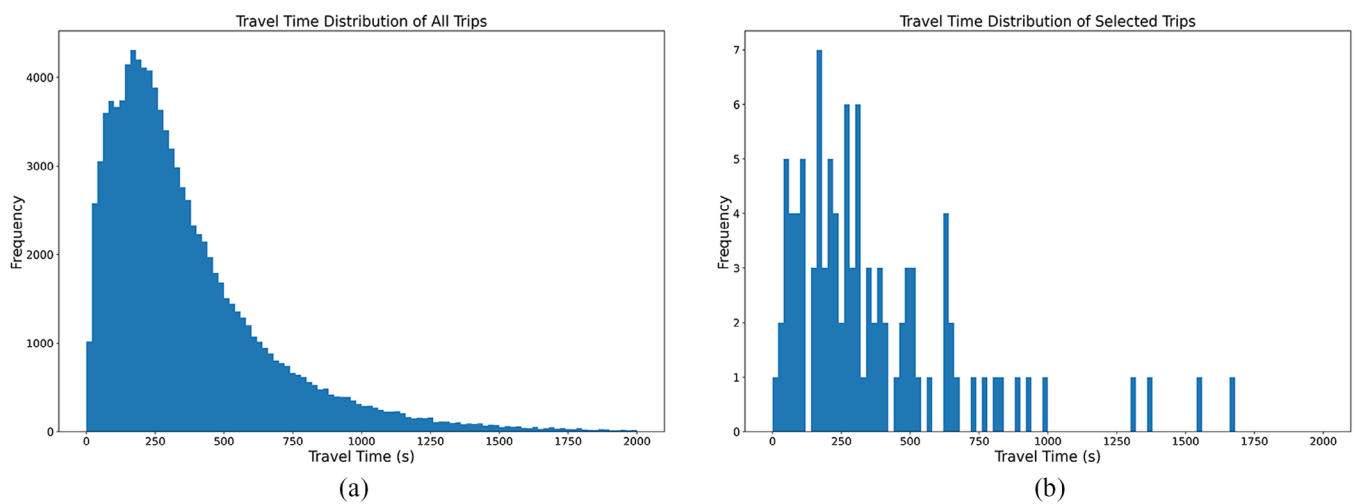


FIGURE F3 Travel time distribution of (a) all trips and (b) selected trips for evaluation.

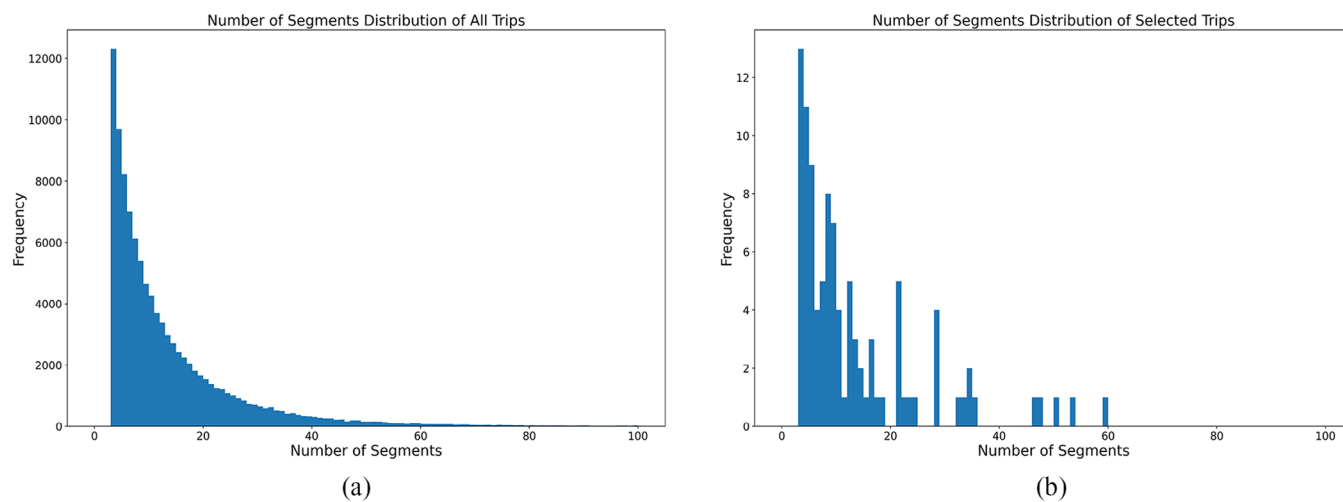


FIGURE F4 Number of trip segment distribution of (a) all trips and (b) selected trips for evaluation.