Journal of Traffic and Transportation Engineering(English Edition)

交通运输工程学报(英文版)

ISSN 2095-7564,CN 61-1494/U

《Journal of Traffic and Transportation Engineering(English Edition)》网络首发论文

题目：	Review and prospect of floating car data research in transportation
作者：	Chi Zhang，Yuming Zhou，Min Zhang，Bo Wang，Yuhan Nie
网络首发日期：	2024-10-25
引用格式：	Chi Zhang，Yuming Zhou，Min Zhang，Bo Wang，Yuhan Nie．Review and prospect of floating car data research in transportation[J/OL]．Journal of Traffic and Transportation Engineering(English Edition). https://link.cnki.net/urlid/61.1494.u.20241024.1144.002

Published online：2024-10-25 11:35:20
URL：https://link.cnki.net/urlid/61.1494.u.20241024.1144.002

Review Article

Review and prospect of floating car data research in transportation

Chi Zhang[a, *], Yuming Zhou[a], Min Zhang[b], Bo Wang[a], Yuhan Nie[b]

[a] *School of Highway, Chang'an University, Xi'an 710064, China*

[b] *College of Transportation Engineering, Chang'an University, Xi'an 710064, China*

Highlights

- Bibliometric analysis of floating car data research in the transportation field was performed using CiteSpace technology.
- Literature publication, keyword burst detection, co-occurrence, and cluster analysis were conducted.
- High-frequency floating car data can be used to model continuous operating speed prediction for heavy trucks.
- Transformer and Graph Neural Networks may capture the spatio-temporal patterns of floating car data.

Abstract

With the advancement of Intelligent Transportation Systems, floating car data, as a crucial source of transportation information, has garnered increasing attention for its applications and development directions within the context of massive traffic data. This study conducts an in-depth literature review analysis of floating car data in the transportation field based on the Web of Science (WOS) database from 2000 to 2023, employing bibliometric methods and knowledge graph technologies. The current research status was visually analyzed through the literature distribution by year, research regions and institutions, research hotspots, and literature clustering using the bibliometric tool CiteSpace. Three major research topics were identified based on the literature clustering analysis. A systematic review of key literature was conducted to address research challenges related to floating car sampling proportions and frequencies, and future research challenges and opportunities were proposed. The result shows an overall parabolic increase in publication volume, with research hotspots mainly focusing on mountainous cities, cluster analysis, machine learning, and deep learning. The three major research clusters include traffic flow state, traffic safety, and route planning. The optimal investment proportion for floating cars is determined to be 3-8%, and the sampling frequency significantly affects the accuracy of vehicle speed and heading angle information, while having a weaker impact on positional parameters. With the trend of large-scale Internet-connected vehicle deployment in the future, a massive amount of floating car data will be generated, prompting in-depth research on the fusion of heterogeneous data sources, including floating car data. Future research could focus on leveraging Transformer and Graph Neural Networks to explore spatiotemporal features of data, developing lightweight real-time floating car data processing algorithms, and constructing multimodal refined models tailored to specific traffic scenarios.

Keywords:

Transportation engineering; Floating car data application; Bibliometric analysis; Freeway; Floating car proportion; Sampling frequency

*Corresponding author.
E-mail addresses: zhangchi@chd.edu.cn (C. Zhang), 17502968661@163.com (Y. Zhou), minzhang@chd.edu.cn (M. Zhang), wb1010110wb@chd.edu.cn (B. Wang), 2024034010@chd.edu.cn (Y. Nie).

1 Introduction

With the rapid development of economies and societies globally, the transportation industry, as a fundamental and leading sector of national economies worldwide, is facing unprecedented opportunities and challenges. However, with the continuous growth in travel demand, concerns regarding traffic safety and efficiency are becoming more pronounced, presenting significant challenges to the sustainable advancement of the transportation sector on a global scale. Intelligent Transportation Systems (ITS), as effective means to enhance traffic management and address issues of safety and efficiency, have achieved significant success internationally. The progress and innovation of ITS heavily rely on real-time, massive, and multi-source traffic data support. In recent years, with advancements in traffic detection technology, modern data collection methods such as loop detectors, surveillance cameras, and floating cars have gradually become important data sources for intelligent transportation systems globally. These advanced data collection technologies provide solid data support for real-time monitoring and analysis of traffic conditions, traffic flow optimization, road capacity improvement, and ensuring traffic safety. In particular, floating car data, with its strengths of high dynamism, wide coverage, and large information volume, has become an indispensable part of intelligent transportation systems worldwide. Through in-depth analysis and exploration of floating car data, the refinement of traffic management can be effectively enhanced, providing a scientific basis for traffic planning, road design, emergency response, and more. Therefore, strengthening the collection, processing, and analysis of floating car data is of significant importance in driving the development of intelligent transportation systems globally and elevating the overall level of the transportation industry on an international scale.

Surveillance cameras, loop detectors, and gantry systems serve as stationary data collection tools. Their fixed locations present challenges in capturing comprehensive traffic data over extended road segments or networks, and they may miss traffic incidents between two fixed detectors. Additionally, these fixed detectors are usually exposed to outdoor elements, making them vulnerable to damage from diverse weather conditions and environmental factors. The substantial maintenance expenses linked to frequent repairs contribute to the high overall cost of maintaining fixed detectors(Darwish and Abu Bakar, 2015). Unmanned aerial vehicle (UAV) technology holds significant potential for advancement in road traffic data collection. It can overcome spatial distance limitations without affecting road traffic operations, enabling the acquisition of macroscopic traffic data such as density, flow, and speed. Nevertheless, UAVs are vulnerable to adverse weather conditions and may encounter challenges in obtaining precise microscopic traffic flow data, often due to issues like video distortion.

Floating cars are spatial dynamic data collectors capable of providing real-time high-precision spatiotemporal traffic data(Del Serrone et al., 2023). These data collection devices are immune to adverse weather conditions and have relatively low maintenance costs. Equipped with satellite positioning and wireless communication, a fleet of vehicles gathers spatial positions, vehicle speeds, vehicle IDs, time stamps, and heading angles information, transmitting vast amounts of real-time data to the central control room. This data serves as a vital information source for intelligent transportation systems. Widely used in the transportation sector, floating car data offers high precision and spatial resolution. Literature highlights various applications of floating car technology, including traffic state prediction (Mena-Oreja and Gozalvez, 2021), travel time estimation(Li et al., 2022), abnormal event detection(Pascale et al., 2015), fleet management(Dabbas et al., 2021), among others. With the widespread use of floating car data in the transportation field, two main issues have emerged: the lack of uniform standards for utilizing floating car data, leading to heterogeneity in attributes such as sampling proportion and sampling frequency, affecting data quality and model generalization. The absence of studies using bibliometrics and knowledge graphs to systematically summarize the primary theoretical research on floating car data in transportation, specifying identifying the transportation issues best suited for such data. It is essential to conduct thorough research and systematic analysis of the theoretical underpinnings of floating car technology in transportation from a data application perspective. This involves summarizing the relationship between heterogeneous attributes of floating car data and model effectiveness, exploring theoretical advancements in floating car data within the transportation domain, and their applicability to transportation challenges. This process is crucial for delineating the future trajectory of development.

This paper provides a systematic review of the research progress of floating car data in transportation from 2000 to 2023. The article meticulously examines the major theoretical innovations and research accomplishments in this domain and, in consideration of the distinctive characteristics of floating car data, delineates the corresponding research directions. Furthermore, it rigorously explores the standardization of the discussion on the fundamental attributes of floating car data - sampling ratio and sampling frequency. Ultimately, the paper presents a forward-looking perspective on the challenges and opportunities that research on floating car data may encounter in the transportation field in the future.

Although there are existing review articles on floating car data, certain distinguishing characteristics justify the inclusion of this review in the literature on floating car data:

1. Firstly, this review specifically focuses on the narrower research scope of floating car data, allowing for a more detailed systematic review of literature from the past 20 years. In contrast, other similar literature tends to emphasize "big data," leading to more generalized reviews that may lack detailed descriptions, with the literature analysis usually covering the past 5 or 10 years(Lian et al., 2020; Neilson et al., 2019).

2. This study also addresses a significant deficiency in previous research by providing the latest analyses and incorporating recent advancements. It also presents the current research challenges and constructive solutions. In contrast, other literature tends to focus more on identifying research challenges with fewer constructive solutions offered.

3. A notable feature of this review, unlike previous reviews, is the adoption of bibliometric analysis techniques, including journal analysis, co-occurrence analysis, and clustering analysis. It also includes analyses of publication distribution by year, research regions and institutions, and the importance of keywords. Such analyses provide readers with a macroscopic understanding of the research field before delving into a systematic literature review. The latter part of the review follows a more traditional literature review approach.

The rest of the paper is organized as follows. In Section 2, the study outlines the data sources and research methods employed. Section 3 showcases the annual distribution of literature, research regions, and research hotspots. Section 4, the paper summarizes the clustering outcomes of the reviewed articles, including the new theories and conclusions within each cluster. Section 5. provides a concise analysis of the fundamental attributes of floating car data, particularly focusing on sampling ratio and frequency. Section 6 delves into future research challenges and opportunities. Finally, section 7 succinctly concludes the paper with overarching conclusions from the reviewed studies.

2 Data sources and research methods

This section introduces the sources, quantity, and time span of the literature, and provides the principles and methods of co-citation analysis, burst analysis, and clustering analysis.

2.1 Data sources

The literature was sourced from the Web of Science (WOS) Core Collection database. The bibliometric analysis method is only compatible with the WOS database. In order to accommodate the limitations of the methodology, we chose the WOS database for our analysis. However, the cluster analysis in Chapter 4 incorporated significant literature from Science Direct and IEEE Xplore databases to avoid overlooking important literature. The search query used was based on the topic = "Floating car data" OR "Floating car technology" OR "Probe vehicle data" OR "FCD", limited to "Research Areas = Engineering or Transportation", with a search period from 2000 to 2023, and selecting document types as "Article OR Proceeding Paper OR Review Article", resulting in a total of 769 relevant articles. The reason the paper did not consider the related keywords "GPS data" and "connected vehicle data" in our search formula is that adding these two keywords would significantly increase the number of retrieved literature to 3,233 articles. When there is an overwhelming amount of literature, it tends to lead to a more macroscopic and generalized analysis, which may hinder a detailed examination of the literature. Therefore, we opted to focus solely on research within the scope of "floating car data".

2.2 Research methods

CiteSpace is a citation visualization analysis software that has gradually developed in the context of scientometrics and data visualization. Its working principle involves placing the entire research field in a three-dimensional citation space, visualizing scientific knowledge structures, patterns, and distributions through visual means, known as a scientific knowledge graph. Its main analysis methods include co-citation analysis, burst analysis, and clustering analysis.

2.2.1. Co-citation analysis

Co-citation analysis originates from the citation network, which is generated from the mutual references between papers. Scholars often cite previous literature in their research process, and when two articles are cited together, a connection line is added between them, forming a citation network where multiple documents refer to each other. Co-citation measures the similarity between two papers based on the number of times they are co-cited by other papers. This relationship strengthens as the number of citations increases. When co-citation analysis is applied to keywords, a keyword co-occurrence network is formed, where keywords act as nodes in the network, and the similarity between keywords is represented by the thickness of the edges. The formula for calculating the association strength is shown as follows:

$$\mathrm{Cosine}(x,y)=\frac{XY}{\|X\|\|Y\|}=C_xC_y\Big/\sqrt{(\sum_{i=1}C_{x_i}{}^2)(\sum_{i=1}C_{y_i}{}^2)}$$

2.2.2. Burst analysis

CiteSpace offers the "Burst Detection" function to identify keywords with significant changes in citation frequency over a specific period. This feature is used to analyze the rise and fall of keywords, determining research hotspots in different time frames.

2.2.3. Cluster analysis

CiteSpace utilizes the Modularity (Q value) and Weighted Mean Silhouette (S value) to jointly characterize the overall clustering effectiveness based on network structure and clustering clarity. Typically, a Q value greater than 0.3 indicates a significant clustering result, while an S value greater than 0.5 signifies reasonable clustering clusters and an S value exceeding 0.7 indicates excellent clustering effectiveness.

3 Statistical analysis

3.1 Annual distribution of literature

The literature distribution by year is illustrated in Fig. 1. The overall publications show a parabolic upward trend. The years 2015 and 2019 represent two peak periods in the publication. The first peak may be attributed to advancements in communication technology, while the second peak could be driven by the rise of data fusion technologies, leading to the widespread application of floating car data. To further analyze the developmental context of the research, the time-zone analysis in Citespace was utilized to arrange keywords in chronological order. Based on the volume of publications and research year, the study of floating car data applications can be roughly categorized into the stages of initiation, rapid development, and steady development.

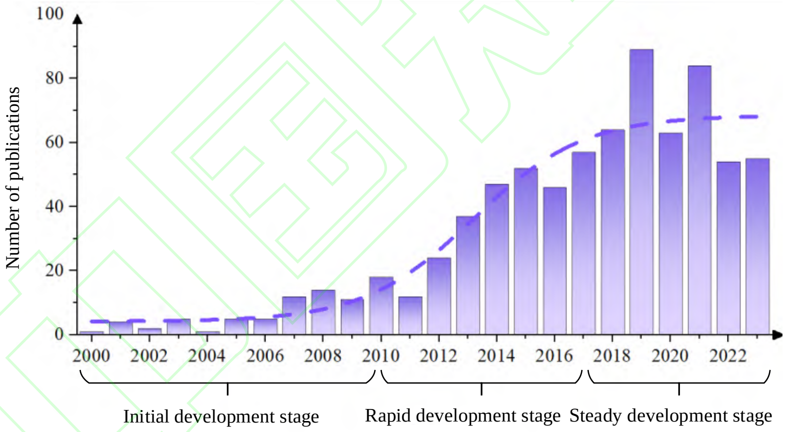


Fig. 1 Yearly distribution of literature.

Initial development stage (2000 to 2010). Floating car data was initially applied in the field of transportation in the year 2000, and by 2010, there were relatively few research outcomes, with an average annual publication volume of only 10 papers, indicating the stage of initiation. Research topics during this phase included traffic dynamic models, driver behavior, traffic OD matrix prediction, studies on data fusion technologies considering floating car data, further breakthroughs in global positioning system (GPS) technology, and initial attempts and applications of intelligent transportation systems. Floating car data was not widely utilized during this stage, and positioning technologies related to floating cars were still in the developmental phase.

Rapid development stage (2011 to 2017). With an average annual publication volume of only 30 papers, this stage represents a period of rapid development. The primary focus during this phase was on the core technology of map matching. Other research areas included traffic flow speed prediction for urban traffic networks and highways, traffic wave propagation modeling, and travel time prediction. Research methods tended to lean towards utilizing big data-driven clustering algorithms to mine floating car data.

Steady development stage (2018 to present). With an average annual publication volume of 60 papers, this stage represents a period of steady development. Research directions in this phase show a trend towards diversified development. High-precision extended floating car data has emerged, which can be generated in real-time by connected vehicles. Research topics include traffic flow state prediction, macroscopic fundamental diagram estimation, trajectory data reconstruction, capacity estimation, travel time prediction, queue length estimation, congestion pattern recognition, traffic safety studies, and reliability research. The research scope encompasses both path-level and road network-level roads. Research methods include machine learning and deep learning, with commonly used algorithms being convolutional neural networks and generative adversarial networks.

3.2 Research areas and institutions

The distribution of countries and institutions is shown in Table 1. The table presents the top 15 countries based on publication volume, with the percentage representing the proportion of publications from each country to the total. It can be observed that due to national policies and the advantage in the number of researchers, China and the United States are the largest contributors, with research output percentages of 28% and 21%, respectively. Following closely are Germany, Italy, and Japan, with percentages of 12%, 7%, and 6%, respectively. Additionally, the Netherlands, Belgium, France, Spain, Sweden, Austria, Canada, India, South Korea, and Australia each account for 2%-4% of the total publication volume.

In addition, the main research institutions in each country were also analyzed. The internal institutions within each country are ranked from most to least prominent. In China, the primary research institutions include Beijing Jiao tong University, Tongji University, Tsinghua University, Wuhan University, and

Chang'an University. In the United States, a key player in international competition, prominent institutions include Florida State University, University of California, Purdue University, and Virginia Tech. In conclusion, research on floating car data in the transportation field is widespread globally, with numerous internationally renowned research institutions involved. This indicates that the related research is valued and focused on from an international perspective, leading to the production of abundant research outcomes.

Table 1 Country and organization distribution of the FCD application literature.

Index	Flag	Country	Publication (%)	Main research institutions
1	🇨🇳	China	28	Beijing Jiao tong University, Tongji University, Tsinghua University, Wuhan University, Chang'an University
2	🇺🇸	United States	21	Florida State University, University of California, Purdue University, Virginia Tech
3	🇩🇪	Germany	12	Technical University of Munich, Helmholtz Association, German Aerospace Center, University of Duisburg-Essen, Brunswick University of Technology
4	🇮🇹	Italy	7	Polytechnic University of Milan, Sapienza University of Rome
5	🇯🇵	Japan	6	University of Tokyo, Tokyo Institute of Technology, Nagoya University
6	🇦🇹	Austria	4	Austrian Institute of Technology, University of Salzburg
7	🇫🇷	France	4	Grenoble Alpes University, École Normale Supérieure Paris
8	🇨🇦	Canada	3	University of Waterloo, University of Alberta, University of Calgary, University of Toronto
9	🇳🇱	Netherlands	3	Delft University of Technology
10	🇮🇳	India	2	Indian Institute of Technology, Indiana Department of Transportation
11	🇰🇷	South Korea	2	Seoul National University
12	🇧🇪	Belgium	2	Ghent University
13	🇸🇪	Sweden	2	Royal Institute of Technology
14	🇪🇸	Spain	2	Polytechnic University of Madrid
15	🇦🇺	Australia	2	University of Queensland, University of New South Wales

3.3 Keyword importance

The higher the frequency of a keyword, the more publications related to that keyword were published within the analyzed year, and high-frequency keywords can reflect the mainstream research directions. Centrality refers to the centrality of a keyword in the co-occurrence network. A higher value indicates that the keyword is directly connected to more other keywords, implying a higher importance of the keyword in the co-word network. The top ten keywords by frequency and centrality from the keyword co-occurrence analysis are presented in Table 2. The numerical values in the last column in parentheses indicate the ranking based on centrality.

Table 2 Top 10 High-Frequency keywords in Web of Science.

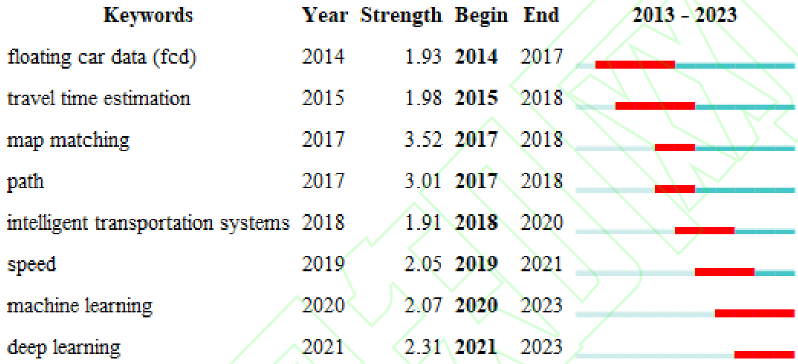
Index	Keyword	Frequency	Centrality
1	floating car data	114	0.18（2）
2	prediction	23	0.07（6）
3	big data	15	0.01（7）
4	traffic flow	14	0.16（3）
5	machine learning	9	0.00（8）
6	car sharing	9	0.27（1）
7	speed	8	0.07（6）
8	neural network	7	0.01（7）
9	map matching	7	0.11（4）
10	travel time estimation	7	0.09（5）

It can be observed that the keywords with higher frequencies include prediction, big data, traffic flow, machine learning, vehicle sharing, speed, neural networks, map matching, and travel time prediction. It is reasonable to infer that mainstream research often adopts big data-driven machine learning and deep learning methods, focusing primarily on predictive issues related to speed, traffic flow, and travel time(Winfrey et al., 2023).

There are significant changes in centrality ranking compared to frequency ranking. It is noteworthy that the centrality ranking of vehicle sharing has surged to first place, indicating that recent mainstream research often integrates the concept of vehicle sharing. Following that are traffic flow, map matching, travel time prediction, speed, big data, neural networks, and machine learning.

3.4 Keywords popping up

Keyword importance analysis focuses on the cumulative frequency during the research period, which cannot illustrate the changes over the years. The latter can reflect the differences in research hotspots in different years. To analyze the research hotspots of floating car data in different years, this study utilizes the "Burst detection" function of CiteSpace to detect the rise and fall of keywords from 2010 to 2023. The keyword burst function captures the moments of significant increase and decline in keyword frequency in different years to determine the research hotspots in different periods. By sorting the starting year of keyword bursts from the furthest to the nearest, the keyword burst results are presented in Fig. 2.



Keywords	Year	Strength	Begin	End	2013 - 2023
floating car data (fcd)	2014	1.93	**2014**	2017	
travel time estimation	2015	1.98	**2015**	2018	
map matching	2017	3.52	**2017**	2018	
path	2017	3.01	**2017**	2018	
intelligent transportation systems	2018	1.91	**2018**	2020	
speed	2019	2.05	**2019**	2021	
machine learning	2020	2.07	**2020**	2023	
deep learning	2021	2.31	**2021**	2023	

Fig. 2 Keywords popping up in the Top 8 of Web of Science.

It can be observed that the research hotspots before 2018 were travel time prediction, map matching, route planning, and intelligent transportation systems. After 2018, speed, machine learning, and deep learning have become research hotspots. As of 2023, travel time prediction, and machine learning have remained research hotspots for a continuous period of three years, while other research hotspots disappeared within two years.

4 Cluster analysis

4.1 Keyword clustering results

To explore the mainstream research topics of floating car data, keyword clustering analysis was conducted to identify clusters of literature using the "Association strength" normalization method with a co-citation frequency threshold set at 5. The clustering results are shown in Fig. 3. It can be observed that the keywords are clustered into three categories, with keywords in each cluster sharing similar research themes. The research focuses on three main themes: traffic flow state, traffic safety, and route planning. A systematic review of the literature focused on the different clustering directions will follow, a process that includes further searches of the literature from the Science Direct, and IEEE Xplore databases, and the incorporation of some new literature and possibly the collection of some of the older literature. However, our ultimate goal is to minimize the risk of overlooking any relevant literature.

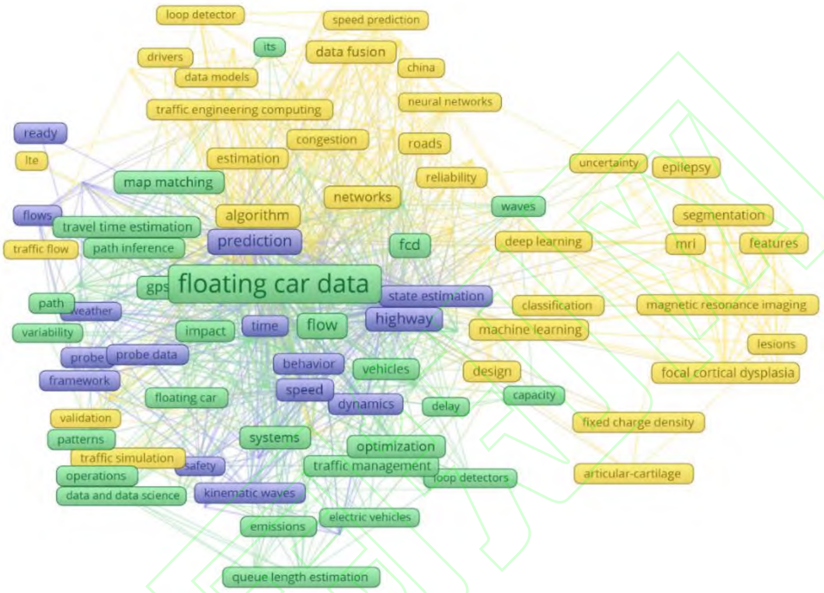


Fig. 3 Clustering results map.

4.2 *Traffic flow status (Cluster one)*

With technological advancements, the availability of floating car data continues to improve, encompassing not only historical trajectory data but also real-time data streams, laying the foundation for enhancing the level of traffic management services. Among various traffic flow management services, the prediction of traffic flow status is of significant importance for traffic planning and traffic safety research. While most studies rely on loop detectors to collect traffic data, this method poses challenges such as high costs, time-consuming processes, and limitations due to sensor deployment locations. In contrast, floating car data can effectively overcome these constraints. Subsequently, this paper will focus on introducing research on traffic flow status based on floating car data, covering four subtopics: traffic volume, traffic speed, traffic density, and macroscopic fundamental diagram. The main areas of research focus are detailed in Table 3.

Table 3 Research hotspots of floating vehicle data in the field of traffic state prediction.

Research topics	Related literature	Main research content
Traffic Volume	Sekula et al. (2018); Antoniou et al. (2016)	Estimation of historical traffic volume between sparse traffic sensors
Traffic Speed	Pavlyuk and Jackson (2022); Jiang et al. (2017); Rempe et al. (2022)	Fusion of multi-source heterogeneous data to enhance sparse data density, improvement of estimation accuracy through neural network algorithms
Traffic Density	Kyriacou et al. (2022); (Kyriacou et al. 2021)	Extension of floating car data combined with Bayesian paradigm and maximum a posteriori estimation to derive traffic density
Macroscopic Fundamental Diagram	Kong et al. (2018); Jiang et al. (2021)	Validation and improvement of macroscopic fundamental diagram through a fusion of floating car data and loop detector data

4.2.1 Traffic volume

Traffic volume estimation is an essential component of traffic management systems. Traditional road detectors, such as loop detectors and cameras, can directly detect traffic volume. However, these devices are susceptible to adverse weather conditions, leading to equipment wear and high maintenance costs. Floating car technology addresses this challenge by collecting data unaffected by adverse weather conditions, resulting in relatively lower maintenance costs. The primary approach to traffic volume estimation is based on estimated vehicle travel times. To achieve traffic volume prediction tasks, travel time data extracted from floating car data is often used as model input, while historical traffic volume data obtained from road detectors is used as output to establish a nonlinear relationship between the two (Li et al., 2019). Fig. 4 illustrates the process of obtaining travel times using Google Maps floating car data, obtaining traffic volume using road sensors, and constructing Gaussian Process Regression (GPR) models for both. Sekula et al. (2018) and (Zercher et al. 2024) studied the issue of estimating historical traffic flow between sparse traffic sensors and proposed a new method that combines neural networks with existing section methods, quantifying the value of estimating traffic volume using floating car data. Seo et al. (2015) developed a method to estimate flow, density, and speed based on detected vehicle data without requiring additional assumptions about traffic flow characteristics (e.g., fundamental diagram). Antoniou estimated OD flows in a real network.

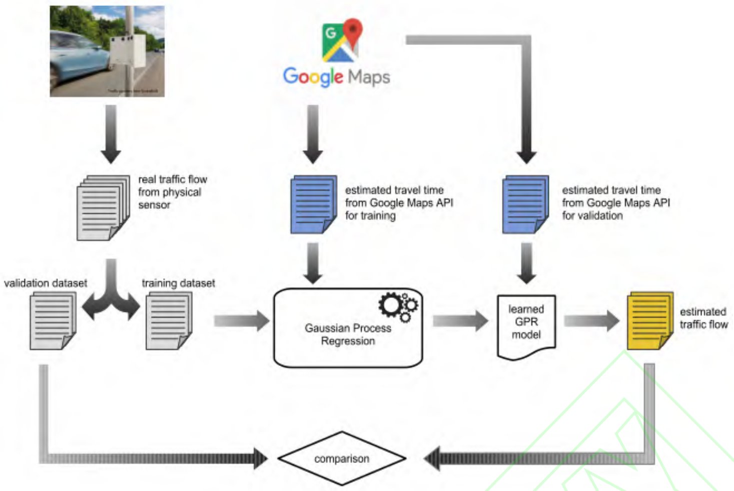

Google Maps
real traffic flow from physical sensor
estimated travel time from Google Maps API for training
estimated travel time from Google Maps API for validation
validation dataset
training dataset
Gaussian Process Regression
learned GPR model
estimated traffic flow
comparison

Fig. 4 Floating car data training process.

4.2.2 Traffic speed

Accurately predicting urban road traffic speed is of significant importance for the precise implementation of intelligent transportation systems The primary challenge in reconstructing traffic speed is the spatiotemporal sparsity of floating car data and the dynamic changes in penetration rate, which affect the accuracy of estimation. This challenge can be addressed from both a data perspective and a methodological perspective.

From a data perspective, floating car data belongs to the spatiotemporal two-dimensional data type. Existing fusional data types include temporal dimension data, spatial dimension data, and spatiotemporal dimension data. Among them, spatiotemporal dimension data is consistent with a floating car data type, and data fusion can yield better results. Visual traffic data collected by video cameras installed on floating cars falls under spatiotemporal data and can be used to enhance traditional FCD. Empirical results show that visually enhanced FCD significantly enhances the accuracy of traffic speed estimation(Pavlyuk and Jackson, 2022). Combining complementary information from multiple sources typically enhances accuracy, strengthens robustness, and reduces ambiguity. (Jiang et al., 2017) fused floating car data and loop detectors to reconstruct the speed field based on heterogeneous data. Deng et al. (2013) studied how to use multiple data sources (including loop detector counts, AVI Bluetooth travel time readings, and GPS location samples) to estimate the macroscopic traffic state of homogeneous highway sections. Although data assimilation techniques for reconstructing and predicting traffic states from multiple data sources have made significant progress, these methods are mostly data-driven and do not fully utilize the value of physical models (Nantes et al., 2016).

Improvements at the methodological level may further address the sparsity and dynamic penetration rate issues. Since floating car data may exhibit different penetration rates at different spatiotemporal locations, the spatiotemporal domain can be decomposed into smaller grid units for separate processing, where grid speed data and grid occupancy used as input, may be effective. Rempe et al. (2022) improved the architecture of a deep CNN U-net to address the sparsity of floating car data and investigated the problem of spatiotemporal highway traffic speed estimation based on deep convolutional neural networks. The improved network yielded good spatiotemporal distribution results of speed, as shown in Fig. 5. From the graph, it can be seen that the data of speed in the space-time range is continuous and complete, and no longer sparse. If it were still sparse, Fig. 5 would show a discrete speed-space-time graph. Hara et al. (2018) proposed a model for estimating unobserved highway path speeds using historically detected vehicle data through machine learning techniques. Zhang and Yang (2020) utilized regression machine learning algorithms to estimate traffic speed using floating car data and sensor detector data. Rempe et al. (2017) introduced a novel method for estimating highway speeds based on floating car data, which only requires speed data from floating cars without relying on other data inputs such as density or flow information.

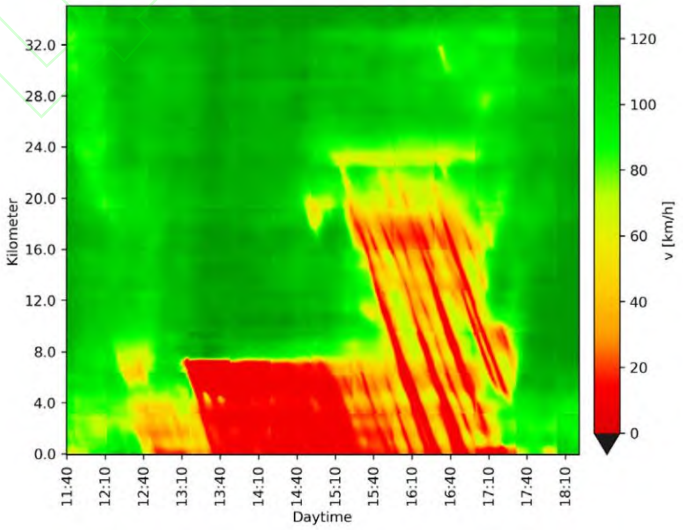


Fig. 5 Spatio-temporal speed distribution considering data sparsity (Rempe et al., 2022).

In the field of traffic speed estimation, floating car trajectory data is often combined with deep learning methods, utilizing recurrent or convolutional neural networks to fully explore the spatiotemporal two-dimensional features. The main challenge faced is the sparsity of spatiotemporal floating car data, prompting scholars to often use multi-source heterogeneous data fusion as model inputs to enhance data density. Further research is needed on how to accurately reconstruct speed with fewer data (Yoon et al., 2018).

4.2.3 Traffic density

The emergence of connected and autonomous vehicles (CAVs) has provided new possibilities for utilizing Extended Floating Car Data (XFCD) for traffic state estimation (Kyriacou et al. 2022). Compared to traditional floating car data, XFCD not only has higher road network coverage but also has the capability to real-time measuring the distance between floating cars and surrounding vehicles. The distance information is closely related to traffic density estimation, and XFCD combined with the Bayesian paradigm and Maximum A Posteriori estimation can derive traffic density (Kyriacou et al., 2021). Heshami and Kattan (2024) proposed a Bayesian traffic state estimation method for estimating traffic density based on XFCD. In addition, traffic density can also be derived from flow and speed information (Rodriguez-Vega et al., 2021). Predicting traffic volume based on mainly relies on travel time information in floating car data. When predicting traffic speed, research challenges such as data sparsity and dynamic penetration rate exist and can be improved through data fusion and spatiotemporal unitization methods. In addition to traditional reliance on speed and flow data calculations, predicting traffic density can also make full use of the vehicle spacing information in XFCD.

4.2.4 Macroscopic fundamental diagram

The Macroscopic Fundamental Diagram (MFD) is a graph that characterizes the relationship between traffic volume and density and is one of the important tools for evaluating traffic states (Carlos et al., 2007; Daganzo and Geroliminis, 2008). Floating car trajectory data, by providing detailed traffic volume and speed information, helps in fine-tuning the modeling of traffic flow operation states. Currently, using floating car trajectory data for MFD estimation is a common method.

Most studies combine FCD with Loop Detector Data (LDD) to validate and estimate the Macroscopic Fundamental Diagram (MFD). The required traffic volume is obtained from LDD, while traffic density and speed are obtained from FCD (Ji et al., 2018). Previous studies have conducted extensive work on validating MFD based on floating car data. Kong et al. (2018) proposed a method for estimating queue length based on different percentages of floating vehicles using a Back Propagation neural network. Based on estimating queue length, the relationship between average queue length and average flow, speed, and density at intersections was fitted, demonstrating the existence of MFD. Knoop et al., (2018) obtained aggregated data for one month in 2015 using Google Maps. They collected aggregated speed and volume every 5 minutes from 7 a.m. to 10 p.m., proving the existence, shape, and robustness of the Amsterdam MFD. Jiang et al. (2021) extracted speed from FCD collected from over 60,000 GPS-equipped taxis in Beijing and used 2-minute aggregated traffic volume from 44 RTMS detectors on the Beijing Ring Expressway to propose an improved fundamental diagram considering spatiotemporal changes in traffic volume and speed. The research showed that combining data can reduce MFD estimation errors to 0.04 (Ji et al., 2018). FCD is often used as one of the fusion data sources in MFD estimation research, complementing and correcting each other with non-homogeneous data, which can reduce errors caused by insufficient quality and accuracy, thereby improving the accuracy and reliability of MFD estimation.

4.3 Traffic safety (Cluster two)

Real-time prediction of collision risk can provide support for traffic accident management, offering important information for allocating resources to professionals to actively address anticipated traffic accidents. Understanding the relationship between traffic operating conditions and collision risk and further implementing safety measures is crucial. Speed differential is often considered to be associated with high collision risk. However, using sparse and intermittent loop sensors can often fail to capture continuous speed variations and detailed vehicle formations near collision locations. Microscopic and high-resolution floating car data can address this issue. The main research hotspots of this cluster are shown in Table 4.

Table 4 Hot research topics of floating car data in traffic anomalous event recognition and prediction topic.

Application directions	Research topics	Related literature	Main research content
Traffic Accidents	Traffic Accident Detection	Houbraken et al. (2017); Asakura et al. (2017); Jalali and Torfeh Nejad (2020); Yang et al. (2017)	Real-time monitoring of accident locations, prediction of congestion time and locations caused by accidents, and identification of secondary collisions.
	Traffic Accident Prediction	Fu et al. (2019); (Zhang et al. 2024; Yu et al. 2021)	Predicting collision risks on highways based on pre-collision traffic dynamics (such as average vehicle speed and speed reduction) and static properties of highways; identification of risky drivers.
Traffic Congestion	Traffic Congestion Identification	Sun et al. (2019); D'andre and Marcelloni (2017); Tang et al. (2016); Song et al. (2019); Laranjeiro et al. (2019)	Identification of traffic congestion using traffic performance index, speed, travel time, density, and queue length.
	Traffic Congestion Prediction	Houbraken et al. (2018); Erdelić et al. (2021)	Adaptive adjustment of traffic congestion prediction models based on the spatiotemporal dynamic characteristics of floating car data.

4.3.1 Traffic accidents

In recent years, substantial advancements have been made in traffic accident modeling. However, due to limitations in data collection methods, numerous studies have struggled to deliver reliable predictive outcomes (Oh et al., 2005; Lee et al., 2002). Floating car technology, as a real-time collision tendency data collection method, provides high data accuracy and wide coverage in space and time. Therefore, numerous scholars have endeavored to leverage it for traffic accident prediction.

Most studies use traffic data from the location of accidents to conduct research, but this data is often difficult to obtain. In addition, the traffic environment in which the accident vehicles are located before and during the accident can influence the accident outcome (Zabat et al., 1995). Therefore, filtering out floating cars that were in the same traffic environment as the accident vehicles before or during the accident and extracting their speed and other relevant information, such as speed differential ratio, average speed difference, speed fluctuation, and speed variance, to represent the traffic conditions at the accident scene (Zabat et al., 1995). Xu et al. (2019) used floating car data with a sampling frequency of 10 seconds to extract speed differential information during the morning peak hours on urban elevated highways. They derived the standard deviation of the Sectional Density-Constrained Speed Mean (SDCSM) and the Mean-Constrained Speed Standard Deviation (MCSSD) and developed a hierarchical and non-hierarchical Poisson-Gamma measurement error model to simulate collision frequency on highways. Yu et al. (2021) conducted collision risk analysis using discrete and comprehensive traffic data (such as loop detector data and detected vehicle data) and identified collision events as the prediction target. They used the Safety Surrogate Indicator with a Collision Correction Time (MTTC) of less than 2 seconds to identify high-risk events extracted traffic operation characteristics in the 5 seconds before the event occurred based on vehicle trajectory data, and developed three different logistic regression models. Zhang et al. (2024) utilized high-resolution real-time traffic speed data from crowdsourced floating cars on the Alabama highway network, employed machine learning models, and predicted collision risk on highways based on pre-collision traffic dynamics (such as average vehicle speed, and speed differential) and static highway properties.

4.3.2 Traffic congestion

Traffic congestion prediction involves utilizing historical data and current information to forecast the traffic congestion situation in a specific local area or the entire road network for a certain period, given the known or partially known traffic flow state. The automatic identification of traffic congestion status is an essential component of intelligent transportation systems and serves as a prerequisite for urban expressway monitoring and intelligent traffic control. Both domestic and international research has demonstrated that congestion on one or several roads is sufficient to cause congestion throughout the entire city road network. Therefore, identifying the bottleneck locations of these congested roads and implementing effective measures can gradually address the issue of traffic congestion across the entire road network.

Advanced traffic congestion estimation methods are mostly based on floating car data (Kan et al., 2019; Zhao and Hu, 2019), while some also utilize data from automatic license plate recognition systems (Shi et al., 2018; Zheng and Liu, 2017; Beliakov et al., 2018). Common indicators for discerning traffic congestion include Traffic Performance Index (TPI) (Sun et al., 2019), speed (D'andre and Marcelloni, 2017), travel time (Tang et al., 2016), density (Laranjeiro et al., 2019), and queue length (Song et al., 2019). Floating car data can directly provide high-precision information on speed, travel time, and queue length, making it highly favored in traffic congestion estimation.

Domestic and international scholars mainly utilize speed information extracted from floating car data to accurately identify and locate road congestion bottlenecks. By converting average speed data into traffic state information, the traffic bottleneck locations can be determined. Altintasi et al. (2017) collected Floating Car Data (FCD) at 1-minute intervals from a main urban road in Ankara, extracting average speed information from the FCD. They defined qualitative 4-scale state parameters based on average speed according to the service level of urban roads. Zhu et al. (2021) calculated the average travel speed based on travel time in floating car data and used it as an indicator for categorizing congestion levels. By incorporating fuzzy theory to address the issue of inaccurate speed thresholds, they established a congestion discrimination model. Zhang et al. (2014) using floating car data at a 5-minute time granularity, established 3-level discrimination indicators based on the spatiotemporal characteristics of recurrent congestion, including congestion thresholds, the ratio of congestion duration to frequent periods, and frequency of recurrent congestion. Zhang et al. (2018) utilized 5-minute aggregated floating car data from Beijing to extract speed differentials as congestion discrimination indicators to identify traffic bottlenecks, with speed-capacity ratio used as an auxiliary indicator to distinguish true bottlenecks from non-bottlenecks. The identified traffic bottlenecks are shown in Fig. 6. The longitudinal axis in the figure corresponds to the identification numbers of bottlenecked road segments, while the horizontal axis represents the time range during which traffic bottlenecks occur on the corresponding segments. It can be observed that the numerical values marked on the longitudinal axis represent the key road segments prone to traffic bottlenecks within the scope of the study. Segment 91 stands out as the road segment with the longest duration of traffic bottlenecks within a day, thus warranting significant attention from the transportation management authorities. These studies primarily predict traffic congestion using average speed information from floating car data, focusing on the static identification and evaluation of traffic congestion while neglecting the spatiotemporal heterogeneity of traffic congestion.

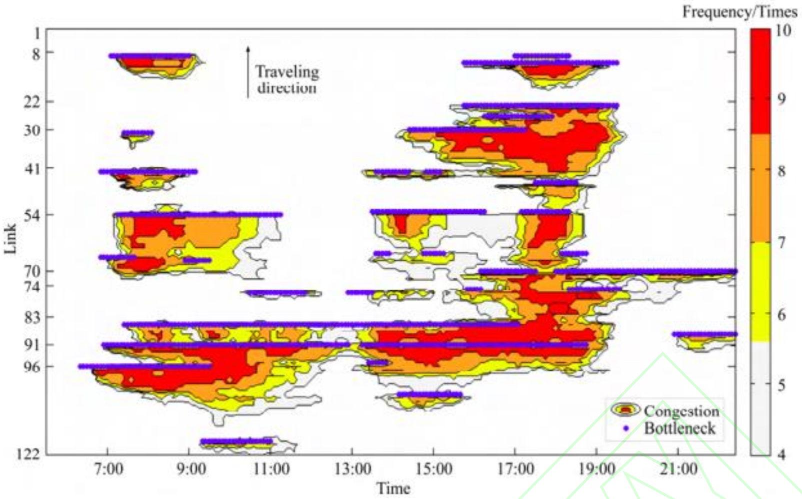


Fig. 6 Identification of spatio-temporal traffic bottlenecks in the third ring road of Beijing (Zhang et al., 2018).

In recent years, significant advancements have been made in spatiotemporal prediction of traffic congestion based on floating car trajectory data. To better consider the spatiotemporal characteristics of congestion models, it is necessary to construct spatiotemporal grid units. Some scholars, based on the matching of floating car data with spatiotemporal road network data, have introduced the principle of probability density segmentation to segment vehicle speeds in spatial and temporal distances, aiming to establish congestion models with adaptive adjustment capabilities (Chen et al., 2022). Prior to inputting the spatiotemporal grid of roads into the model, clustering processing is typically performed to aid in extracting data features for the model, thereby constructing a more accurate multimodal congestion model.

The spatiotemporal grid units constructed based on speed and travel time information derived from extensive historical floating car data enable the analysis of congestion patterns from both macro and micro perspectives. At the micro level, congestion segments can be delineated using speed profiles, whereas at the macro level, congestion zones are generally identified using travel time data. The identification of congestion zones based on travel time highlights the spatiotemporal heterogeneity of traffic congestion, indicating that congestion boundaries exhibit spatial variations across different periods, as depicted in Fig. 7.

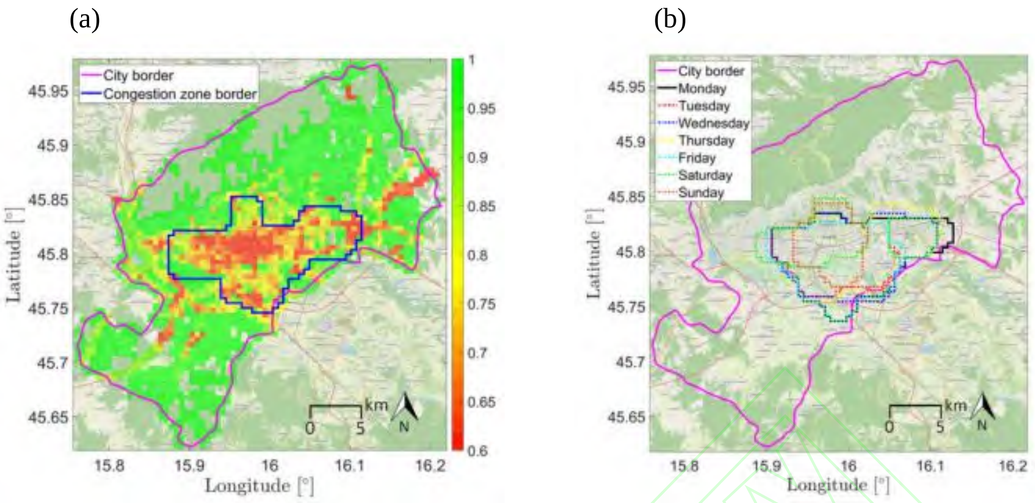


Fig. 7 Spatial characteristics of congestion and changes in congestion boundaries in the day-week dimension (Erdelić et al., 2021). (a) Congestion zone of the City of Zagreb-S. (b) Congestion zones per day in the week in the City of Zagreb-R.

The congestion propagation process at a finer time scale is illustrated in Fig. 8, where the study integrates FCD with LDD using speed as a congestion discrimination indicator (Houbraken et al., 2018). The red section demonstrates the procession of congestion (lower speed) starting at 8km and propagating backward over time. A substantial amount of floating car data is essential for tasks related to identifying traffic congestion. For instance, in one study, the historical FCD data of 4908 vehicles collected over 5 years totaled 6.55 billion records, with moving vehicles sampled approximately once every 100 meters and stationary vehicles sampled every 5 minutes (Erdelić et al., 2021). It is apparent that congestion indicators often include average speed, speed differentials, and congestion indices. Previous studies predominantly focused on static congestion, categorizing congestion states based on indicator thresholds. Recent research has delved into the multidimensional spatiotemporal characteristics, evolution processes, and spatiotemporal multimodal models of traffic congestion.

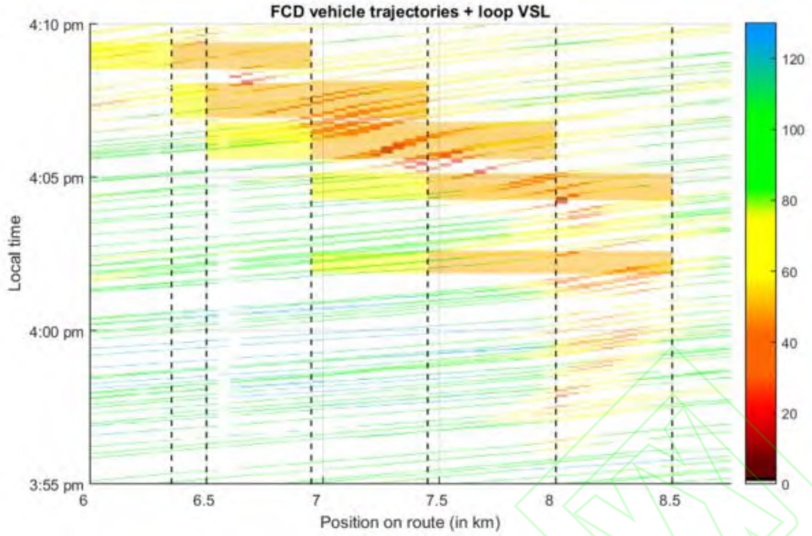


Fig. 8 Congestion transfer process (color represents speed) (Houbraken et al., 2018).

4.4 Route planning (Cluster three)

Route planning and travel time prediction are typical scenarios for utilizing floating car data. Floating car data inherently includes timestamp information, enabling direct measurement of travel time. The primary research focal points within this cluster are summarized in Table 5

Table 5 Research hotspots of floating car data in the field of path planning and travel time prediction.

Application directions	**Research topics**	**Related literature**	**Main research content**
Route planning	Path identification	Dabbas et al. (2021); Croce et al. (2020); Comi and Polimeni (2022)	Utilizing floating car data to obtain actual travel times and identifying the actual path by comparing estimated travel times of potential paths
	Path optimization	Shen and Ban (2016); Jayol et al. (2022)	Researching path selection models with travel time or traffic emissions as optimization objectives
Travel time prediction	Segment level	Zheng and Van Zuylen (2013); Song et al. (2013); Zhang, et al. (2017); Jiang et al. (2014); Jenelius and Koutsopoulos (2013)	Employing theoretical analytical methods (such as neural networks, Kalman filtering models, and maximum likelihood models) or directly estimating travel times in practical engineering applications
	Intersections	Liu et al. (2016); Luo et al. (2019); Gan et al. (2017)	Estimating queue lengths at signalized intersections and turning delays at signalized intersections
	Path level	Rahmani et al. (2015); Qin and Yun (2018); Li et al. (2022)	Using Bayesian methods with a particle filter framework, convolutional neural networks, and long short-term memory networks to extract spatiotemporal traffic flow features for predicting path-level travel times
	Road network level	Li, et al. (2022); Jiang et al. (2014)	Studying the input proportion and sampling frequency of floating car data based on the dynamic changes in traffic flow characteristics of the road network

4.4.1 Identification of actual paths

Path identification typically relies on a substantial volume of historical traffic data, combined with machine learning and deep learning techniques, to identify drivers' actual driving paths by analyzing the correlation patterns among various locations within the traffic network. The primary approach for path identification based on floating car data involves initially extracting travel time details, where the actual driving route is identified as the one with estimated travel times closest to the actual travel time among all feasible paths. Floating car data and vehicle RFID data are commonly used to extract travel time information. Dabbas et al. (2021) aggregated floating car data with a sampling rate of 50% below 15 seconds and 75% below 30 seconds, finding that aggregated FCD can accurately estimate routes. Bracci et al. (2021) constructed a set of approximately 10,000 actual floating car data vehicle driving routes.

4.4.2 Optimal route planning

Vehicle route selection typically aims to minimize travel time and reduce fuel consumption as optimization objectives, considering factors like traffic congestion, road conditions, and vehicle characteristics. Optimization algorithms are used to find the best route. Floating car trajectory data can extract travel time information and is commonly used to develop optimal route selection models to reduce congestion. Additionally, speed information can be leveraged to analyze route selection models to minimize traffic pollutant emissions.

With the goal of minimizing travel time, Shen and Ban (2016) proposed a route analysis algorithm based on Floating Car Data (FCD) that considers dynamic edge weights for the shortest travel time. Yang et al. (2015) using GPS data obtained from the taxi dispatch system in Beijing, extracted Origin-Destination (OD) pairs for travel time information. By comprehensively considering the influence of road network conditions and traffic conditions on driver route selection behavior, they constructed a taxi multi-path probability selection model based on the Path-Size Logit (PSL) model. Davoodi and Mesgari (2015) introduced an ant colony optimization algorithm that includes the fusion of non-homogeneous traffic data, including floating car data, to determine the optimal route.

With the aim of reducing traffic emissions, analyzing speeds on road segments every fifteen minutes can help identify a less polluting travel route by utilizing a dynamic speed map, COPERT emission factors, and a time-dependent Dijkstra algorithm (Jayol et al., 2022).

4.4.3 Segment-level travel time prediction

Methods for predicting travel times on road segments based on floating car data mainly consist of theoretical analytical methods and practical engineering methods.

Theoretical analytical methods encompass neural network models (Zheng and Van Zuylen, 2013; Song et al., 2013; Zhang, Zhu, et al., 2017), Kalman filtering models (Jiang, Zhang, & Xia, 2014), maximum likelihood models (Jenelius and Koutsopoulos, 2013), etc. The travel time extracted from floating car data or the ratio of travel times between the target segment and adjacent segments can serve as the model output, while the input can include the relationship features between the target segment and adjacent segments (Zhang, Zhu, et al., 2017), road feature parameters (speed limits, road grades, etc.) (Jenelius and Koutsopoulos, 2013), road environment conditions (traffic volume changes over a period, current season and weather conditions, etc.) (Zhong et al., 2021), and intersection signal timing information (Jenelius and Koutsopoulos, 2013).

Theoretical analytical methods often involve numerous assumptions and intricate modeling, rendering them largely impractical for engineering applications. Additionally, many theoretical analytical methods are required to necessitate intersection signal timing data. Given the restricted sharing of traffic data among urban traffic management departments in China, crucial data, especially intersection signal timing data, is often inaccessible, greatly impeding the implementation of these methods.

Field engineering models include direct methods and indirect methods. Direct methods utilize the positional coordinates of FCD points on both sides of a road segment boundary to estimate the moment when vehicles cross the segment boundary through interpolation, thereby calculating the travel time for a single vehicle on the road segment. Indirect methods use the instantaneous speed sequence of FCD points to estimate the average speed of vehicles passing through the road segment and then calculate the travel time for a single vehicle on the road segment. As the sampling interval of FCD data increases, the estimation errors of travel times obtained by both direct and indirect methods show an increasing trend. When the sampling interval is less than 7 seconds, the estimation errors of direct and indirect methods are quite similar. However, if the sampling interval surpasses 7 seconds, the direct method demonstrates notably superior estimation performance compared to the indirect method (Jiang et al., 2009).

The studies mentioned highlight that theoretical analytical methods entail various assumptions, and complex modeling, and are challenging to implement, yet they provide high prediction accuracy, often utilizing travel time information from floating car data as the model output. Conversely, practical engineering models are straightforward to implement, but they assume vehicles travel at a constant speed between adjacent floating car data points, which may not reflect real-world conditions accurately. Additionally, the implementation conditions assume the presence of floating car data points near the boundaries of road segments, but most data have relatively low sampling frequencies, which do not meet the assumption.

4.4.4 Intersection delay time

Vehicle queues and turning delays at signalized intersections directly contribute to an increase in travel time (Axer and Friedrich, 2016). Some scholars have utilized floating car data to study queue lengths and turning delays at signalized intersections.

The location information of the tail-end floating car is a key factor in estimating the real-time queue length at signalized intersections. (Zhuang et al., 2013). However, solely determining queue length based on the position of the tail-end floating car relies on the ideal assumption that the floating car is at the tail end. Liu et al. (2016) improved upon this by using the tail-end floating car position as the base and employing a weighted average arrival rate to represent the arrival rate of vehicles after the tail-end floating car, thereby calculating the maximum queue length of vehicles within a signal cycle. This method not only inherits the estimation concept based on the tail-end floating car position but also introduces the concept of arrival rate, making the model more reflective of real-world scenarios.

In terms of intersection turning delays, Liu et al. (2013) proposed a technical framework for calculating turning delays at intersections using floating car data, combining the mainline method. Some scholars have applied deep learning methods to estimate intersection delays. Liu et al. (2013) computed the necessary parameters for delay estimation by analyzing the spatiotemporal trajectories of floating cars, determined intersection signal timing through K-S testing and density clustering algorithms, identified the total number of blocked vehicles at the intersection based on traffic wave theory, and calculated the average delay by integrating traffic flow mechanics at intersections. Liu et al. (2013) utilized loop detector data and floating car travel times to calculate the average delay time at signal-controlled intersections.

4.4.5 Path-Level travel time prediction

The methods for predicting path-level travel time using floating car data are similar to those for segment-level travel time. However, due to the increased complexity of path research compared to segments, predicting travel times at the path level faces challenges in research accuracy due to low frequency and low penetration rates. Scholars have attempted to address this issue by employing non-parametric path travel time distribution estimation (Rahmani et al., 2015), Bayesian methods within a particle filter framework (Qin and Yun, 2018), and combined methods using convolutional neural networks and long short-term memory networks (Li et al., 2022). Among these, the deep learning network model that combines convolutional neural networks and long short-term memory networks considers spatial dependencies, temporal dependencies of road segments, and the issue of temporal drift in coarse-grained data, achieving a prediction accuracy of over 90%, with the method's efficiency currently at a leading level among similar models. Zhang et al. (2017) based on a large amount of floating car data, extracted spatiotemporal traffic features using gray-level co-occurrence matrices and predicted future travel times by combining a negative exponential weighted combination of empirical travel times within a given departure time. Additionally, Rahmani et al. (2017) proposed a fixed-point formula for the joint problem of path travel time prediction and path inference, simultaneously solving both problems.

4.4.6 Network-Level travel time prediction

Predicting network-level travel times typically relies on historical traffic data, real-time traffic information, road topology, and other relevant factors. Various machine learning, deep learning, and data mining techniques are utilized for modeling. Traffic volume, occupancy, and segment travel time parameters serve as inputs for road network travel time prediction models (Li, Wang, et al., 2022; Jiang, Zhang, Xu, et al., 2014). Predicting network-level travel times requires consideration of whether the data adequately captures the dynamic characteristics of the road network. Dynamic factors such as traffic volume and spatial environment on the roads evolve over time and directly impact network travel times. Selecting the appropriate proportion of floating cars and the sampling frequency is also essential for capturing these dynamic characteristics.

5 FCD data quality analysis

Based on the high-frequency floating car data obtained from future large-scale connected vehicles, the emphasis is on its application in highway scenarios. This section utilizes the high-frequency floating car data from a logistics delivery truck of a freight company that routinely travels on highways as a sample for data collection in future large-scale connected vehicles. The characteristics of high-frequency floating car data are examined, and the aspects of sampling proportion and sampling frequency are investigated.

5.1 Sampling proportion

Sampling proportion refers to the ratio of the number of floating cars to the total traffic volume. Generally, a too small proportion of floating cars can increase the impact of individual randomness, thereby affecting the reliability of traffic data; while a proportion that is too large can increase the installation and maintenance costs of the system and the information processing load (An et al., 2021).

Previous research has extensively studied how to determine the optimal floating car sampling proportion. The common method used currently is to determine the number of floating cars based on the confidence level of a particular traffic characteristic, which can meet the needs of selecting an appropriate number of floating cars for different research purposes (Tu et al., 2006). Quiroga and Bullock (1998), Srinivasan et al. (1996), and Ygnace et al. (2001) analyzed the optimal floating car proportion based on the accuracy of estimating average speeds on road segments, the reliability of estimating travel times on road segments, and floating car positioning technology, respectively.

Sunderrajan et al. (2016) suggested that at least 5%-10% of floating cars are needed to achieve reliable traffic state estimation. Houbraken et al. (2018) based on the Netherlands' national floating car proportion of 6%-8%, demonstrated that the collected FCD at this proportion could serve as an alternative to loop detector data in dynamic traffic management systems. Kerner et al. (2005) found that a high-quality reconstruction of actual travel times in the road network could be achieved when the penetration rate of floating cars among all vehicles is 1.5%.

The methods mentioned above have limited consideration for traffic characteristics. Determining the optimal floating car proportion based on simulation experiments is more in line with the characteristics of traffic systems and is also more convenient and flexible (Li et al., 2008; Cheu et al., 2010; Lin et al., 2008). Lin et al. (2008) used VISSIM simulation experiments and their results showed that when the penetration rate of floating cars was between 3% to 5%, the accuracy could reach over 95%. Tang et al. (2014) also based on VISSIM simulations, found that considering economic costs, using a floating car proportion lower than 8% was more reasonable, resulting in relatively small speed relative errors. The variation of speed relative errors with the floating car proportion in his study is shown in Fig. 9.

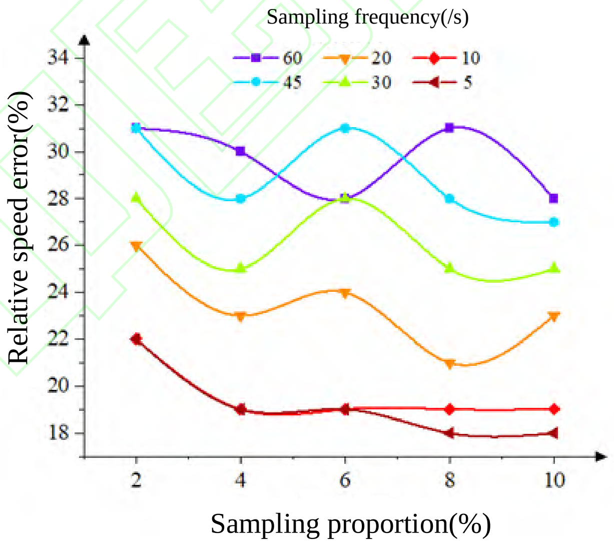


Fig. 9 Relative errors in speed for different floating car proportions.

An et al. (2021) based on TransModeler simulations, found that as the floating car proportion increased, the relative speed errors of each road segment significantly decreased. At higher floating car proportions, the errors gradually tended to be the same, with a turning point in speed relative errors at around 5%. Altintasi et al. (2022) proposed a method to evaluate the proportion of commercial Floating

Car Data (FCD) by assessing the quality of floating car speed data by comparing it with actual data. The results showed that reliable speed values could be obtained when the floating car proportion was above 15%. The recommended floating car proportions by previous researchers are shown in Table 6. The range of floating car proportions is around 3-8%, with specific values determined through simulation experiments based on actual traffic conditions and research purposes.

Table 6 Recommended floating car proportions.

References	Year	Methods	Recommended proportions	Evaluation criteria
Kerner et al. (2005)	2005	Optimal single-vehicle cost	1.50%	Identify traffic events with the least amount of FCD information.
Lin et al. (2008)	2008	VISSIM simulation	Greater than 3%	Exhibit relatively small errors in travel time.
Tang et al. (2014)	2014	VISSIM simulation	Less than 8%	Demonstrate relatively small errors in median speed.
Houbraken et al. (2018)	2018	Comparison with loop detector data	6%-8%	At this proportion, it can serve as a substitute for loop detector data.
An et al. (2021)	2021	TransModeler simulation	5%-8%	Show relatively small errors in speed.

Some scholars have proposed that integrating data can reduce the data errors caused by the FCD proportion and have studied the speed-relative errors of fused data from loop detectors and floating cars (Ambühl and Menendez, 2016). They found that even when fusing the two types of data with small sampling proportions, significant estimation errors may still occur, as shown in Fig. 10.

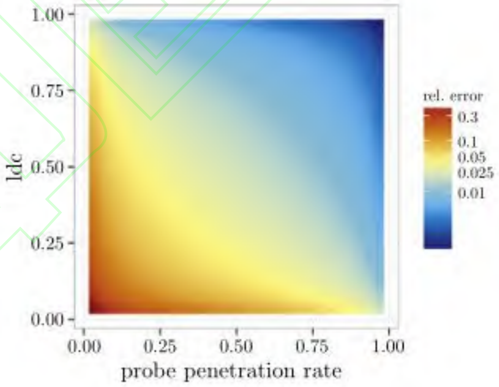


Fig. 10 Relative error of 95% speed for multi-source data fusion (Ambühl and Menendez, 2016).

5.2 *Sampling frequency*

The way in which floating cars send data is known as the sampling strategy, and there are at least three different strategies (Sense, 2008; Herrera, 2009): time-based sampling strategy, space-based sampling strategy, and event-based sampling strategy. Time-based sampling strategy involves frequent data transmission over time. The time interval between two consecutive data packets is referred to as the sampling interval. Space-based sampling involves sending data to specific geographic locations that may be of interest (Hoh et al., 2011). Lastly, in event-based or trigger-based sampling, data is sent after a specific action is performed, such as braking (detected by an accelerometer) or horn honking (detected by a microphone).

This section only discusses the frequency of the time-based sampling strategy. The sampling frequency of floating car data refers to the duration between sending data back to the central processing unit, as shown in the specific transmission path in Fig. 11. The data transmission frequency depends on the available bandwidth and energy consumption of the underlying communication system. Floating car data can have low, medium, and high sampling frequencies. Generally, a sampling time interval greater than 30 seconds is considered low-frequency sampling, less than 10 seconds is high-frequency sampling, and between the two is medium-frequency sampling (Houbraken et al., 2018; Chen et al., 2022). In large and medium-sized cities in China, the average sampling interval of actual floating car data is typically around 30 seconds (Li et al., 2014). Most of the literature within the scope of this study uses medium to low-frequency floating car data, with fixed sampling intervals including 15 seconds, 30 seconds, 45 seconds, and 1 minute. Some studies involve dynamic sampling, with sampling frequencies ranging from 20 seconds to 3 minutes.

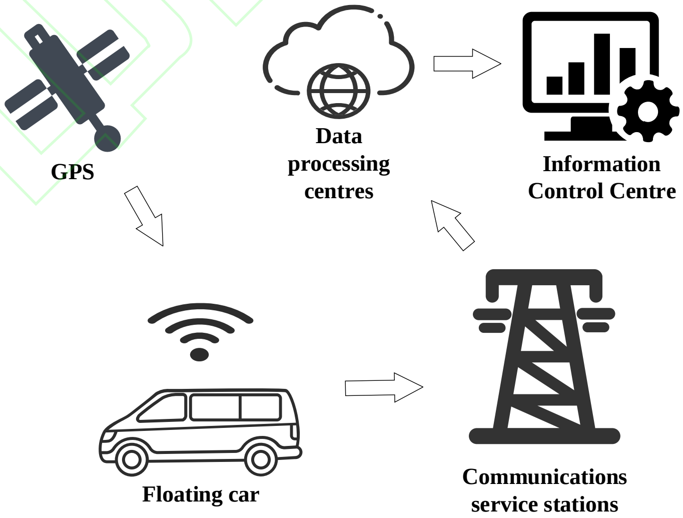


Fig. 11 Floating car data transfer path.

The sampling frequency impacts data collection accuracy, which further leads to error propagation. High sampling frequency gathers data with precise spatiotemporal information and extensive coverage of the road network, but they also entail substantial increases in communication costs and data processing volume. Conversely, low-frequency sampling data may have significant information gaps and high uncertainty, leading to larger errors when compared to actual operating conditions. Research has found that the data sampling frequency primarily affects vehicle speed and direction information, with minimal impact on spatial position information (Ranacher et al., 2016).

Due to privacy constraints in accessing floating car data, most researchers can only obtain low-frequency data aggregated from raw data. For example, Google Maps GPS data is aggregated every 5 minutes, and most taxi floating car data is aggregated every 30 seconds to 1 minute(Fu and Liu, 2023). As the utilization of connected vehicles continues to rise, acquiring high-frequency publicly available floating car datasets will no longer pose a challenge in the future. In this context, researchers need to consider how to select the sampling frequency for floating car data collection. Zhang et al. (2007) used Shannon's sampling theorem to determine the optimal sampling frequency on city roads as 10 seconds and on highways as 20 seconds. Tang et al. (2014) found that for traffic state estimation tasks, the optimal floating car sampling frequency is every 10 seconds, as higher sampling frequencies do not enhance the accuracy of traffic state estimation.

Currently, urban traffic information collection based on floating cars typically uses equidistant sampling, which does not optimize the sampling interval based on the differences in geometric conditions and states of the road network. To overcome the limitations of existing sampling algorithms, Cao and Peng (2014) focused on the actual road network. They took into account the topological relationships and geometric characteristics of the road network and used the spectral information of historical floating car speeds to dynamically optimize the sampling interval based on the cutoff frequency. This approach not only ensures data accuracy but also reduces data volume.

The studies mentioned above indicate that the sampling frequency has a significant impact on the accuracy of speed and vehicle heading angle information. In studies involving speed and heading angle, researchers need to consider the sampling frequency. Previous research has suggested optimal sampling frequencies of 10 seconds for city roads and 20 seconds for highways. Additionally, dynamic optimization of the sampling interval based on actual road conditions can be performed. For research work that only involves the spatial position information of floating cars, the sampling frequency of floating cars may not be a primary concern. Currently, research methods often assume that the raw data is high-frequency sampling data with Gaussian noise(Kong et al., 2018), however, these studies typically utilize medium to low-frequency data, resulting in a discrepancy between the methodologies and assumptions. Datasets with high sampling frequency facilitate for retrieving of vehicle information with higher temporal precision, enabling the capture of high-precision vehicle travel trajectories(Wang et al., 2024). High-frequency sampling, to some extent, diminishes the workload of map matching and data interpolation errors, enhances data certainty, and contributes to the precise execution of tasks such as traffic congestion and traffic accident prediction in Intelligent Transportation Systems (Cerqueira et al., 2018; Hofleitner et al., 2012).

6 Challenges and opportunities

Although a single source of floating car data can provide a wealth of traffic information, its limitations cannot be ignored. To overcome this limitation, a common practice is to integrate loop detector data with floating car data. However, from the perspective of data source diversity, in addition to loop detectors, other traffic data sources like ETC gantries, fixed cameras, etc., can be effectively integrated using the highway milepost system. Furthermore, environmental information such as weather data, traffic accident data, and construction zone data can also be fused based on the milepost system to achieve more accurate traffic state analysis. Through multi-source data fusion technology, not only can the sparsity and errors that may arise from a single data source be reduced technically, but also in practical terms, the perceptual capabilities of vehicles towards their surroundings can be significantly enhanced. This comprehensive understanding of road conditions provides more solid data support for traffic management and decision-making. Future research needs to further explore data feature extraction techniques and fusion algorithm selection in the integration of non-homogeneous data sources to enhance the accuracy and efficiency of information fusion. Specifically, spatial data fusion is achieved through the Markov model. The fusion of temporal data and spatial data requires spatiotemporal tensor transformation. For example, Constrained tensor fusion and unified tensor fusion (UTF) model. Additionally, improving the fusion algorithm to enhance fusion performance is crucial. The introduction of modern statistical theory, random set theory, fuzzy set theory, rough set theory, Bayes theory, evidence theory, support vector machine, and other intelligent computing technologies will present new development opportunities for state estimation of nonlinear non-Gaussian systems and heterogeneous data fusion.

High-frequency floating car data, characterized by its high sampling frequency, enables the collection of a large number of data sampling points on the mainline curved sections of highways. This feature provides the possibility to delve into the impact mechanisms of road geometric indicators on traffic safety and offers data support for the design and optimization of these indicators. The study could focus on developing effective data extraction and analysis methods to identify key indicators closely related to highway geometry from high-frequency floating car data, such as curvature, slope, and road surface conditions, among others. Furthermore, it could delve into how these geometric indicators influence safety and comfort, and how the design of rational geometric indicators can elevate the design standards of highways. Moreover, the research scope could extend to specific scenarios like highway interchanges, construction zones, and tunnels, analyzing the uniqueness of geometric indicators in these contexts and their influencing factors. Through these studies, a more precise scientific basis can be provided for the planning, design, and management of highways, thereby enhancing road operational efficiency and safety, and ultimately offering drivers a more comfortable and safer driving experience.

Floating car data provides direct speed information, supporting research topics related to speed. However, it was found in this study that nearly 80% of the floating car data consists of urban taxi data, predominantly small passenger cars, indicating a research gap in the study of heavy trucks. Considering trucks as heterogeneous road users with distinct driving characteristics from regular road users, and their tendency to cause more severe traffic accidents, they should be a focal point for researchers(Zubaidi et al., 2022; Zhang et al., 2024). Therefore, this study proposes initiating a systematic investigation of heavy trucks from a traffic safety perspective, including but not limited to the following aspects. Analyzing speed characteristics of heavy trucks in different road sections and traffic conditions, deciphering their driving patterns and influencing factors. Studying the interaction of heavy trucks with other vehicles in traffic flow, analyzing their impact on overall traffic flow. Comparing the safety differences between heavy trucks and small passenger cars, and exploring strategies for traffic management tailored to the characteristics of heavy trucks to enhance road safety. Through in-depth research on floating car data related to heavy trucks, it is anticipated that comprehensive data support and decision-making foundations can be provided for the field of traffic safety.

Establishing a lightweight model for rapid mining and analysis of high spatiotemporal resolution floating car data. Floating cars, serving as mobile sensors on highways, provide enhanced spatiotemporal resolution and real-time capabilities compared to fixed sensors. There is a critical need for in-depth research on expediting the mining and analysis of real-time data, and designing lightweight real-time data processing algorithms or systems to support real-time applications such as traffic management and intelligent navigation. Additionally, current research utilizing floating car data could explore the spatiotemporal characteristics of the data. For example, exploring methods like the CNN+LSTM approach, Transformer, and graph neural networks can shed light on the efficacy of various deep learning models in capturing spatiotemporal patterns, allowing for a nuanced understanding of their strengths and limitations.

Proposing a system of evaluation metrics and optimization methods for enhancing the quality of floating car data. In the era of big data, data-driven methods such as deep learning and reinforcement learning have gained prominence. However, the optimization and improvement of these data-driven methods also rely on continuous training with high-quality data. Data quality serves as the fundamental factor influencing prediction errors in deep learning models. For instance, a lower sampling proportion and frequency of floating car data result in diminished data quality, leading directly to inaccuracies in outcomes such as speed and travel time. Exploring evaluation metrics for floating car data quality, including but not limited to data completeness, accuracy, consistency, timeliness, etc., and investigating how to comprehensively consider multiple metrics to assess the quality of floating car data. Studying the impact of different metrics on prediction errors in deep learning models and how optimizing data quality can enhance model performance and generalization capabilities.

7 Conclusions

The paper provides a review of the research outcomes on floating car data in the transportation field from 2000 to 2023. It introduces the data sources and analysis methods, analyzes the distribution of literature by year, research areas, and institutions, discusses research hotspots, and primarily reviews the research outcomes of three major clustering groups: traffic flow state, traffic safety, and route planning. It further analyzes the influencing factors of floating car data quality, and finally proposes five challenges and opportunities that high-frequency floating car data may face in highway scenarios. The main research conclusions are as follows.

Research on floating car data can be divided into three stages: initial development stage, rapid development stage, and steady development stage. China and the United States have become the largest research entities in floating car data. Major research institutions include Beijing Jiao Tong University, Tongji University, Tsinghua University in China, Florida State University, and the University of California in the United States. The mainstream research focuses on machine learning and deep learning methods, with research topics revolving around traffic flow state and travel time prediction.

These research studies can be categorized into three types: traffic flow state, traffic safety, and route planning. The primary issues include the oversight of data quality defects caused by data sampling proportions and sampling frequencies, as well as the problem of data misalignment with the assumptions of the methods. The review indicates that the recommended range for sampling proportions is 3-8%. The sampling frequency significantly affects the accuracy of speed and heading angle data, while exerting minimal impact on positional parameters. The fusion of multi-source heterogeneous data can improve the quality deficiencies of individual floating car data, and deep neural networks can effectively capture the spatiotemporal characteristics of floating car data.

The research is limited by the lack of quantitative studies on sampling frequency and proportions. Future research topics could include investigating methods for multi-source heterogeneous information fusion based on the highway milepost system, analyzing the impact mechanisms of high-frequency floating car data on highway alignment indicators, developing continuous operating speed prediction models for heavy trucks in interchange scenarios, establishing lightweight models for efficient mining and analysis of high spatiotemporal resolution floating car data, and proposing an evaluation metrics system and optimization methods for enhancing floating car data quality.

Conflict of interest

The authors do not have any conflict of interest with other entities or researchers.

Acknowledgments

The study was supported by the Natural Science Foundation of Shaanxi Province (2023-JC-YB-391), the National Key Research & Development Program of China (2020YFC1512003); Sichuan Province Science and Technology Support Program (2022YFG0048); Science and Technology Project of Sichuan Transportation Department (2022-ZL-04).

References

Altintasi, O., Tuydes-Yaman, H.,Tuncay, K., 2017. Detection of urban traffic patterns from floating car data (FCD). Transportation Research Procedia 22, 382-391.

Altintasi, O., Tuydes-Yaman, H.,Tuncay, K., 2022. A method to estimate traffic penetration rates of commercial floating car data using speed information. Transport 37(3), 161-176.

Ambühl, L., Menendez, M., 2016. Data fusion algorithm for macroscopic fundamental diagram estimation. Transportation Research Part C: Emerging Technologies 71, 184-197.

An, Y., Jiao, P., Li, Y., et al., 2021. Research on the method of proportional value of floating vehicles based on TransModeler simulation. Journal of Beijing University of Civil Engineering and Architecture 37(3), 56-63.

Antoniou, C., Barcelo, J., Breen, M., et al., 2016. Towards a generic benchmarking platform for origin-destination flows estimation/updating algorithms: Design, demonstration and validation. Transportation Research Part C-Emerging Technologies 66, 79-98.

Asakura, Y., Kusakabe, T., Long Xuan, N., et al., 2017. Incident detection methods using probe vehicles with on-board GPS equipment. Transportation Research Part C-Emerging Technologies 81, 330-341.

Axer, S., Friedrich, B., 2016. A methodology for signal timing estimation based on low frequency floating car data: Analysis of needed sample sizes and influencing factors. Transportation Research Procedia 15, 220-232.

Beliakov, G., Gagolewski, M., James, S., et al., 2018. Measuring traffic congestion: An approach based on learning weighted inequality, spread and aggregation indices from comparison data. Applied Soft Computing 67, 910-919.

Bracci, A., Colombaroni, C., Fusco, G., et al. Investigation and modeling on drivers' route and departure time choices from a big data set of floating car data. In: The 7th International Conference on Models and Technologies for Intelligent Transportation Systems, Heraklion, 2021.

Cao, W., Peng, X., 2014. Real road network oriented optimization method of floating car sampling interval. Journal of Data Acquisition and Processing 29(5), 770-776.

Carlos, F., Daganzo, 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. Transportation Research Part B Methodological 41(1), 49-62.

Cerqueira, V., Moreira-Matias, L., Khiari, J., et al., 2018. On evaluating floating car data quality for knowledge discovery. IEEE Transactions on Intelligent Transportation Systems 19(11), 3749-3760.

Chen, D., Zhou, S.-t., Chen, Y., et al., 2022. Traffic performance identification method based on adaptive congestion index. Journal of Transportation Systems Engineering and Information Technology 22(2), 137-144.

Cheu, R.L., Xie, C., Lee, D., 2010. Probe vehicle population and sample size for arterial speed estimation. Computer-aided Civil & Infrastructure Engineering 17(1), 53-60.

Comi, A., Polimeni, A., 2022. Estimating path choice models through floating car data. Forecasting 4(2), 525-537.

Croce, A.I., Musolino, G., Rindone, C., et al., 2020. Route and path choices of freight vehicles: A case study with floating car data. Sustainability 12(20), 8557.

D'Andre, E., Marcelloni, F., 2017. Detection of traffic congestion and incidents from GPS trace analysis. Expert Systems with Applications 73, 43-56.

Dabbas, H., Fourati, W., Friedrich, B., 2021. Using floating car data in route choice modelling-field study. Transportation Research Procedia 52, 700-707.

Daganzo, C.F., Geroliminis, N., 2008. An analytical approximation for the macroscopic fundamental diagram of urban traffic. Transportation Research Part B: Methodological 42(9), 771-781.

Darwish, T., Abu Bakar, K., 2015. Traffic density estimation in vehicular ad hoc networks: A review. Ad Hoc Networks 24, 337-351.

Davoodi, M., Mesgari, M.S., 2015. GIS-based route finding using ant colony optimization and urban traffic data from different sources. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 40, 129-133.

Del Serrone, G., Cantisani, G., Peluso, P., 2023. Blending of floating car data and point-based sensor data to deduce operating speeds under different traffic flow conditions. European Transport/Trasporti Europei (91), 1-11.

Deng, W., Lei, H., Zhou, X., 2013. Traffic state estimation and uncertainty quantification based on heterogeneous data sources: A three detector approach. Transportation Research Part B-Methodological 57, 132-157.

Erdelić, T., Carić, T., Erdelić, M., et al., 2021. Estimating congestion zones and travel time indexes based on the floating car data. Computers, Environment and Urban Systems 87, 101604.

Fu, R., Tong, L., Guo, Y., et al., 2019. A case study in China to determine whether GPS data and derivative indicator can be used to identify risky drivers. Journal of Advanced Transportation 2019(1): 9072531.

Fu, C., Liu, H., 2023. Investigating distance halo effect of fixed automated speed camera based on taxi GPS trajectory data. Journal of traffic and transportation engineering (English edition), 10(1), 70-85.

Gan, Q., Gomes, G., Bayen, A., 2017. Estimation of performance metrics at signalized intersections using loop detector data and probe travel times. IEEE Transactions on Intelligent Transportation Systems 18(11), 2939-2949.

Hara, Y., Suzuki, J., Kuwahara, M., 2018. Network-wide traffic state estimation using a mixture Gaussian graphical model and graphical lasso. Transportation Research Part C-Emerging Technologies 86, 622-638.

Herrera, J.C., 2009. Assessment of GPS-enabled smartphone data and its use in traffic state estimation for highways. University of California, Berkeley.

Heshami, S., Kattan, L., 2024. A stochastic microscopic based freeway traffic state and spatial-temporal pattern prediction in a connected vehicle environment. Journal of Intelligent Transportation Systems, 28(3): 313-339.

Hofleitner, A., Herring, R., Bayen, A., 2012. Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. Transportation Research Part B-Methodological 46(9), 1097-1122.

Hoh, B., Iwuchukwu, T., Jacobson, Q., et al., 2011. Enhancing privacy and accuracy in probe vehicle-based traffic monitoring via virtual trip lines. IEEE Transactions on Mobile Computing 11(5), 849-864.

Houbraken, M., Logghe, S., Schreuder, M., et al., 2017. Automated incident detection using real-time floating car data. Journal of Advanced Transportation 2017, 8241545.

Houbraken, M., Logghe, S., Audenaert, P., et al., 2018. Examining the potential of floating car data for dynamic traffic management. Iet Intelligent Transport Systems 12(5), 335-344.

Jalali, A., Torfeh Nejad, H., 2020. Incident detection in freeway based on autocorrelation factor of GPS probe data. International Journal of Intelligent Transportation Systems Research 18(1), 174-182.

Jayol, A., Lejri, D., Leclercq, L., 2022. Routes alternatives with reduced emissions: Large-scale statistical analysis of probe vehicle data in Lyon. Atmosphere 13(10), 1681.

Jenelius, E., Koutsopoulos, H.N., 2013. Travel time estimation for urban road networks using low frequency probe vehicle data. Transportation Research Part B-Methodological 53, 64-81.

Ji, Y.B.B., Xu, M.W., Li, J., et al., 2018. Determining the macroscopic fundamental diagram from mixed and partial traffic data. Promet-Traffic & Transportation 30(3), 267-279.

Jiang, G., Chang, A., Zhang, W., 2009. Comparison of link travel-time estimation methods based on GPS equipped floating car. In: International Conference on Transportation Engineering, Chengdu, 2009.

Jiang, Y., Song, G.H., Zhang, Z.Y., et al., 2021. Estimation of hourly traffic flows from floating car data for vehicle emission estimation. Journal of Advanced Transportation 2021, 1-11.

Jiang, Z., Zhang, C., Xia, Y., 2014. Travel time prediction model for urban road network based on multi-source data. Procedia-Social and Behavioral Sciences 138, 811-818.

Jiang, Z., Zhang, C., Xu, Z., et al., 2014. Development of a travel time prediction model for urban road network using multi-source data. Journal of Transport Information and Safety 32(3), 27-31.

Jiang, Z., Chen, X., Ouyang, Y., 2017. Traffic state and emission estimation for urban expressways based on heterogeneous data. Transportation Research Part D-Transport and Environment 53, 440-453.

Kan, Z.H., Tang, L.L., Kwan, M.P., et al., 2019. Traffic congestion analysis at the turn level using taxis' GPS trajectory data. Computers Environment and Urban Systems 74, 229-243.

Kerner, B.S., Demir, C., Herrtwich, R.G., et al. Traffic state detection with floating car data in road networks. In: Proceedings. 2005 IEEE Intelligent Transportation Systems, Vienna, 2005.

Knoop, V.L., van Erp, P. B., Leclercq, L., et al. Empirical MFDs using Google traffic data. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, 2018.

Kong, J., Hou, Z., Ren, Y. Findings on queue length based macroscopic fundamental diagrams with enhanced floating car estimation method. In: 2018 37th Chinese Control Conference, Wuhan, 2018.

Kyriacou, V., Englezou, Y., Panayiotou, C.G., et al. Estimating the posterior predictive distribution of the traffic density in multi-lane highways using spacing measurements. In: 2021 IEEE International Intelligent Transportation Systems Conference, Indianapolis, 2021.

Kyriacou, V., Englezou, Y., Panayiotou, C.G., et al., 2022. Bayesian traffic state estimation using extended floating car data. IEEE Transactions on Intelligent Transportation Systems 24(2), 1518-1532.

Laranjeiro, P.F., Merchan, D., Godoy, L.A., et al., 2019. Using GPS data to explore speed patterns and temporal fluctuations in urban logistics: The case of Sao Paulo, Brazil. Journal of Transport Geography 76, 114-129.

Lee, C., Saccomanno, F., Hellinga, B., et al., 2002. Analysis of crash precursors on instrumented freeways. Transportation Research Record 1784 (1), 1-8.

Li, H., Yang, X., Luo, L., 2014. Mining method of floating car data based on link travel time estimation. Journal of Traffic and Transportation Engineering 14(6), 100-109.

Li, J.J., Boonaert, J., Doniec, A., et al. Traffic flow multi-model with machine learning method based on floating car data. In: 2019 6th International Conference on Control, Decision and Information Technologies, Paris, 2019.

Li, Q., Miwa, T., Morikawa, T., 2008. Preliminary analysis on link travel time for probe-based estimation method by microscopic simulation. International Journal of ITS Research 6(1), 21-27.

Li, X., Quan, W., Wang, H., et al., 2022. Route travel time prediction on deep learning model through spatiotemporal features. Journal of Jilin University (Engineering and Technology Edition) 52(3), 557-563.

Li, Y.Y., Wang, S.Q., Zhang, X.R., et al., 2022. Estimation and reliability research of post-earthquake traffic travel time distribution based on floating car data. Applied Sciences-Basel 12(18), 9129.

Lian, Y.Q., Zhang, G.Q., Lee, J., et al., 2020. Review on big data applications in safety research of intelligent transportation systems and connected/automated vehicles. Accident Analysis and Prevention 146 (2020): 105711.

Lin, S., Xu, J., Xiong, W., 2008. Determination method of sample capacity of GPS floating vehicle based on simulation. Journal of Transport Information and Safety (4), 72-74.

Liu, X.L., Lu, F., Zhang, H.C., et al., 2013. Intersection delay estimation from floating car data via principal curves: A case study on Beijing's road network. Frontiers Of Earth Science 7(2), 206-216.

Liu, Z., Wang, W., Guo, H., et al., 2016. Queue length estimation of signalized intersections based on floating car data. Journal of Transportation Engineering 16(4), 7-11.

Luo, X., Cao, Y., Liu, B., et al., 2019. Accurate estimation simulation of signal delay at urban road network intersections. Computer Simulation 36(3), 202-206.

Mena-Oreja, J., Gozalvez, J., 2021. On the impact of floating car data and data fusion on the prediction of the traffic density, flow and speed using an error recurrent convolutional neural network. IEEE Access 9, 133710-133724.

Nantes, A., Ngoduy, D., Bhaskar, A., et al., 2016. Real-time traffic state estimation in urban corridors from heterogeneous data. Transportation Research Part C-Emerging Technologies 66, 99-118.

Neilson, A., Indratmo, Daniel, B., et al., 2019. Systematic review of the literature on big data in the transportation domain: Concepts and applications. Big Data Research 17, 35-44.

Oh, J.S., Oh, C., Ritchie, S.G., et al., 2005. Real-time estimation of accident likelihood for safety enhancement. Journal of Transportation Engineering 131(5), 358-363.

Pascale, A., Deflorio, F., Nicoli, M., et al., 2015. Motorway speed pattern identification from floating vehicle data for freight applications. Transportation Research Part C-Emerging Technologies 51, 104-119.

Pavlyuk, D., Jackson, I., 2022. Potential of vision-enhanced floating car data for urban traffic estimation. Transportation Research Procedia 62, 366-373.

Qin, W.W., Yun, M.P., 2018. Estimation of urban link travel time distribution using Markov chains and Bayesian approaches. Journal of Advanced Transportation 2018(1), 5148085.

Quiroga, C., Bullock, D., 1998. Determination of sample sizes for travel time studies. ITE Journal 68(8), 92-98.

Rahmani, M., Jenelius, E., Koutsopoulos, H.N., 2015. Non-parametric estimation of route travel time distributions from low-frequency floating car data. Transportation Research Part C-Emerging Technologies 58, 343-362.

Rahmani, M., Koutsopoulos, H.N., Jenelius, E., 2017. Travel time estimation from sparse floating car data with consistent path inference: A fixed point approach. Transportation Research Part C: Emerging Technologies 85, 628-643.

Ranacher, P., Brunauer, R., van der Spek, S., et al., 2016. What is an appropriate temporal sampling rate to record floating car data with a GPS? ISPRS International Journal of Geo-information 5(1), 1-17.

Rempe, F., Franeck, P., Fastenrath, U., et al., 2017. A phase-based smoothing method for accurate traffic speed estimation with floating car data. Transportation Research Part C-Emerging Technologies 85, 644-663.

Rempe, F., Franeck, P., Bogenberger, K., 2022. On the estimation of traffic speeds with deep convolutional neural networks given probe data. Transportation Research Part C-Emerging Technologies 134, 103448.

Rodriguez-Vega, M., Canudas-de-Wit, C., Fourati, H., 2021. Urban network traffic state estimation using a data-based approach. IFAC-PapersOnLine 54(2), 278-283.

Sekula, P., Markovic, N., Vander Laan, Z., et al., 2018. Estimating historical hourly traffic volumes via machine learning and vehicle probe data: A Maryland case study. Transportation Research Part C-Emerging Technologies 97, 147-158.

Sense, T., 2008. Rich monitoring of road and traffic conditions using mobile smartphones. MSR-TR-2008-59. Microsoft Research.

Seo, T., Kusakabe, T., Asakura, Y., 2015. Estimation of flow and density using probe vehicles with spacing measurement equipment. Transportation Research Part C-Emerging Technologies 53, 134-150.

Shen, J.W., Ban, Y.F., 2016. Route choice of the shortest travel time based on floating car data. Journal of Sensors 2016, 1-11.

Shi, Y., Deng, M., Yang, X.X., et al., 2018. Detecting anomalies in spatio-temporal flow data by constructing dynamic neighbourhoods. Computers Environment and Urban Systems 67, 80-96.

Song, B.I., Wang, Z.J., Han, C.W., et al., 2013. Estimation of left-turning travel time at traffic intersection. The Journal of China Universities of Posts and Telecommunications 20, 10-14.

Song, J.C., Zhao, C.L., Zhong, S.P., et al., 2019. Mapping spatio-temporal patterns and detecting the factors of traffic congestion with multi-source data fusion and mining techniques. Computers Environment and Urban Systems 77, 101364.

Srinivasan, Karthik, Jovanis, et al., 1996. Determination of number of probe vehicles required for reliable travel time measurement in urban network. Journal of the Transportation Research Board 1537(1), 15-22.

Sun, Q.X., Sun, Y.X., Sun, L., et al., 2019. Research on traffic congestion characteristics of city business circles based on TPI data: The case of Qingdao, China. Physica A: Statistical Mechanics and Its Applications 534, 122214.

Sunderrajan, A., Viswanathan, V., Cai, W.T., et al., 2016. Traffic state estimation using floating car data. Procedia Computer Science 80 (2016): 2008-2018.

Tang, K., Mei, Y., Li, K., 2014. A simulation-based evaluation of traffic state estimation accuracy by using floating car data in complex road networks. Journal of Tongji University (Natural Science) 42(9), 1347-1351.

Tang, L., Kan, Z., Zhang, X., et al., 2016. Travel time estimation at intersections based on low-frequency spatial-temporal GPS trajectory big data. American Cartographer 43(5), 417-426.

Tu, Z., Li, H., Yao, C., et al., 2006. Study on the route coverage and the update cycle of transportation information based on the minimum samples of floating car. China Railway Science (5), 127-131.

Winfrey, C., Meleby, P., Miao, L., 2023. Using big data and machine learning to rank traffic signals in Tennessee. Journal of traffic and transportation engineering (English edition), 10(5), 918-933.

Wang, B., Wong, Y. D., Zhang, C., et al., 2024. Exploring the impact of rainfall on vehicle trajectory patterns and sideslip risk: an empirical investigation. Journal of Advanced Transportation, 2024(1), 3138719.

Xu, C., Wang, X., Yang, H., et al., 2019. Exploring the impacts of speed variances on safety performance of urban elevated expressways using GPS data. Accident Analysis & Prevention 123, 29-38.

Yang, H., Wang, Z., Xie, K., et al., 2017. Use of ubiquitous probe vehicle data for identifying secondary crashes. Transportation Research Part C-Emerging Technologies 82, 138-160.

Yang, Y., Yao, E.-j., Pan, L., et al., 2015. Taxi route choice behavior modeling based on GPS data. Journal of Transportation Systems Engineering and Information Technology 15(1), 81-86.

Ygnace, J.L., Drane, C., IEEE, et al. Cellular telecommunication and transportation convergence: a case study of a research conducted in California and in France on cellular positioning techniques and transportation issues. In: 2001 IEEE Intelligent Transportation Systems-Proceedings, Oakland, 2001.

Yoon, J., Jordon, J., van der Schaar, M. Gain: Missing data imputation using generative adversarial nets. In: International conference on machine learning, Stockholm, 2018.

Yu, R., Han, L., Zhang, H., 2021. Trajectory data based freeway high-risk events prediction and its influencing factors analyses. Accident Analysis and Prevention 154, 106085.

Zabat, M., Stabile, N., Farascaroli, S., et al. (1995). The aerodynamic performance of platoons: A final report.

Zercher, B.P., Feng, Y., Bush, M.F., 2024. Towards IMn with electrostatic drift fields: Resetting the potential of trapped ions between dimensions of ion mobility. International Journal of Mass Spectrometry 495, 117163.

Zhang, C., Yang, X., Yan, X., 2007. Method for floating cars sampling cycle optimization. Journal of Transportation Systems Engineering and Information Technology (3), 100-104.

Zhang, F., Zhu, X., Guo, W., et al., 2017. Sparse link travel time estimation using big data of floating car. Geomatics and Information Science of Wuhan University 42(1), 56-62.

Zhang, J.B., Song, G.H., Yu, L., et al., 2018. Identification and characteristics analysis of bottlenecks on urban expressways based on floating car data. Journal of Central South University 25(8), 2014-2024.

Zhang, X., Song, G., Lin, Z., et al., 2014. FCD-based identification method for urban recurrent traffic congestions. Journal of Transport Information and Safety 32(1), 5-9.

Zhang, Z., Wang, Y., Chen, P., et al., 2017. Probe data-driven travel time forecasting for urban expressways by matching similar spatiotemporal traffic patterns. Transportation Research Part C-Emerging Technologies 85, 476-493.

Zhang, Z., Yang, X., 2020. Freeway traffic speed estimation by regression machine-learning techniques using probe vehicle and sensor detector data. Journal of Transportation Engineering Part A-Systems 146(12), 04020138.

Zhang, Z., Nie, Q., Liu, J., et al., 2024. Machine learning based real-time prediction of freeway crash risk using crowdsourced probe vehicle data. Journal of Intelligent Transportation Systems 28(1), 84-102.

Zhao, P.J., Hu, H.Y., 2019. Geographical patterns of traffic congestion in growing megacities: Big data analytics from Beijing. Cities 92, 164-174.

Zheng, F.F., Van Zuylen, H., 2013. Urban link travel time estimation based on sparse probe vehicle data. Transportation Research Part C-Emerging Technologies 31, 145-157.

Zheng, J.F., Liu, H.X., 2017. Estimating traffic volumes for signalized intersections using connected vehicle data. Transportation Research Part C-Emerging Technologies 79, 347-362.

Zhong, S., He, J., Zhu, K., et al., 2021. Travel time estimation based on built environment and low frequency floating car data. Journal of Transportation Systems Engineering and Information Technology 21(4), 125-131.

Zhu, X., Wen, X., Zhang, J., et al., 2021. Discrimination of urban traffic congestion segment based on bus floating vehicle data. Journal of Wuhan University of Technology (Transportation Science & Engineering) 45(4), 666-671.

Zhuang, L., He, Z., Ye, W., et al., 2013. Queue length estimation based on floating car data. Journal of Transportation Systems Engineering and Information Technology 13(3), 78-84.

Zhang, M., Nie, Y., Zhang, C., et al., 2024. Analysis of the Duration of Mandatory Lane Changes for Heavy-Duty Trucks at Interchanges. Sustainability (2071-1050), 16(14).

Zubaidi, H., Alnedawi, A., Obaid, I., et al., 2022. Injury severities from heavy vehicle accidents: An exploratory empirical analysis. Journal of traffic and transportation engineering (English edition), 9(6), 991-1002.

Author biographies and photos

Chi Zhang, Ph.D. in Engineering, postdoctoral experience, professor at Chang'an University, recipient of the Wu Fu Zhenhua Transportation Education Outstanding Teacher Award, doctoral/international student supervisor, postdoctoral collaborative supervisor, visiting scholar at Clemson University in the United States, registered road engineer, engaged in research and teaching work in road overall design, road digitization, traffic safety, and interdisciplinary fields. Hosted and participated in over 40 scientific research projects, including national key research and development programs. Edited and collaborated on the publication of 2 textbooks and academic monographs. Published over 100 academic papers as first author/corresponding author, including more than 50 SCI/EI indexed papers and 10 authorized invention patents. Received 8 national and provincial-level teaching achievement awards, and 9 provincial-level scientific research awards or honors, and guided students to receive 15 national and provincial-level awards.

Zhou Yuming is a female researcher from Xi'an, Shaanxi Province, China. She obtained her Bachelor's degree in Traffic Engineering from the School of Transportation Engineering at Chang'an University in June 2021. In the same year, she was admitted to the Master's program in Transportation Engineering (Academic) at the School of Highways, Chang'an University, without taking the entrance examination. Currently, she is pursuing a Ph.D. degree in Road and Railway Engineering, specializing in traffic big data mining and the application of mathematical statistics and machine learning methods in studying road safety and traffic accidents. Zhou has been recognized for her academic achievements, including receiving the National Inspirational Scholarship and being acknowledged as an Outstanding Youth League Member. She has published four research papers and holds one patent.

Min Zhang is an associate professor and holds a PhD in Engineering in Transportation Planning and Management from Chang'an University. My research areas include traffic safety theory and technology, traffic planning and design, comprehensive transportation system analysis, etc. I have led or participated in national key research and development programs, central university high-tech research and cultivation projects, World Bank urban transportation cooperation projects (Xi'an Public Transport Operation Improvement Research), Xi'an urban transportation development research, and other scientific research projects. Published over 20 academic papers, won 1 third prize in science and technology from the China Highway Society, authorized 2 national invention patents, 1 utility model patent, and 3 software copyrights.

Wang Bo, a male member of the Communist Party of China, is a highly accomplished individual from Shangluo, Shaanxi Province. He holds a joint Ph.D. degree from Nanyang Technological University in Singapore and has been recognized for his outstanding achievements in various fields. Wang Bo has held several important positions, including Chairman of the Student Union and Graduate Student Union, as well as serving as a part-time counselor for graduate students. He has received numerous accolades, such as the "National Scholarship for Ph.D. Students" and the "Outstanding Youth of Shaanxi" award. Additionally, Wang Bo has made significant contributions to academic research, with publications in esteemed journals and multiple patents to his name. His dedication to self-improvement and volunteer work has also been acknowledged, further highlighting his commitment to serving the community.

Yuhan Nie graduated from Changsha University of Science and Technology with a major in Transportation Engineering and began pursuing a doctoral degree in Transportation Engineering at Chang'an University in 2024. He mainly engages in the development of traffic engineering software, as well as research on traffic big data technology and traffic safety.