TMSeer: Visual Analysis of City-level Travel Modes Using Cellular Signaling Data

Yuanzhi Zeng[1,a], Shuxian Gu[1,a], Shiwei Chen[a], Yong Wang[b] and Haipeng Zeng[*,a]

[a]School of Intelligent Systems Engineering, Sun Yat-Sen University, Shenzhen, 518107
[b]College of Computing and Data Science, Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Visualization
Travel modes
Urban mobility
Cellular signaling data
Visual analytics

ABSTRACT

Travel behavior analysis is crucial for various applications such as urban planning, transportation management, social behavior analysis, and business intelligence. Owing to the wide use of mobile phones, the massive cellular signaling data (CSD) provides us with an unprecedented opportunity to investigate people's travel modes. However, it is challenging to explore such massive CSD due to its intrinsic sparsity, large size, and complexity. To fill this gap, we propose *TMSeer*, a visual analytics approach to help domain experts explore CSD and analyze city-level travel modes of people. Specifically, we first design an unsupervised method that combines rule-based heuristics (RBH) and Gaussian Mixture Model (GMM) to infer travel modes from CSD. We also take advantage of GMM's characteristics to assess the uncertainty of inferred results. Then, we present novel visualizations to enable interactive multi-level exploration of CSD and in-depth analysis of travel modes: a map view with heatmaps to show the spatial distribution of urban traffic, a region view with hybrid radial diagrams based on clock and directional metaphors to display the traffic of different travel modes at the regional level; and a path view to visualize the detailed paths of different modes between regions. In particular, an enhanced Sankey diagram is designed to visualize the details of the movement, together with the novel band design to show the efficiency of travel modes in the path. We conducted two case studies and expert interviews with domain experts to evaluate our approach. The results demonstrate the effectiveness and usability of *TMSeer* in analyzing travel modes of people and urban mobility.

1. Introduction

With the fast development of the economy and technology, there have been various transportation choices for people to travel to different places. Even within the same city, people can travel between two locations via walking, driving, taking a bus, or the subway. The in-depth analysis of such traveling behaviors is crucial for many downstream tasks like transportation management, urban planning, social behavior analysis, and business intelligence. For example, to solve the traffic congestion problems within crowded cities, it is necessary to first understand the detailed travel behaviors of different people from different regions at various time periods [21]. When policymakers and urban planners design new policies to encourage people to travel by public transportation or bicycle instead of driving, they also need to gain a deep understanding of the detailed travel behaviors of people [4, 3]. When business owners want to find the best locations for their business, the traveling modes of people around the locations can also help them make the optimal decisions [39].

Various research has been done to analyze the travel modes of people. The most straightforward way to investigate the travel behaviors of people is to collect people's feedback on their travel preferences via questionnaires or online surveys [4, 3]. But such methods are often time-consuming and only provide limited rough information on people's overall travel preferences. It cannot enable a detailed and accurate analysis of people's travel modes across various regions and time ranges. Then, with the availability of taxi data, more research studies have focused on analyzing the taxi data to gain insights into the travel behaviors of crowds within cities. For example, Ferreira et al. [24] provided in-depth insight into mobility patterns by analyzing taxi trips. Du et al. [19] delved into the travel patterns of online taxis. Compared with the traditional questionnaire or survey-based methods, such approaches can help decision-makers and urban planners gain a detailed understanding of how people travel within cities via taxi. However, they cannot enable a comprehensive analysis of other travel modes like bicycles, walking, and subways.

Given the wide usage of mobile phones, cellular signaling data (CSD) emerge as a particularly suitable choice for analyzing travel modes. The massive volume of CSD data enables extensive population coverage, capturing mobility patterns across diverse user groups and geographic areas at a relatively low cost. This wide coverage makes CSD an invaluable resource for studying travel behaviors on a large scale. Previous studies have demonstrated the utility of CSD in travel analysis. For example, Aguiléra et al. [1] measured the quality of service of Paris transit system. Gundlegård et al. [26] estimated the travel demand based on CSD. Chiou et al. [12] analyzed the home and work locations, and travel purposes of people. However, they can not achieve a detailed analysis of people's travel patterns according to the travel mode. By collaborating with telecommunication companies

*Corresponding author.
zengyzh25@mail2.sysu.edu.cn (Y. Zeng[1,]); gushx3@mail2.sysu.edu.cn (S. Gu[1,]); chenshw39@mail2.sysu.edu.cn (S. Chen); yong-wang@ntu.edu.sg (Y. Wang); zenghp5@mail.sysu.edu.cn (H. Zeng[*,])
ORCID(s): 0000-0002-0339-0361 (H. Zeng[*,])
[1]The two authors contribute equally to this work.

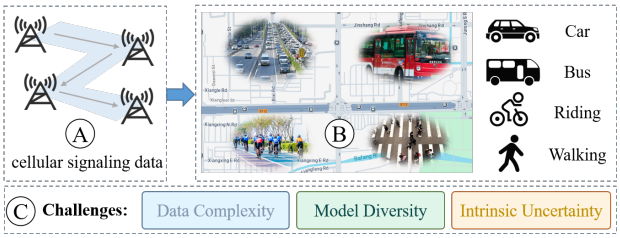

Figure 1: Motivation Illustration. (A) Cellular signaling data as the main data source. (B) The goal is to analyze four regional-level urban travel modes: car, bus, riding, and walking. (C) The three major challenges addressed in this study.

and employing proper data anonymization techniques, we aim to gain deep insights into the overall travel patterns of crowds within the city.

Fig. 1 serves as a motivation illustration, highlighting the overall goal of this research: to analyze city-level travel modes of urban residents at the regional scale using CSD. However, this is a non-trivial task due to three major challenges, as outlined below. **(1) Essential complexity in terms of the data scale, data dimensions, and dynamic evolution.** The analysis of large-scale, multi-attribute spatiotemporal data is a challenge. CSD is larger in scale and more redundant. In addition to spatio-temporal attributes, the consideration of travel modes increases the difficulty of multivariate data analysis. Each mode has its own set of attributes, such as speed, capacity, and comfort, which contribute to the overall decision-making process for travelers. **(2) Significant diversity of different travel modes.** The comparative analysis of the efficiency of different travel modes lays a solid foundation for optimal urban planning and traffic management. However, the choices of different travel modes result in an increasing difficulty of travel analysis. Multiple travel modes span across different temporal and spatial scales. It is challenging to effectively compare different travel modes and discover the relationship between them. **(3) Intrinsic uncertainty in the travel mode identification.** Inferring the travel mode people adopted using CSD is extremely challenging, due to the high spatio-temporal sparsity of CSD. This leads to fewer spatio-temporal characteristics in the trajectory. Although studies have been conducted using CSD to identify possible modes of travel, the accuracy of the identification is much lower than using GPS data [50, 11]. There is still potential to improve the accuracy. In other words, there is a great deal of uncertainty in the identification results. Due to these challenges, a fully automated analysis of travel modes is relatively difficult, lacking considerable experience and knowledge. Visual analysis, combined with both advanced computational power and human cognitive abilities, can be an effective solution for analyzing travel modes.

To address these challenges, we propose *TMSeer*, a visual analytics system designed to support the interactive exploration of large-scale cellular signaling data and facilitate city-level analysis of multiple travel modes. To manage the data complexity stemming from large-scale, multi-attribute, and dynamic CSD, *TMSeer* incorporates coordinated visual components, including a hybrid radial diagram and interactive filtering techniques. These allow users to explore high-volume spatiotemporal data across multiple travel modes in a scalable and organized manner. To address the diversity of travel modes, we develop novel visualization designs that support intuitive comparison across different transportation types. In particular, we introduce an enhanced Sankey diagram with a band representation that enables users to compare inter-regional travel efficiency, mode composition, and flow structure at a glance. To mitigate the uncertainty in travel mode identification, we propose a hybrid method that combines rule-based heuristics (RBH) and Gaussian Mixture Models (GMM) to infer travel modes from sparse CSD. GMM provides probabilistic outputs, which are further visualized to help users assess the confidence level of inferred results and focus on reliable insights. The contributions of this paper are summarized as follows:

- We develop a visual analytics system, *TMSeer*, to support the interactive exploration of large-scale CSD and analysis of city-level traffic based on multiple travel modes. An enhanced Sankey diagram is adopted to facilitate the exploration of inter-regional paths. Further, we propose a novel band design to show the details of different travel modes in the paths.
- We combine RBH and GMM to identify travel modes from CSD and take advantage of the outputs of GMM to assess the uncertainty in inferred outcomes.
- We conduct case studies and expert interviews using real-world CSD to evaluate our approach. The results demonstrate its usefulness and effectiveness in facilitating intuitive and in-depth analysis of travel modes within a city.

2. Related Work

2.1. Traffic Analysis of Travel Modes

The traffic analysis of traveling modes has been a hot issue for urban planning and traffic management. A large number of studies have been conducted on the mining of traffic data, providing guidance for the revision of transportation planning and policies, such as fare promotion [42] and bus route planning [54].

The study of massive-scale human movement in cities plays an important role in solving many of the problems that modern cities face. Many researchers focus on the movement of people within an urban area using various travel modes, namely urban mobility. Some research in this domain often focused on a single travel mode. Ferreira et al. [24] proposed a visual analytics system to explore spatio-temporal patterns of taxi movement within the city. Huang et al. [29] developed a graph-based approach to model the structure of traffic and discovered the hubs and backbone areas used by taxis, which benefit traffic management. Additionally, Zhao et al. [62] used the K-means++ algorithm to group people with similar

mobility patterns when taking the subway. Alhumoud et al. [2] uses CDRs to analyze human mobility patterns and predict metro usage, demonstrating how mobile phone data can enhance transport planning in emerging urban environments.

To broaden the scope, some researchers began to incorporate more travel modes. Zhong et al. [64] have enhanced the ability to distinguish similar modes (e.g., bus vs. car) through weakly-labeled integration of cellular/CDR with transit telemetry. Chen et al. [10] enriched movement trajectories with semantic information to identify intercity travel patterns. Nevertheless, fine-grained identification of multiple travel modes within urban areas, which involves more complex and overlapping mobility behaviors, remained underexplored. Another line of inquiry has examined interactions between different modes. Ma and Knaap [38] studied the impact of bike-share on rail transit ridership, finding it depends on the position of the metro rail. Chen et al. [7] unraveling the transit patterns between bus and subway. Yet, it is not able to dynamically present the spatio-temporal patterns between different travel modes, and it is not friendly to data analysts who do not have relevant knowledge.

Recent advances have begun to bridge the gap between single-mode analysis and city-scale multi-mode understanding, However, the traffic analysis of multiple travel modes is still in need of further research. There remains a shortage of studies that leverage large-scale urban mobility data, such as mobile phone signaling data, to jointly analyze and visually compare city-level travel patterns across multiple modes. Most existing approaches are either confined to a narrow spectrum of travel modes or to dynamic and interactive exploration. In this paper, we address these gaps by using city-scale CSD to analyze travel behavior across multiple modes, supported by an interactive visual analytics framework. Our approach enables the exploration of travel mode choices under real traffic conditions and facilitates intuitive spatio-temporal comparisons of multi-modal mobility patterns.

2.2. Identification of Travel Modes Based on CSD

Analysis of travel modes has always been a crucial component of transportation planning and policy making [56]. As traditional traffic surveys [45] are time-consuming, there have been many researchers working on automatic techniques for travel modes identification. With the increasing ubiquity of mobile phones, CSD have become a prominent data source for travel mode identification due to their low collection cost and broad population coverage. Unlike GPS data, which provide high spatial accuracy but require user consent and active tracking, CSD data are passively collected by mobile network operators, making them suitable for large-scale and cost-effective applications.

Since CSD record the location where the phone connects to the cellular tower, the spatial accuracy of CSD is significantly lower than that of GPS data. Many researchers have made efforts to this end. Early studies focused on extracting motion characteristics suitable for the sparse and noisy nature of CSD. For example, Chen et al. [9] designed a set of domain-specific features suitable for CSD, while Ding et al. [17] utilized several multidimensional mobility features to extract interpretable motion characteristics from CSD. Beyond handcrafted features, Jiang et al. [32] emphasized the importance of data preprocessing by eliminating outliers in raw CSD to enhance data quality before feature extraction. They proposed a novel travel pattern recognition framework based on deep neural networks, which modified the traditional data cleaning methods. These studies generally relied on heuristic or statistical features, such as travel speed, acceleration, and trajectory similarity, which are sensitive to data sparsity and often fail to generalize across diverse urban environments.

In addition to feature engineering, researchers have long incorporated external geographic data to compensate for the limited spatial accuracy of CSD. For instance, public transport route data were integrated to align bus trajectories with inferred paths, significantly improving recognition accuracy for public transit modes [11]. Similarly, online map APIs have been widely adopted to infer travel modes by matching coarse cell-tower sequences to road networks [44, 9]. However, reliance on these APIs introduces scalability and adaptability constraints, as they are often commercial, region-specific, and unsuitable for large-scale or real-time applications. Shen et al. [48] proposed a map matching technology based on deep reinforcement learning, which can project coarse cell-tower sequences onto road networks without commercial map APIs, substantially improving the downstream separability of mode-specific trajectories. However, methods that leverage GIS data, such as fine-grained maps or transport network layouts, often assume high data completeness and precision, which may not hold in areas with sparse base stations or complex road networks.

When it comes to data labels, the scarcity of mode-annotated CSD has been a major obstacle to supervised learning methods. To mitigate this, Chen et al. [9] employed small-scale labeled data collected by volunteers to train mode detection models. However, such datasets are limited in scale and demographic representativeness. Unsupervised methods, such as the PAM and k-medoids clustering algorithms used by Chin et al. [11], were proposed to bypass the need for ground truth. But these methods generally exhibit low accuracy. Due to the challenges in limited labels, recent work on CSD-based travel-mode identification pushes toward label-efficient pipelines at city scale. Transfer learning on real mobile phone signalling data (MSD) enables fine-grained travel mode identification with minimal labels and better cross-city generalization [30]. Also, integrated frameworks fuse signalling with public data (surveys, networks, census) to infer multimodal choices while keeping ground-truth demands low [37]. For transit-rich settings, fine-grained metro-trip detection directly from cellular trajectories shows that robust preprocessing and temporal regularization can recover station/line-level patterns without GPS traces [35].

In summary, while prior research on CSD-based travel mode identification has achieved notable progress, it still

faces several persistent limitations. In particular, labeling strategies often face inherent trade-offs between annotation cost and model accuracy, which motivates the exploration of effective unsupervised techniques suitable for large-scale applications. Furthermore, the inherent uncertainty in identification outcomes remains a critical yet under-addressed challenge. In this paper, we derived a set of features and developed an unsupervised method that combines RBH and GMM. Our method does not depend on map APIs and can identify travel modes of large-scale CSD. We also take advantage of the outputs of GMM to assess the uncertainty of identification results. Combined with visual interactions and domain knowledge, it aims to decrease the uncertainty of mode identification results.

2.3. Traffic Data Visualization

Traffic data records the movement of the urban population and generates a lot of multidimensional heterogeneous data. Visual analytics is widely used to analyze traffic data.

The visualizations used in urban visual analytics studies can be categorized as spatial, temporal, and other property visualizations according to data properties [63, 16]. The visualization of spatial context is the basis of urban analytics. Map-based visualizations enable urban analysis to be performed in a geographic context. Data dots represent geographic locations and spatial events [29], while lines represent trajectories [20]. Heatmap is a smooth representation of aggregated geographically located objects [16, 23]. Glyph can summarize and combine complex data [58, 5, 22]. Recent systems further couple map-based exploration with semantic urban functions and performance measures, e.g., SenseMap [8]. The flow map abstract and summarize massive crowd movement on a large spatial scale [34, 23]. Automated techniques like REA-FM improve the generation of natural-looking flow maps via tree layouts [53]. Temporal visualizations display temporal features along a timeline. The axis-based design is the most popular method thanks to its simplicity and understandability [18]. The radial layout can help to reveal the cyclic character of time [57]. What's more, urban data may also contain high-dimensional, relational, and semantic information, including numerical properties, categorical properties and textual properties. Prior research has explored appropriate visual channels to encode these attributes, such as parallel coordinate plot and matrix [54]. In addition, frameworks like Curio [41] coordinate attribute, spatial, and temporal visualizations within a unified dataflow framework. Beyond 2D layouts, a systematizes visual analytics for 3D urban data and discusses task–data–technique alignments in city-scale scenarios [40].

Besides, uncertainty visualization has become increasingly important due to the ubiquitous nature of uncertainty in data. This study specifically focuses on uncertainty caused by the model and parameters applied to the data analysis. Uncertainty visualization involves three components: identifying quantification, visual representation, and reducing uncertainty when possible. Uncertainty quantification mainly uses statistics metrics, such as variance and probability [15, 33], and confidence intervals [25]. Since uncertainty varies with data types, it drives varied visualization designs. Senaratne et al. [47] introduced space-time prisms with uncertain markers to show the uncertainty of activity area. Zeng et al. [60] encoded the uncertainty of emotion recognition by the height of bars to explore influencing factors. Deng et al. [15] utilized the saturation of colors to visualize the reliability and uncertainty of cascaded inference results. Recent work has examined cognitive impacts of uncertainty encodings and task pressure in spatial decision-making [6], and proposed congestion summaries that explicitly encode uncertainty in city-wide traffic patterns [5]. Wang et al. [52] encoded the effects of differential privacy noise in a grid-based correlation view to reveal uncertainty in correlation patterns. In addition, to reduce uncertainty, deductive reasoning [14] and interactive visualization systems [61] have shown promise in improving interpretability and reducing uncertainty.

In this study, the consideration of multiple travel modes leads to a more complex visualization task. We extend the original Sankey diagram to display the spatial time series among locations.

3. Problem Analysis

The analysis of travel modes based on CSD is a complex task that can rarely be solved using only automatic data mining methods. Due to the large, complex, and dynamic traffic data, these methods may not perform well without the involvement of domain experts. Although automatic and artificial intelligence technologies have made significant progress, there is still no substitute for human judgment and experience on some complex or subtle tasks. For example, automatic methods are difficult to understand complex traffic anomalies. Visualization as a bridge promotes the the collaboration between humans and AI [36].

In this section, we first surveyed the travel modes analysis from the transportation aspect and then analyzed why existing methods can not solve the problem. We further determined the limitations for using the existing transportation visualization systems, and propose our solutions from the visualization aspect.

3.1. Challenges of City-level Travel Mode Analysis

We tried to understand the difficulties of analyzing city-level travel modes with existing methods. There have been many studies devoted to mode identification, trajectory estimation and data mining. Due to the large, complex, and dynamic traffic data, the data mining methods cannot perform well without the involvement of domain experts. There are difficulties to analyze city-level travel modes with these common methods.

We conducted in-depth interviews with three experts (E1, E2 and E3). E1 is an urban planning expert from a traffic planning bureau in Guangdong Province of China, with over 5 years of experience studying data-driven solutions for urban problems. E2 is an urban road design expert

with over 2 years of experience and has been involved in several large traffic planning projects, including integrated transportation hub planning, and road network optimization. E3 is a researcher who has long been engaged in intelligent transportation research. We asked experts questions such as: How do multiple travel modes move spatially? What are the characteristics of travel behavior in different travel modes in terms of time? What is the uncertainty in the analysis process? By summarizing their feedback, we derived the major challenges of city-level travel mode analysis from the experts:

1) The scale of CSD and the diversity of travel modes pose challenges for visualization. It is challenging to compare travel modes from a spatio-temporal perspective and to construct visualizations, which is difficult to describe by a digital table or simple graph.

2) Travel is a dynamic process occurring over time, and the notable features of how multiple travel modes changes are different across different time and space scales. The existing visualization cannot support this dynamic comparison well.

3) Trajectory sparsity and travel mode recognition model bring uncertainty to the analysis. CSD cannot be mapped to precise sections of road. It leads to the loss of some mobility information, which hinders microscopic analysis of travel behavior and brings great uncertainty to the recognition result of travel mode.

3.2. Solutions

Visual analytics is a popular approach in facilitating human-AI collaboration which leverages the complementary strengths of humans and AI to enhance the travel mode analysis [36], rather than replacing humans with AI. The combination of data visualization and urban computing methods enables human in the loop, and promotes the efficiently exploration of traffic data [16].

There have been many successful visualization systems solving urban mobility problems. Overall, the spatial distribution can accurately be shown in these systems based on the map, and users can check the travel behavior of different travel modes in online systems. However, users cannot obtain the mobility changes of multiple travel modes at the same time. It is challenging to analyze the movement of different travel modes across multiple regions. Therefore, we have to draw on the visualization work and combine the necessary visual elements to design the visualization module of the travel modes analysis. Based on the challenges and successful visualization concepts, we have proposed a visual analytics system to deal with the difficulties raised by experts. We conducted literature reviews on the related studies and discussed with three domain experts (E1, E2, E3). From the feedback of these experts, we summarize a set of system requirements as shown below.

R1. Provide visual cues of travel mode identification results to facilitate more effective analysis. Firstly define a form to present the uncertainty of travel mode identification (**R1.1**). Moreover, visual coding for uncertainty is necessary to pay attention to correct results and combine expert knowledge (**R1.2**).

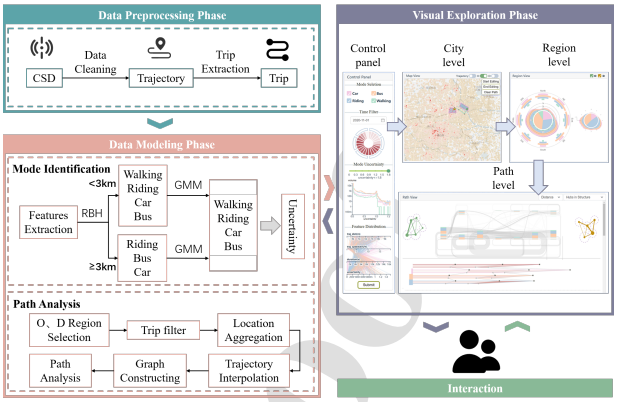

Figure 2: The system pipeline of *TMSeer*. In the data preprocessing phase, raw data is cleaned and trips are extracted. In the data modeling phase, mode identification and path analysis are carried out. In the visual exploration phase, four coordinated views with rich interactions are provided for travel modes exploration.

R2. Obtain the global movements of the whole city traffic. Analysts need to grasp an overview of urban traffic in different regions. Regions with more or less traffic can be discovered intuitively for further exploration.

R3. Implement a visual form to express the spatio-temporal patterns for regional traffic. First, it is necessary to explore different travel modes in the region (**R3.1**), such as exploring the number of people leaving or arriving in the region at different times of the day, and the travel modes used by people. Second, the traffic between the focused region and other surrounding areas is also noteworthy (**R3.2**).

R4. Visualize detailed information of movement along paths between two focused regions. An interpolation method that compensate for sparse trajectories can effectively promote mobility analysis (**R4.1**). In addition, it is necessary to visualize how it is actually used by different travel modes. The comparison of paths can help to find important or abnormal paths (**R4.2**), such as those with lower travel efficiency.

We designed the system *TMSeer*, to address the aforementioned requirements, combining multiple modeling methods and probability theory. As shown in Fig. 2, the analytical pipeline consists of three phases, namely, data preprocessing, data modeling, and visual exploration. In the data preprocessing phase, we clean CSD and extract trips. In the data modeling phase, we identify travel modes of trips (**R1.1**) and handle path analysis requests (**R4.1**). In the visual exploration phase, we implement our visual analytics system, *TMSeer*, based on the Vue.js[2] front-end framework and the Flask[3] back-end framework, which consists of four coordinated views. The control panel provides visual

[2] https://vuejs.org/
[3] https://flask.palletsprojects.com/en/3.0.x/

encoding showing uncertainty (**R1.2**). The map view shows an overview of the traffic density at the city level through several types of heatmaps (**R2**). The region view displays the traffic of different travel modes at the regional level with hybrid radial diagrams based on metaphors (**R3**). The path view provides detailed exploration for paths between two regions (**R4.2**).

4. Data Description and Processing

In this section, we first describe the types of data used in our approach and define the relevant terms. Then, we introduce the data processing and data modeling.

4.1. Data Description

We mainly used two types of data collected in Foshan, Guangdong Province, China. The detailed information is described as follows:

Cellular Signaling Data (CSD). In order to ensure the quality of communication services, telecommunication operators need to record the location of their users and collect a lot of data every day, which is called cellular signaling data (CSD). The location of the cell station connected by the phone is recorded, which can be seen as a kind of trajectory data. The collected data contains trajectory records of mobile phones and information on cell stations. Each record of a trajectory is defined as $tr = (pid, slo, t)$, where pid is the encrypted mobile phone ID, slo denotes the location of the connected cell station, and t is the timestamp. The data used in this work is provided by a large mobile phone service company in China, covering the trajectories of 244,928 users for the whole day of November 1st, 2020, which has been stripped of identifying information. It is used for large-scale urban travel mode exploration. In addition, a small-scale CSD dataset was constructed from two sources. First, 10 individual volunteers participated in data collection over a 3-day period. Among them, 3 volunteers were assigned to collect walking data, while the remaining 7 focused on bus travel data. All volunteers installed a custom mobile application that automatically recorded their connected cell tower information and timestamps while the phone was online and within coverage. Second, car travel data was collected over the course of one month from a single taxi vehicle using the same mobile application. The app continuously captured base station connection logs without requiring any manual input, ensuring consistent and passive data acquisition during regular driving. As a test set, the effectiveness of the travel mode identification algorithm and trajectory interpolation method proposed in this paper is evaluated.

Bus Route Data. The bus route data records detailed information about all bus lines and bus stops in Foshan, including the line name, stop name, and location. The data used in this work comprises 1,342 bus routes and 38,480 stops.

4.2. Definitions

We introduce the following important concepts defined in this paper:

- **Travel mode** refers to the method or means of transportation used by residents to travel, such as walking, riding, bus, car, and so on.
- **A trajectory** contains a set of locations of cell stations connected by phone users for a certain period of time, and consists of a sequence of records:

$$Tr = \{loc_1, loc_2, ..., loc_l\} \tag{1}$$

  where $loc_i = (pid_i, slo_i, t_i)$ and l is the record number.
- **A trip** refers to an individual route for the phone user that moves from origin (O) to destination (D). A trip includes the trajectory Tr_{od} from origin to destination and the inferred travel mode.
- **TripGraph** is a directed graph, $G_T = (N, E)$, to represent a network of trips between the specified origin and destination regions. It aggregates all trips between the specific origin and destination regions. The node n_i represents the location of the cell stations connected by phone users and the edge e_i represents the links between two physically-connected locations.
- **A path**, $P = \{e_1, e_2, ..., e_n\}$, is a finite sequence of edges from origin to destination in the graph G_T, indicating a trajectory that is commonly employed by people to travel from origin to destination.
- **Duration** refers to the time spent moving from one location to another.

4.3. Data Preprocessing

Data Cleaning. Oscillation and drift problems are the most prominent sources that result in the noise/outliers of CSD. We first employ the data cleaning process provided by Chen et al. [9] to remove data noise and obtain the trajectory of each mobile phone.

Trip Extraction. In this step, we try to identify the origin and destination of each trajectory and extract trips. According to the characteristics of human mobility [11], individuals generally remain in the same area for a while after they finish a trip. We identify records as “moving” and “stationary” based on the travel distances and duration. We calculate the travel duration and distance between adjacent trajectory points. If the interval time is more than 30 minutes, we label the point as “stationary". For the rest of the points, if the distance is more than 1 km [11], we label the point as “moving". Otherwise, we aggregate the current point with the next point and update the travel duration and distance that will be further judged for its status. Finally, we extracted 930,867 trips.

4.4. Travel Mode Identification

We first introduce the extracted features and then describe the proposed identification algorithm. At last, we show our methods for uncertainty assessment.

Feature Extraction. Table 1 shows the features we adopted for mode identification. We first investigate the characteristics of the traffic environment. As the public network can effectively aid in mode identification [46], we calculate the proximity to underlying bus route networks

Table 1
Features used for mode identification.

Features	Description
proximity_to_bus_route	The proximity of real trips to the bus route network.
duration	Duration of the entire trip.
traj_dist	Accumulated distance of all sampling points during the entire trip.
traj_speed	*traj_dist* divided by *duration*.

(*proximity_to_bus_route*). We introduce the buffer, which is a kind of sphere of influence or service of geospatial target [11]. We create buffer areas (size=200m, referring to [11]) for the bus route network. Then we measure the proximity to the bus route by calculating the proportion of real trips that fall into the buffer. *proximity_to_bus_route* reflects the overlap between trips and bus lines, which is beneficial to the identification of the bus mode. Cooperated with other features, we can identify bus separated from other vehicles.

Then, we explore spatio-temporal characteristics of travel behavior. In the time dimension, we consider the *duration* of the trip. The choice of travel mode will affect the travel duration of a trip to some extent. We check the distribution of travel time of different travel modes and found obvious differences, which can be used as the characteristics of travel mode identification.

In the space dimension, we consider travel distance. Two features can be extracted from the trip, including the Euclidean distance between origin and destination (*od_dist*), and the accumulated distance of all sampling points during the entire trip (*traj_dist*). We finally adopt *traj_dist* as a feature, as it can reflect the real trajectory of the trip and its distribution difference is more obvious, making it more conducive to travel mode identification. Then in the absence of special cases such as congestion, the moving speed is the most direct factor to infer the travel mode. Considering the above two characteristics of moving distance, two kinds of indicators can be calculated to measure the moving speed: OD speed (*od_speed*) and trajectory speed (*traj_speed*). We choose *traj_speed* to indicate the distance characteristics, which can more comprehensively measure the moving speed of the entire trip.

Mode Identification. We propose an unsupervised method that combines RBH (Rule-Based Heuristic) [46] and GMM (Gaussian Mixture Model) for travel mode identification. Specifically, We assume that every mobile phone user adopts a single transportation mode in his/her trips. As shown in Fig. 2, the RBH component refers to a simple distance-based rule: we use a 3 km threshold [46] to segment all trips into two subsets. Trips longer than 3 km are considered candidates for riding, bus, or car, while those shorter than 3 km also include walking. No further rule-based classification is applied within each subset. We then apply GMM (Gaussian Mixture Model) separately to each of the two subsets, leveraging its strength in modeling elliptical clusters across diverse feature distributions. The number of Gaussian components is set to 1 in each subset. During clustering, GMM estimates the model parameters via maximum likelihood, assigns data points to the most likely mode cluster based on posterior probability, and calculates the confidence of each assignment. For the classification task in this paper, the probability-based GMM can well adapt to the spatial and temporal feature distribution of different travel modes, and obtain the probability that the results belong to each travel mode, so as to measure the uncertainty of the identification results. x is the features of a trip, which is input into the model. K denotes the number of travel modes. The distribution of x is:

$$P(x|\Theta) = \sum_{k=1}^{K} \alpha_k N(x; \mu_k, \sigma_k^2) \tag{2}$$

where Θ denotes all Gaussian model parameters, α_k is the prior probability of the k-th Gaussian model, $\sum_{k=1}^{K} \alpha_k = 1$, μ_k and σ_k are the mean and standard variance in type k, respectively. The output of the model includes the cluster label of each trip, as well as the probability of belonging to each category. The introduction of probability in GMM can effectively handle data in overlapping clusters and facilitate uncertainty measurement.

Uncertainty Assessment. We use information entropy as a metric for quantifying uncertainty [27]. By calculating the probabilities that the trip belongs to each mode of travel obtained from the GMM, the uncertainty of the mode identification result for each trip is determined (**R1.1**). The equation is listed below:

$$H(x) = -\sum_{i} P(x_i) \log_b P(x_i) \tag{3}$$

where $P(x_i)$ denotes for the probability of trip x using travel mode i. The lower $H(x)$ indicates lower uncertainty, meaning the result is more credible.

4.5. Trajectory Interpolation and TripGraph Generation

To enable detailed path analysis of inter-regional travels, we propose a graph-based approach to interpolate paths, aiming to discover the main routes between different regions and uncover potential patterns (**R4.1**). It consists of four steps as shown below.

Location Aggregation. Considering the uncertainty in the spatial coverage of CSD, we first aggregate the locations of the neighboring cell stations to better extract the major paths. First, we performed a K-means clustering of all the cells C based on their latitude and longitude coordinates. Here we chose K-means for three reasons: i) it performs well when dealing with high-dimensional datasets [59], ii) the clustering efficiency is high and meets the requirement of real-time analysis of our system, and iii) it is credited with implementation simplicity, and it has been used widely for clustering in many domains [31]. In the following process,

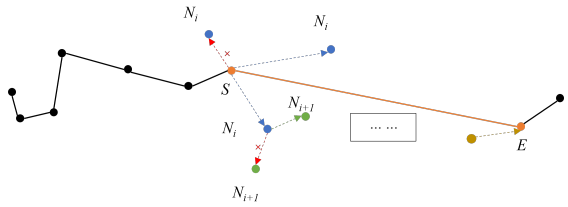

Figure 3: Schematic diagram of trajectory interpolation. Our method is based on the historical trajectory movement information of various travel modes. It searches potential paths through graph construction and graph traversal algorithm to realize sparse trajectory interpolation.

we replace each cell with its cluster center, and the trajectory dataset represented by the cluster centers is defined as C_k. At last, an entire trip is simplified into a series of lines of clustering centers. In this way, the clustered areas better reflect the potential location of the phone, and the clutter in the path is reduced.

Trajectory Interpolation. After aggregating locations, it is possible that some portions of trajectories are still quite sparse, which hinders the extraction of complete paths and makes it not conducive to our analysis. Thus, we aim to estimate the complete trajectories by designing a trajectory interpolation method for CSD.

We assume that: 1) The same travel mode usually has a common movement pattern when passing through a local area. 2) The majority of people usually have similar movement patterns between a pair of origin point and destination point. Inspired by the trajectory interpolation or recovery algorithms [13, 55, 51], we propose a simple trajectory interpolation method for CSD.

We use C_k to interpolate. For each node in C_k, there are a finite number of nodes that can be next connected, and the features of trajectories through each next node can be counted which is denoted as $n.f$. From this, we can naturally reason that the closer the features of the current trajectory to be interpolated are to $n.f$, the higher the probability that the trajectory to be interpolated will actually pass through the node. In addition to this, the more traffic that passes through a particular node, the more likely that the unknown trajectory will pass through that node.

Based on the above reasoning, we further extract the features of the C_k, including *time consumption*, *traffic flow*, O_i, D_i, and *travel mode*. The probability density function can be obtained using Kernel density estimation to fit the continuous features. Given a pair of origin and destination O, D to be interpolated, and its continuous features are denoted as f. The probability of the corresponding feature can be calculated by padding f. For O_i, D_i, we calculate the proportion of origin and destination in O_i, D_i that is within 1000m of O, D. We assume that each feature is mutually exclusive, the probability of selecting each next node can be obtained by multiplying the probability values of each feature. Also, we add positional constraints to ensure that the interpolation is valid and the overall direction of the interpolated path is consistent with the actual trajectory. A binary tree is constructed by selecting the top two nodes with the highest probability for each iteration. In the end, we obtain an OD-determined network due to the existence of loops. The probability of each path can be calculated based on the Markov assumption. At last, a depth-first search algorithm is applied to obtain the path with the highest probability as our interpolation path. Fig. 3 shows the process of trajectory interpolation. The pseudo-code is shown in Algorithm 1.

Algorithm 1: Trajectory interpolation algorithm

Data: start and end for the part to be interpolated S, E, origin and destination of the trajectory to be interpolate O, D, history trajectory C
Output: Interpolated trajectory T_i
preprocess C using K-means $\rightarrow N_k$;
Use KDE on each node to get $f.pdf$ for continuous features;
$n_{current} = S$;
$G \leftarrow S$;
while $n_{current} != E$ **do**
  for *each node in* $n_{current}.next$ **do**
    if *node meet constrains* **then**
      $P(node) = \prod^{f \in n.f} \int_{f-pad}^{f+pad} f.pdf(t)dt$;
      count the proportion of O_i, D_i within 1000m from $O, D \rightarrow P_{OD\cdot}$;
      $P(node) = P(node) \times P_{OD}$;
    normalize $P(node)$;
  identify the top two nodes in $P(node) \rightarrow n_1, n_2$;
  $n_{current} \leftarrow n_1, n_2$;
  $createGraph(G, n_{current})$;
calculate all possible route with start S and end E in $G \rightarrow Paths$;
for *each path in* $Paths$ **do**
  $P_i = \prod^{n \in path}(P(n))$;
$T_i = Paths[\arg\max(P)]$;

Graph constructing. We construct the TripGraph G_T to represent the network of trips between the specified origin and destination regions. When generating G_T, the input is C_k, the aggregated points of all CSD sampling points between origin and destination regions. The vertices of output graph G_T represent the cell station clustering centers, and the edges represent the links between two physically connected cluster centers. Here, edges are added when there exists at least one trip moving from a vertex to another vertex. What's more, three kinds of weight are appended to the edges, including the travel distance, the traffic volume, and the travel time. When an edge is added in the graph construction process, we calculate the traffic volume, average distance and average time by going through all the trajectories segments.

Path Analysis. To study the role of paths in traffic, we score the vertices by calculating the PageRank [43] for each vertex in graph G_T, which is a widely used metric in graph analysis. PageRank is originally developed to assess the significance of a web page on the internet. By utilizing this

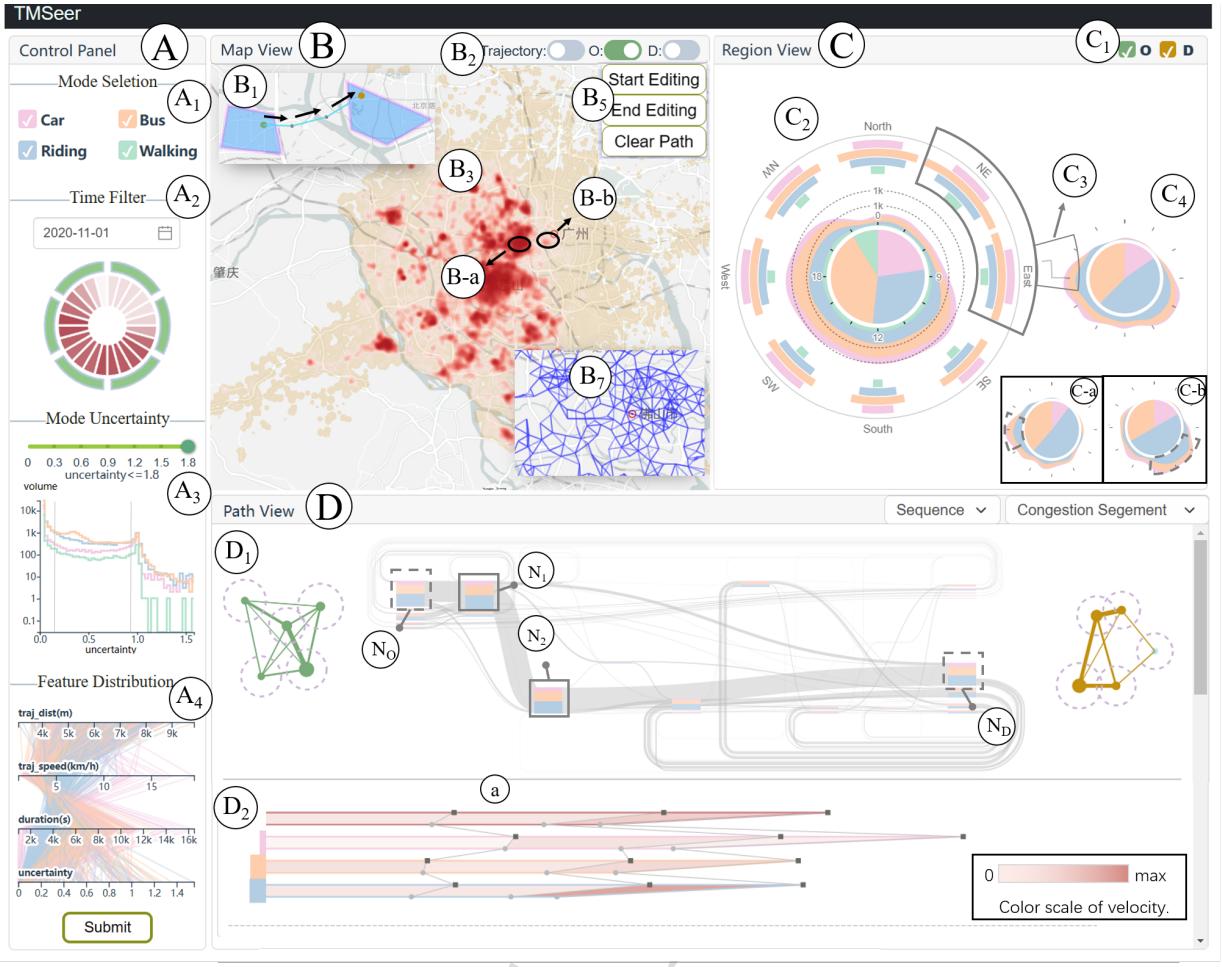

Figure 4: Our visualization system, *TMSeer*, supports exploring urban mobility with four different traffic modes, i.e., walking, riding, bus, and car. The control panel (A) shows the temporal distribution of traffic volume, the relationship between travel mode features, and supports trip filtering. The map view (B) shows the spatial distribution of traffic and provides operations for region selection. The region view (C) demonstrates the spatio-temporal mobility patterns in a focus region, and with other inter-regional traffic. The path view (D) provides a detailed exploration of inter-regional paths and the comparative analysis of paths.

method, the score for the importance of a segment in a path is calculated based on the principle that links to segments with high scores will raise the score more significantly compared to links to segments with low scores. If weight is based on travel distance, a high PageRank shows a hub in the path structure, which is the intersection of major roads. When weight is based on flow, a high score indicates a transportation hub that is actually used by traffic, which is the distribution center of traffic flow. Lastly, if the weight is based on travel time, a high PageRank shows congested segments. It helps to discover important or abnormal sections in the path.

5. Visualization Design

This section describes a set of visualization techniques that assist users in exploring city-level traffic and comparing different travel modes.

5.1. Control Panel

The control panel (Fig. 4A) is designed to help users to filter trips and provide cues for uncertainty (**R1.2**), facilitating better interaction with our system. At the top, a mode selector (Fig. 4A_1) is provided for users to select trips with modes they are interested in. Then a radial chart (Fig. 4A_2) like a clock presents the change in traffic volume for each hour of the day. Darker colors indicate more traffic. Users can filter the trip by clicking the ring of the corresponding period. In Fig. 4A_3, users can filter trips by the uncertainty. The line chart shows the distribution of uncertainty for all mode identification results, with the horizontal axis representing the uncertainty and the vertical axis representing the trip volume. Due to the wide range of data variation in the vertical axis, we use a logarithmic axis to achieve a better presentation. Also, we add two vertical gray lines to indicate the locations of the median and the 95th percentile, which can provide users with hints on choosing the appropriate uncertainty filtering threshold. At the bottom (Fig. 4A_4),

the parallel coordinate graph shows the relationship between the uncertainty of all mode identification results and feature distribution. The color of the lines indicates the travel mode. It helps users to analyze the identification results and infer possible travel modes. Moreover, it can change with the selected trips during the analysis procedure. After finishing the parameters setting, users can click the submit button to accomplish trip filtering.

5.2. Map View

TMSeer provides a spatial overview of the entire traffic in the city (**R2**) through the map view (Fig. 4B), which consists of three parts.

Firstly, a heat map (Fig. 4B_3) is provided for all the connected cell stations, showing the global traffic distribution. After discussion with experts, we provide two heat maps for origins and destinations separately, facilitating the traffic exploration of inflow and outflow.

Besides, experts recommend presenting more information about trajectories that heat maps cannot show. Thus, trajectories are represented by lines connecting cell stations and shown in the map view (Fig. 4B_7). The thickness of the lines indicates the volume of traffic flows. For better visual effects, we aggregate cell stations in close proximity and apply the edge bundling technique [28] to reduce visual clutters.

To assist visual linkage and smoother operations, we present path line (Fig. 4B_1) in map view to provide a spatial context for path analysis. We present each path with nodes and colored lines. To avoid visual confusion, different colors are used to represent different paths. We use two colors to represent origin (green) and destination (orange) points separately. Additionally, animation is added in order to show the continuous movements in the paths, which can effectively reduce the confusion caused by path overlaps. Users can choose to show or hide the graphs they want to explore by clicking the toggle buttons (Fig. 4B_2). Additionally, arbitrary polygon box selection (Fig. 4B_5) operations are supported for following region exploration.

5.3. Region View

After determining their region of interest from the map view, users want to further investigate the mobility pattern of the region (**R3**). The region view (Fig. 4C) is designed to meet the exploration requirements, which consists of two parts. We design a hybrid radial diagram (Fig. 4C_2) to provide the spatio-temporal distribution of different travel modes for a focus region (**R3.1**), complementing the inability of the map view. The region view also display preliminary mobility patterns between the focus region and surrounding region (**R3.2**). Users can choose multiple surrounding regions according to their own interests and make preferential choices among surrounding regions for subsequent detailed investigation. We use a simplified hybrid radial diagram (Fig. 4C_4) to show the temporal distribution of each travel mode, and apply a well-established layout which is based on the relative positions in the map to ensure a better visual effect.

Hybrid Radial Diagram Design. We use the hybrid radial diagram design (Fig. 5A) to show the mobility patterns of the central region. Four colors are employed to encode four travel modes throughout. To display residents' preference for travel mode in the central region, we use a pie chart (Fig. 5A-a) to show the proportion of travel modes and adopt a stack radial area chart to show temporal patterns of traffic mobility in the central region (Fig. 5A-b). For spatial patterns, we calculated the proportion of different travel modes in eight directions, the direction of each trajectory is represented using origin and destination direction. We use a radial bar chart in each corresponding direction and make it axis-symmetrical to avoid confusion and improve visual effect (Fig. 5A-c). For the surrounding region, we use the simplified hybrid radial diagram (Fig. 5B) to display the mobility between the central region and the surrounding region. Similar to Fig. 5A, we also use the pie chart and stack area chart to show the distribution of different travel modes and temporal patterns in the simplified diagram. The radius of each diagram represents traffic flows between the central region and the corresponding regions. Note that the simplified hybrid radial diagram (Fig. 5B) shows traffic flow between two regions while the central hybrid radial diagram (Fig. 5A) indicates all traffic flow with origins or destinations in the central region. Furthermore, users are allowed to configure the central region as "O" or "D" (Fig. 4C_4) to explore the inflow or outflow of the central area. Users can also select both origins and destinations to analysis the overall traffic flow including inflow and outflow.

Layout. To visually map the region view to the map view, we place each hybrid radial diagram based on its original orientation and use a line to connect the surrounding hybrid radial diagram to the central hybrid radial diagram (Fig. 4C_3). In this way, the layout of the region view is consistent with the actual positions of the regions in the map view, making it easy to be explored by users. We use a force-directed layout to maintain the relative position of each hybrid radial diagram while avoiding overlapping.

Alternative Designs. Before adopting the current central hybrid radial diagram design (Fig. 5A) and surrounding hybrid radial diagram design (Fig. 5B), we have considered other alternative designs correspondingly. For example, as for the central hybrid radial diagram design, we counted the flow of different modes of travel over eight separate distances in each of the four directions and used short lines to represent the flow of each mode of travel (Fig. 5C). However, after discussions with experts, it was felt that the information in the different directions was more important than the distance. The design was therefore discarded. As for the surrounding hybrid radial diagram design, we used a radial central symmetrical bar chart to show the traffic flow of different travel modes every two hours (Fig. 5D), but compared with Fig. 5B, it does not provide a visual representation of the temporal distribution of total traffic flow.

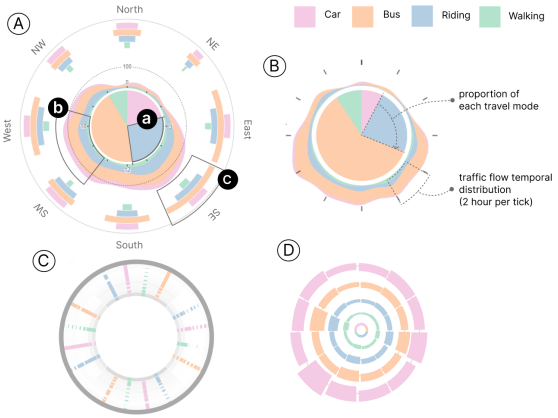

Figure 5: (A) The central hybrid radial diagram design for the central region shows the proportion of travel modes, temporal patterns, and the proportion of travel modes in different directions. (B) The surrounding hybrid radial diagram design, a simplified version, shows the proportion and temporal pattern between the central region and surrounding regions. (C) An alternative design for the central hybrid radial diagram. (D) An alternative design for surrounding hybrid radial diagram.

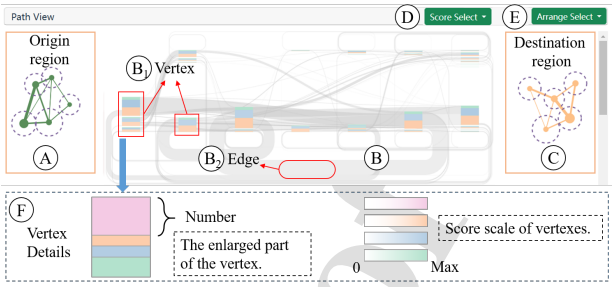

Figure 6: The path graph design in the path view. (A) and (C) show the movement in the origin and destination regions. (B) shows the movement between two regions. (D) provides layout selection for the enhanced Sankey diagram. (E) provides score selection for vertexes. (F) shows detailed vertex composition, including the number and score distribution of travel modes.

5.4. Path View

From the region view, users can explore the traffic modes between pairs of regions, and further analyze details of the paths from one region to another region (**R4.2**). We design the path view for path analysis, comprised of two parts. The path graph (Fig. 4D_1) above displays how each path is used by different travel modes. The bands (Fig. 4D_2) below show the travel efficiency of each path. After path extraction and vertex score calculation in Section 4.3, all the trips between two regions are constructed into a graph, representing the traffic network between the two regions. The score of each vertex is also calculated.

Path Graph. To support the analysis of the path usage for travel modes, a path graph is designed. As shown in Fig. 6, we divide the graph into three parts: a network on the left encoding the movement in the origin region (Fig. 6A), a network on the right encoding the movement in the destination region (Fig. 6C), and an enhanced Sankey diagram in the middle showing the movement connecting the two regions (Fig. 6B). For the origin and destination networks, the size of the vertex represents the traffic passing through, and the thickness of the edge represents the traffic moving between the nodes. The dashed circles indicate the approximate extent of the clusters. The location of the node represents the relative location of the region geographically. For the enhanced Sankey diagram, the rectangles of each vertex (Fig. 6B_1) show the volume of traffic passing through, with height representing the volume and color representing the travel mode. The color of rectangles indicates the scores mentioned in Section 4.5. The darker color shows a higher score for the vertex. What's more, users can switch scores calculated by different weights (Fig. 6E), which have different meanings. For example, in the selection of "hubs in structure", the nodes with darker colors are more likely to be a transportation hub on a geographic road. In the selection of "hubs in traffic", we show the hubs actually used by traffic. While in the selection of "congestion segments", the darker color indicates more congested areas. The thickness of the edges (Fig. 6B_2) indicates the traffic volume between two nodes, with color indicating the average duration spent on the path segment.

Layout. The path graph based on the enhanced Sankey diagram allows users to obtain an overview of paths connecting two regions. As the path graph is complex, we improve the layout of the enhanced Sankey diagram for better analysis. For the arrangement in the horizontal direction, we provide two kinds of options. First, travel is a kind of spatio-temporal sequence. So we first decide the position of nodes by the order of the trip. The vertex reached earlier is placed in the previous column. Second, the relative positions to the origin region can also make the layout clearer. We sort the nodes by the distance to the origin region. For the arrangement in the vertical direction, the nodes with more traffic are placed at the top. Users can select the layout (Fig. 6D) they want. In this way, the important information stands out.

Parallel Band. To better compare different paths, a parallel band design (Fig. 7) is proposed to show the travel efficiency of each path. We first provide a bar chart (Fig. 7A) on the left to show the traffic volume of different travel modes that are encoded by the corresponding mode colors. Then, we design bands (Fig. 7B) for efficiency comparison. Fig. 7C shows the enlarged part of the band. For each band, the above line with dark gray rectangle points on the band indicates the partitioning path segments, with the length showing the average distance of trajectories moving from the previous node to the next. The line with light gray circle points beneath the band displays the duration information. The length between two circles shows the time spent on the corresponding path segment. Also, the color of the quadrangle shows the corresponding average speed. The darker the color is, the higher the speed is. Finally, we place the band representing the summary of all trips at the top, and then the bands for different travel modes are arranged at the

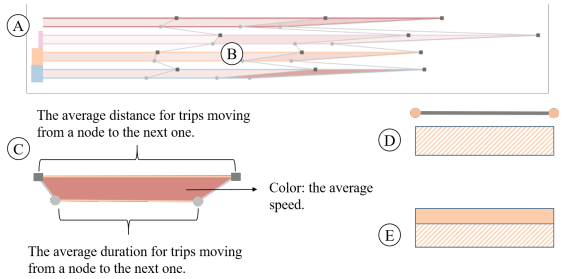

Figure 7: The parallel band design in the path view. (A) The bar chart left shows the volume of travel modes. (B) The band indicates travel efficiency by combining distance and duration. (C) The enlarged part of the band. (D, E) Two alternative designs.

bottom. In such a manner, the efficiency of all travel modes at different segments in the path is shown, including the speed, duration, and distance.

Alternative Designs. To compare the different paths, we considered several alternative designs for the parallel band design. As shown in Fig. 7D, we first considered using the length between nodes to represent the Euclidean distance between cell station clustering centers. Meanwhile, we used the height of the rectangle to represent time. We also used the texture to fill the rectangle, with the tilt to represent speed. However, the clustering of the cell station covered a certain space range, and the Euclidean distance cannot represent the actual distance. This could lead to the wrong information. Next, we tried to add another rectangle on top (Fig. 7E), with the height representing the relative actual distance. However, experts said it was not intuitive and the textures were confusing, making it difficult to capture useful information. Thus, we directly represent the true distance by the length, with additional points indicating nodes. We also use the depth of color instead of the texture to indicate speed, which is more intuitive.

5.5. Interactions

User interactions supported by *TMSeer* are summarized as follows.

Filtering. Rich filtering operations are supported in the control panel (Fig. 4A). Users can filter by travel mode, time period, and identification uncertainty.

Linking. Linking connects all views in the system. For example, the path view is generated by clicking the region of interest in the region view. Users can double-click the nodes of the enhanced Sankey diagram in the path view (Fig. 4D_1), and the corresponding paths will show below. Meanwhile, users can select a path to render it on the map view (Fig. 4B_1).

Highlighting. Highlighting is employed to help users focus on the information of interest. The selected regions will be highlighted in the map view (Fig. 4B_1). In addition, when users hover on the nodes or edges of the enhanced Sankey diagram, the connected nodes and edges will be highlighted. Besides, the same segment belonging to different travel modes in Fig. 7 will be highlighted in red lines when hovering on it.

6. Evaluation

We first conducted quantitative evaluations for the proposed mode identification method using a CSD dataset with ground truth labels annotated by volunteers, then assessed the effectiveness and usability of *TMSeer* through two case studies and interviews with two aforementioned collaborating experts (E1 and E2, who have been introduced in Section 3). Section 6.3 reports the cases conducted by E1 and E2 respectively and Section 6.4 reports the feedback from all experts. The data used in this section is one-day CSD covering the trajectories of 244,928 users for the whole day.

6.1. Mode Identification

This paper evaluates the effectiveness of the approach using a test set collected by volunteers. The test set includes data collected by 10 volunteers over three days and a taxi over one month. We applied the methods introduced in Section 4 to clean the raw data, extract individual trips, and identify travel modes. After preprocessing, the final dataset contains 24 walking samples, 88 bus samples, and 270 car samples, resulting in a total of 382 trajectories used for evaluation.

Baseline and metrics. We implemented the method proposed by Chen et al. [9] as the baseline and compare our model to it, noting that they exclusively employed GMM. We used accuracy, precision, recall, and F1 score to assess the performance of our method. To further verify the effectiveness of the uncertainty assessment method, we evaluated the mode identification results in different uncertainties. Here we refer to the visual cues in Section 5.1, to filter the results and verify the validity of our method. Specifically, we filtered the results according to the 95th percentile and the median and count the indicators separately.

Evaluation results. The left subfigure of Fig. 8 shows the overall accuracy of baseline and our method. Compared with the baseline, the accuracy of our methods is 63.74%, which is higher. The RBH+GMM provides a performance boost. The incorporation of RBH method can avoid some errors. In addition, compared with all results, the accuracy can be effectively improved by filtering trajectories based on the uncertainty according to the system hints. At last, by using the median of uncertainty as the threshold to filter the mode identification results, our approach can achieve an accuracy of 71.73%. This proves that our method can assist to filter more reliable results based on uncertainty, which facilitating the entire analysis process.

The right subfigure of Fig. 8 displays the F1 score of our methods on each travel mode. It is shown that our method performed well on car and bus modes among all the travel modes in terms of the F1 score. However, it is relatively more difficult to infer the walking mode. As shown in the confusion matrix (Table 2), our method sometimes misclassified walking as bus, since two of the volunteers' walking

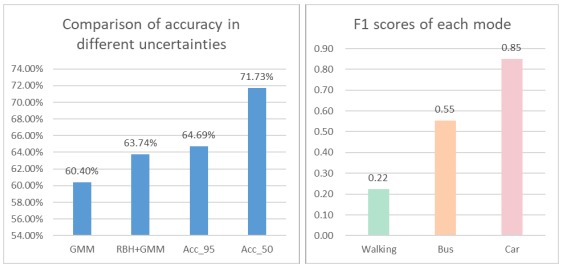

Figure 8: The evaluation results of mode identification. Left: comparison of the classification accuracy of all modes. Right: F1 scores of each transport mode.

Table 2
Confusion matrix of mode identification results.

Actual \ Predicted	Walking	Bus	Car	Total	Recall
Walking	3	21	0	24	12.5%
Bus	15	65	8	88	73.86%
Car	3	61	206	270	76.30%
Total	21	147	214	382	
Precision	14.29%	44.22%	96.26%		

trajectories were collected near bus stations. In general, our method performs well.

6.2. Trajectory Interpolation

To evaluate the interpolation performance, we construct a test set based on the CSD trajectories collected from both volunteers and a taxi. Sparsity is simulated by randomly removing a portion of location records from each trajectory, while preserving the complete version as ground truth for evaluation. Each trajectory spans no less than 10 minutes in duration, and a total of 8,650 trajectories are extracted for testing. This approach enables the creation of realistic sparse trajectory data while ensuring accurate benchmarks for comparison.

Baselines. Two existing models are implemented as baselines, as shown below.

1) Hidden Markov Model (HMM): HMM is a classical probabilistic approach used to model serial data and is widely used to estimate and predict the position during human movement.

2) DeepTransport [49]: DeepTransport is a method using LSTM to infer the detailed trajectory of mobile phone users, and reconstructs the trajectory by feeding individual mobility records to the model.

Metrics. We use great-circle distance, i.e., the length of the shortest path from one point on the sphere to another point on the sphere, to measure the estimated error between the true position and the interpolated position. The calculation formula is as follows:

$$\alpha = \sin^2(\Delta\phi/2) + \cos\phi_1 \cdot \cos\phi_2 \cdot \sin^2(\Delta\lambda/2), \tag{4}$$

$$C = 2 \cdot atan2(\sqrt{\alpha}, \sqrt{1-\alpha}), \tag{5}$$

Table 3
Evaluation results of trajectory interpolation method.

Model	Average Error (m)
HMM	304.35
DeepTransport	182.6
Our method	**163.17**

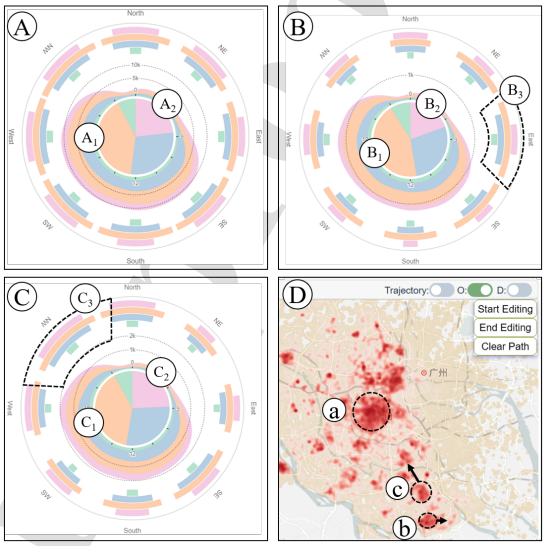

Figure 9: The traffic patterns in different regions. (A) The main travel mode in Region a is the bus (orange). (B) Region b has more bus trips (orange) and fewer car trips (pink), and more traffic flow in the east. (C) Region c has fewer bus trips and more car trips, compared with b, and more traffic flow in the northwest. (D) Selected hot spot regions.

$$\text{dist} = R \cdot C. \tag{6}$$

where ϕ denotes the latitude, λ is the longitude, and R denotes the radius of the earth.

Evaluation results. The average error of different methods is summarized in Table 3. Our proposed method achieves the lowest average error of **163.17** meters, outperforming both HMM (304.35 m) and DeepTransport (182.6 m). This demonstrates the effectiveness of our approach in trajectory interpolation. However, we find that some individual interpolation results deviate from the overall error distribution with larger error value. Through the analysis found that this is because the limited by historical trajectories. After analysis, we found that the historical trajectory provides rich movement information, but also has limitations. Filtering of historical trajectories may lead to a small scale and an increasing sparsity of the filtered trajectories, resulting in unsatisfactory interpolation result. In general, our method is superior to the two baseline models.

6.3. Case Study

6.3.1. Case I: Explore travel modes in different regions

One of E1's daily works is road capacity demand analysis, which needs to analyze the traffic in different areas. So E1 tried to use *TMSeer* to explore travel modes in different regions. First, he filtered trips for exploration in the control panel. As he wanted to explore all travel modes in different regions, so he selected all checkboxes (Fig. 4A_1) for all travel modes exploration. Then he observed that the color change in some sectors was not obvious in the radial chart (Fig. 4A_2), indicating that the traffic volume did not change much during the daytime. So he selected the all-day traffic. Furthermore, he checked the uncertainty distribution of the travel mode recognition results in Fig. 4A_3, which is important for the analysis. As shown in Fig. 4A_3, the number of trips dropped sharply near the uncertainty of 1.0. The trips with uncertainty within 1.0 had a better mode identification effect. Therefore, for a more accurate analysis, he chose the trips for the full day with an uncertainty of less than 1.0. After finishing the configuration, he clicked the "submit" button, then the corresponding views were updated. In the map view, after activating the origin and destination heatmap by clicking the corresponding toggle buttons (Fig. 4B_2), he immediately found several hot spots with more traffic (Fig. 9D).

As Region-a (Fig. 9D-a) was the densest in the city, he selected Region-a for exploration and found that it was the downtown area. After selecting Region-a, the corresponding hybrid radial diagram showing travel patterns within Region a was generated in the region view (Fig. 9A). He selected both "O" and "D" checkboxes (Fig. 4C_4), for analyzing all trips that moved in and moved out of the region. As shown in Fig. 9A_1, it was observed that the orange area representing the bus accounted for the largest proportion, indicating the number of trips taking the bus was the largest. In addition, the outer bars with the same color in different directions were of similar length. E1 thought that Region-a in the downtown attracted the people around it more evenly.

To explore the patterns of travel modes in the suburbs around Region-a (Fig. 9D-a), thus he further selected another hot spot, Region-b (Fig. 9D-b) for further analysis. Figure 9B showed the corresponding hybrid radial diagram summarizing the travel modes in Region-b. He was surprised to find that the orange area Fig. 9B_1 was larger than that in Fig. 9A_1 while the pink area (Fig. 9B_2) was smaller than Fig. 9A_2. It indicated that there were more people using the bus rather than using the car. In addition, the bars on the right side (Fig. 9B_3) were significantly longer. It showed that more traffic flows were moving to or coming from the east. He went back to the position of Region-b in the heat map (Fig. 9D-b) and noticed a small hot spot in the east of Region-b, which was not far. E1 explained that the long distance from Region-b to the city center may cause the movement of people in the suburban area to be more concentrated nearby. "In short-distance travel, cost-effective and convenient buses are more popular," commented E1.

Finally, he wanted to explore whether the suburbs have the same patterns. He found another hot spot, Region-c (Fig. 9D-c), which was near Rregion-b and was also a suburban area. Fig. 9C showed the pattern of travel modes in Region-c. In contrast to Region-b, although it was also close to the suburban area, Region-c had a different pattern for travel modes. It can be seen that the proportion of the orange area in Fig. 9C_1 was smaller than that in Fig. 9B_1 and the pink area (Fig. 9C_2) was larger, indicating fewer trips using bus and more trips using cars. To explain this question, he examined the other parts of the hybrid radial diagram carefully. Immediately, he observed in Fig. 9C_3 that the bar in the upper left was longer, indicating more traffic in the northwest. He backed to Fig. 9D-c and found that the northwest of Region-c is downtown. Perhaps because Region-c was a little bit closer to the downtown than Region-b, so there was more traffic to the downtown rather than nearby, he explained. In his view, the advantage of bus travel diminished as the distance increased. So the change in bus and car proportion in Region-c might result from the long distance to the downtown compared to Region-b (Fig. 9B), which had more trips nearby. He commented, "Measures need to be taken in Region-b, to optimize transportation structure and to promote the usage of the bus. Such as optimizing bus routes and implementing preferential policies."

6.3.2. Case II: Analyze paths based on uncertainty

As E2's daily work was devoted to the design and retrofit of the road, she was interested in using *TMSeer* to analyze the paths with different modes in Foshan City. First of all, she selected four travel modes, all time periods, and did not conduct any uncertainty filtering in the control panel (Fig. 4A), to explore the complete travel pattern throughout the day. After trip filtering, she further explored the map view for an overview of traffic.

As E2 was familiar with this city, she had always been interested in the hot spot region shown in Fig. 4B-a. E2 mentioned that, as a suburban area, there were many people working in another neighboring city. So there was a lot of cross-city travel, which is prone to cause traffic congestion and electric bicycle accidents all year round. Thus, she selected Fig. 4B-a as the first region. Then she chose an area of about the same size across the bridge on the east side as the second region (Fig. 4B-b) because it is common for residents to commute to another city. After the two regions were selected, the region view generated the corresponding hybrid radial diagrams, showing the patterns of travel modes in the first region (Fig. 4C_2) and the traffic between two regions (Fig. 4C_4).

As shown in Fig. 4C_4, she first noticed that the bus (orange) and riding (blue) were the main travel modes between the two regions. Then she immediately pointed out the obvious rise in the area chart in morning and evening rush hours. She then switched the "O" and "D" checkboxes to investigate inflow and outflow separately and found out that outflow had its peak hours in the morning (Fig. 4C-a) while inflow has its peak hours in the evening (Fig. 4C-b),

which was in line with people's commuting characteristics. When she clicked the small hybrid radial diagram (Fig. 4C_4), the parallel coordinate graph (Fig. 4A_4) in the control panel changed to show the feature distribution of selected trips. The path view (Fig. 4D) was also updated. From the parallel coordinate graph (Fig. 4A_4), she discovered that the duration of the bus (orange) and car (pink) usually lasted longer, and the speed of some riding trips (blue) was higher than the normal bicycle. Thus she guessed that riding here was more likely to include the electrical bicycle.

Furthermore, E2 turned to the path view (Fig. 4D) for path analysis. She first checked the two node-link diagrams on both sides, which represented the traffic flow within origin and destination regions. She found some big nodes with thick edges, indicating more traffic volume here. She said, "It might be backbones that many traffic had to pass through for travel." Then she turned to the enhanced Sankey diagram to explore the paths connecting the origin and destination regions. She tried different selections of layouts and scores for the enhanced Sankey diagram. Of all the rectangles representing nodes, she saw two particularly conspicuous nodes (Nodes N_1 and N_2 in Fig. 4D) with the darker colors in the enhanced Sankey diagram, indicating that these nodes were most likely to be hubs and congestion areas. To find the reason for the congestion, she selected a pair of origin and destination nodes with the most traffic (Nodes N_O and N_D in Fig. 4D) passing through two nodes for further analysis.

As shown in the band graph (Fig. 4D_2), E2 first observed the highest flow on riding (blue) in the bar chart on the left. Then she went through each band. The first two segments were lighter in color than the last, indicating a lower speed. This was consistent with the scores shown in the enhanced Sankey diagram. What's more, the line indicating the duration of riding was obviously shorter than others, especially in the second segment. It showed that riding could be more efficient than buses or cars. She then clicked the path to see its location in the map view (Fig. 4B_1), which corresponded to the Guangfo Road. "In fact, traffic jams occur here all year round and are in urgent need of control," said E2.

She explained that because there were so many riding trips, it could affect the efficiency of vehicles to some extent. "Measures need to be implemented to achieve the separation of motor and non-motor vehicles, which can effectively improve travel efficiency and reduce traffic accidents," said E2.

Through this exploration process, she recognized our system, which she felt was very meaningful to understanding the traffic patterns of citizens and was helpful to her work.

6.4. Expert Interview

We conducted one-on-one interviews with domain experts (E1 and E2) and collected their feedback after the case studies. The feedback is summarized as follows.

Visual Design. The experts confirmed that our system is well-designed and user-friendly. They believed that the system could be easily understood by users with different backgrounds. In particular, both E1 and E2 praised the hybrid radial diagram design in region view. "A summary of the travel modes in a region is clearly given," said E1. E2 added, "The function of selecting multiple regions facilitates comparative analysis." Besides, the experts also acknowledged the smoothness of interactions.

Usability and Improvements. The experts praised our system and thought the functions provided by *TMSeer* were really helpful. E1 highly praised that *TMSeer* provides an opportunity to comprehensively analyze the travel modes in the city. E1 commented, "It is low-cost and efficient, which can be a strong reference for urban planning." E2 indicated that path analysis can help to find areas or sections that need to be improved. Apart from the aforementioned, our experts also provided some valuable suggestions to improve the usability of the system. E1 mentioned that medium- and long-term analysis of transportation planning also played an important role, such as in the past decade. "We hope to use it to predict future traffic changes," he suggested. E2 further suggested a more concise visual presentation in the path view, "It will help a lot when making reports."

7. Discussion

In this section, we reflect on the implications of our work, discuss the generalizability of *TMSeer*, and outline current limitations and future research directions.

Implications. This study leverages large-scale CSD for the city-level analysis of travel modes, which aims to gain deep insights into the overall travel patterns of city-level crowds. *TMSeer* comprises novel visualizations and travel mode identification methods to facilitate the exploration of multiple travel modes. The case studies show that our approach can reveal insightful patterns of different modes that can help urban planners and policy-makers with their decision making of urban planning and other relevant policies. Our approach marks an initial effort to apply data visualization and visual analytics techniques in support of urban planning and decision-making, laying the groundwork for future research in this direction.

Generalizability. *TMSeer* is designed for the analysis of traffic patterns for travel modes using CSD. However, the analytical pipeline and system architecture are also adaptable to other types of mobility data. For instance, GPS data, despite offering finer spatial granularity, share a similar spatiotemporal structure with CSD—such as trajectories, origin-destination pairs, and speed profiles. This compatibility allows *TMSeer* to be extended to GPS data with minimal modifications in the preprocessing phase, such as adjusting for sampling intervals or spatial resolution. Nevertheless, due to privacy concerns and access restrictions, large-scale GPS datasets at the city level are not publicly available and typically require coordination with government agencies, which limits our current ability to evaluate this adaptation empirically. In addition, smart card data can be incorporated for analyzing public transit flows, offering another potential application scenario. Beyond transportation analysis, *TMSeer* can also assist in business site selection, as different

travel modes influence commercial activity in distinct ways. Understanding mobility patterns around candidate locations can support decisions such as facility placement and parking allocation. Furthermore, the visualization designs in *TMSeer* are applicable to a variety of domains beyond urban mobility. For example, the hybrid radial diagram is well-suited for presenting cyclical or time-based patterns in other fields, while the enhanced Sankey diagram can be generalized for analyzing complex flows and relationships in supply chains, migration studies, and energy systems.

Trajectory Interpolation. Trajectory interpolation in CSD is necessary, since the intrinsic sparsity of CSD affects path analysis and sometimes introduces errors during graph construction. Our aim is to better infer the route of the target movement, rather than the actual location of the connected cell station. In this paper, we interpolate the cellular signaling data based on the historical trajectories. We assume that: 1) The same travel mode usually has a common movement pattern when passing through a local area; 2) The majority of people usually have similar movement patterns between a pair of origin point and destination point. To ensure scalability and real-time interaction, we selectively interpolate only a subset of trajectories during the visual analysis phase. This approach is also guided by practical considerations: interpolating all CSD records would be computationally expensive, and in some cases could distort key features (e.g., duration or frequency) relevant to travel mode identification. Our strategy thus balances analytical accuracy with system performance.

Limitations and Future work. While our evaluation demonstrates the effectiveness of *TMSeer*, several limitations remain that open up opportunities for future research. First, the current analysis primarily focuses on traffic volume and speed. Additional factors such as income levels, travel cost, land use, weather conditions, and real-time congestion are all critical for a more comprehensive understanding of travel behaviors. Integrating these dimensions would enhance the interpretability and decision-making value of the system, and we plan to explore them in future work. Second, uncertainties remain in travel mode identification due to the relatively coarse spatial resolution of CSD, which can limit recognition accuracy compared to high-resolution GPS data. Nevertheless, our enhanced unsupervised method, combined with uncertainty assessments, has substantially improved the reliability of the results. During expert evaluation, the visual outputs were found to be both intuitive and practically informative. Third, this study is currently limited to one-day CSD data, which constrains longitudinal insights. In future work, we plan to incorporate multi-day or long-term datasets to analyze differences in travel patterns across weekdays, weekends, and special events, enabling broader temporal evaluations of urban mobility. Fourth, the absence of an open, city-scale CSD dataset with mode labels limits direct benchmarking and cross-study comparability. Recent CSD-based travel-mode studies still rely on operator/self-collected data, while the publicly available benchmarks are predominantly GPS-based, which differ in sensing modality and sampling from CSD. As future work, we plan to curate a de-identified subset (subject to compliance) and conduct cross-modality validation against public GPS-based corpora. Finally, as large language models (LLMs) such as GPT and DeepSeek continue to advance, we envision their integration into *TMSeer* as intelligent assistants that can help domain experts interpret patterns, formulate hypotheses, and derive insights from complex mobility data. Such integration could enhance the system's interactivity, explanation capability, and accessibility for non-technical users.

8. Conclusion

In this study, we propose *TMSeer*, an interactive visual analytics system to assist the analysis of traffic patterns for multiple travel modes. We apply large-scale CSD to explore multiple travel modes in the city, achieving the analysis of overall traffic patterns. The proposed method can infer the most likely travel mode adopted by a trip and assess the uncertainties of the inferred travel modes. Intuitive visualizations together with coordinated views and interactions enable multi-level exploration of large-scale cellular data. We conduct quantitative evaluations, case studies and expert interviews to demonstrate the effectiveness and usability of *TMSeer*. In the future, we will extend *TMSeer* to support long-term cellular signaling data exploration, and incorporate more factors, such as land use patterns, income, and cost, to enable a more comprehensive analysis of city-level travel behaviors.

Acknowledgments

We would like to thank our domain experts and the anonymous reviewers for their insightful comments. This work is supported by grants from the National Natural Science Foundation of China (No. 62302531) and the Science and Technology Planning Project of Guangdong Province (No. 2023B1212060029).

References

[1] Aguiléra, V., Allio, S., Benezech, V., Combes, F., Milion, C., 2014. Using cell phone data to measure quality of service and passenger flows of paris transit system. Transportation Research Part C: Emerging Technologies 43, 198–211.

[2] Alhumoud, S., 2025. Analysis of transportation patterns through call detail records (cdrs). PLoS One 20, e0330246.

[3] Aziz, H.A., Nagle, N.N., Morton, A.M., Hilliard, M.R., White, D.A., Stewart, R.N., 2018. Exploring the impact of walk–bike infrastructure, safety perception, and built-environment on active transportation mode choice: a random parameter model using new york city commuter data. Transportation 45, 1207–1229. doi:https://doi.org/10.1007/s11116-017-9760-8.

[4] Bai, T., Li, X., Sun, Z., 2017. Effects of cost adjustment on travel mode choice: Analysis and comparison of different logit models. Transportation Research Procedia 25, 2649–2659. doi:https://doi.org/10.1016/j.trpro.2017.05.150.

[5] Baudains, P., Holliman, N.S., 2025. Visual summaries of traffic congestion with uncertainty: Exploring street network distances and vehicle orientation. Computers & Graphics , 104228.

[6] Chakraborty, S., Kiefer, P., Raubal, M., 2024. The influence of uncertainty visualization on cognitive load in a safety-and time-critical decision-making task. International Journal of Geographical Information Science 38, 1583–1610.

[7] Chen, E., Zhang, W., Ye, Z., Yang, M., 2020. Unraveling latent transfer patterns between metro and bus from large-scale smart card data. IEEE Transactions on Intelligent Transportation Systems 23, 3351–3365. doi:10.1109/TITS.2020.3035719.

[8] Chen, J., Huang, Q., Wang, C., Li, C., 2023. Sensemap: Urban performance visualization and analytics via semantic textual similarity. IEEE Transactions on Visualization and Computer Graphics 30, 6275–6290.

[9] Chen, J., Xiong, C., Cai, M., 2022. A travel mode identification framework based on cellular signaling data. Mobile Information Systems 2022. doi:https://doi.org/10.1155/2022/2113213.

[10] Chen, S., Yuan, X., Wang, Z., Guo, C., Liang, J., Wang, Z., Zhang, X., Zhang, J., 2015. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. IEEE Transactions on Visualization and Computer Graphics 22, 270–279. doi:10.1109/TVCG.2015.2467619.

[11] Chin, K., Huang, H., Horn, C., Kasanicky, I., Weibel, R., 2019. Inferring fine-grained transport modes from mobile phone cellular signaling data. Computers, Environment and Urban Systems 77, 101348. doi:https://doi.org/10.1016/j.compenvurbsys.2019.101348.

[12] Chiou, Y.C., Hsieh, C.W., 2021. Travel pattern analytics driven by cellular signaling data. Asian Transport Studies 7, 100042.

[13] Chondrogiannis, T., Bornholdt, J., Bouros, P., Grossniklaus, M., 2022. History oblivious route recovery on road networks. Prod.SIGSPATIAL doi:https://doi.org/10.1145/3557915.3560979.

[14] Clark, H.H., 1969. Linguistic processes in deductive reasoning. Psychological Review 76, 387. doi:https://doi.org/10.1037/h0027578.

[15] Deng, Z., Weng, D., Liang, Y., Bao, J., Zheng, Y., Schreck, T., Xu, M., Wu, Y., 2021. Visual cascade analytics of large-scale spatiotemporal data. IEEE Transactions on Visualization and Computer Graphics 28, 2486–2499. doi:10.1109/TVCG.2021.3071387.

[16] Deng, Z., Weng, D., Liu, S., Tian, Y., Xu, M., Wu, Y., 2023. A survey of urban visual analytics: Advances and future directions. Computational Visual Media 9, 3–39.

[17] Ding, F., Zhang, Y., Peng, J., Ge, Y., Qu, T., Tao, X., Chen, J., 2024. A hybrid method for intercity transport mode identification based on mobility features and sequential relations mined from cellular signaling data. Computer-Aided Civil and Infrastructure Engineering .

[18] Dong, J., Zhang, H., Cui, M., Lin, Y., Wu, H.Y., Bi, C., 2024. TCEVvis: Visual analytics of traffic congestion influencing factors based on explainable machine learning. Visual Informatics 8, 56–66.

[19] Du, J., Rakha, H.A., Breuer, H., 2022. An in-depth spatiotemporal analysis of ride-hailing travel: The chicago case study. Case Studies on Transport Policy 10, 118–129. doi:https://doi.org/10.1016/j.cstp.2021.11.010.

[20] Ersoy, O., Hurter, C., Paulovich, F., Cantareiro, G., Telea, A., 2011. Skeleton-based edge bundling for graph visualization. IEEE Transactions on Visualization and Computer Graphics 17, 2364–2373. doi:10.1109/TVCG.2011.233.

[21] Ewing, R., Hamidi, S., 2015. Compactness versus sprawl: A review of recent evidence from the united states. Journal of Planning Literature 30, 413–432. doi:https://doi.org/10.1177/0885412215595439.

[22] Feng, Z., Jiang, Y., Wang, H., Fan, Z., Ma, Y., Yang, S.H., Qu, H., Song, X., 2024a. Trafps: A shapley-based visual analytics approach to interpret traffic. Computational Visual Media 10, 1101–1119.

[23] Feng, Z., Zhu, F., Wang, H., Hao, J., Yang, S.H., Zeng, W., Qu, H., 2024b. HoLens: A visual analytics design for higher-order movement modeling and visualization. Computational Visual Media 10, 1079–1100.

[24] Ferreira, N., Poco, J., Vo, H.T., Freire, J., Silva, C.T., 2013. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. IEEE Transactions on Visualization and Computer Graphics 19, 2149–2158. doi:10.1109/TVCG.2013.226.

[25] Franke, M., Barczok, R., Koch, S., Weltecke, D., 2019. Confidence as first-class attribute in digital humanities data, in: Proc.VIS4DH Workshop, p. 5.

[26] Gundlegård, D., Rydergren, C., Breyer, N., Rajna, B., 2016. Travel demand estimation and network assignment based on cellular network data. Computer Communications 95, 29–42.

[27] Holliman, N.S., Coltekin, A., Fernstad, S.J., McLaughlin, L., Simpson, M.D., Woods, A.J., 2019. Visual entropy and the visualization of uncertainty. arXiv preprint arXiv:1907.12879 doi:https://doi.org/10.48550/arXiv.1907.12879.

[28] Holten, D., 2006. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. IEEE Transactions on visualization and computer graphics 12, 741–748.

[29] Huang, X., Zhao, Y., Ma, C., Yang, J., Ye, X., Zhang, C., 2015. Trajgraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. IEEE Transactions on Visualization and Computer Graphics 22, 160–169. doi:10.1109/TVCG.2015.2467771.

[30] Huang, Y., Wang, D., Zhang, M., Zeng, J., Cai, Z., 2025. Transfer learning-based approach for fine-grained travel mode identification on mobile phone signalling data. Transportmetrica A: Transport Science , 2522992.

[31] Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J., 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences 622, 178–210.

[32] Jiang, Z., Huang, A., Qi, G., Guan, W., 2023. A framework of travel mode identification fusing deep learning and map-matching algorithm. IEEE Transactions on Intelligent Transportation Systems 24, 6401–6415.

[33] Kao, D., Luo, A., Dungan, J.L., Pang, A., 2002. Visualizing spatially varying distribution data, in: Proc.IV, IEEE. pp. 219–225. doi:10.1109/IV.2002.1028780.

[34] Kim, S., Jeong, S., Woo, I., Jang, Y., Maciejewski, R., Ebert, D.S., 2017. Data flow analysis and visualization for spatiotemporal statistical data without trajectory information. IEEE transactions on visualization and computer graphics 24, 1287–1300.

[35] Li, G., Xu, R., Shi, T., Deng, X., Liu, Y., Di, D., Zhao, C., Liu, G., 2024a. Fine-grained metro-trip detection from cellular trajectory data using local and global spatial–temporal characteristics. ISPRS International Journal of Geo-Information 13, 314.

[36] Li, H., Wang, Y., Qu, H., 2024b. Where are we so far? understanding data storytelling tools from the perspective of human-ai collaboration, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–19.

[37] Liu, Y., Liao, F., Wang, W., Wang, Y., Chen, J., 2025. An integrated method for inferring multimodal travel mode choices using mobile network data. Transportation Research Part C: Emerging Technologies 179, 105305.

[38] Ma, T., Knaap, G.J., 2019. Estimating the impacts of capital bikeshare on metrorail ridership in the washington metropolitan area. Transportation Research Record 2673, 371–379. doi:https://doi.org/10.1177/0361198119849407.

[39] McQuaid, R.W., Greig, M., Smyth, A., Cooper, J.A., 2003. The importance of transport in business location decisions-scoping study. .

[40] Miranda, F., Ortner, T., Moreira, G., Hosseini, M., Vuckovic, M., Biljecki, F., Silva, C.T., Lage, M., Ferreira, N., 2024. The state of the art in visual analytics for 3d urban data, in: Computer Graphics Forum, Wiley Online Library. p. e15112.

[41] Moreira, G., Hosseini, M., Veiga, C., Alexandre, L., Colaninno, N., de Oliveira, D., Ferreira, N., Lage, M., Miranda, F., 2024. Curio: A dataflow-based framework for collaborative urban visual analytics. IEEE Transactions on Visualization and Computer Graphics .

[42] Nahmias-Biran, B.h., Sharaby, N., Shiftan, Y., 2014. Equity aspects in transportation projects: Case study of transit fare change in haifa. International Journal of Sustainable Transportation 8, 69–83. doi:https://doi.org/10.1080/15568318.2012.758525.

[43] Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The PageRank citation ranking: Bringing order to the web. Technical Report. Stanford infolab.

[44] Peng, Z., Bai, G., Wu, H., Liu, L., Yu, Y., 2021. Travel mode recognition of urban residents using mobile phone data and mapapi. Environment and Planning B: Urban Analytics and City Science 48, 2574–2589. doi:10.1177/2399808320983001.

[45] Prelipcean, A.C., Gidófalvi, G., Susilo, Y.O., 2017. Transportation mode detection–an in-depth review of applicability and reliability. Transport Reviews 37, 442–464. doi:https://doi.org/10.1080/01441647.2016.1246489.

[46] Qu, Y., Gong, H., Wang, P., 2015. Transportation mode split with mobile phone data, in: Proc.ITSC, IEEE. pp. 285–289. doi:10.1109/ITSC.2015.56.

[47] Senaratne, H., Mueller, M., Behrisch, M., Lalanne, F., Bustos-Jiménez, J., Schneidewind, J., Keim, D., Schreck, T., 2017. Urban mobility analysis with mobile network data: A visual analytics approach. IEEE Transactions on Intelligent Transportation Systems 19, 1537–1546. doi:10.1109/TITS.2017.2727281.

[48] Shen, Z., Yang, K., Zhao, X., Zou, J., Du, W., Wu, J., 2024. Dmm: A deep reinforcement learning based map matching framework for cellular data. IEEE Transactions on Knowledge and Data Engineering 36, 5120–5137.

[49] Song, X., Kanasugi, H., Shibasaki, R., 2016. Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level, in: Proceedings of the twenty-fifth international joint conference on artificial intelligence, pp. 2618–2624.

[50] Stenneth, L., Wolfson, O., Yu, P.S., Xu, B., 2011. Transportation mode detection using mobile phones and gis information, in: Proc.SIGSPATIAL, pp. 54–63. doi:https://doi.org/10.1145/2093973.2093982.

[51] Wang, H., Zhang, Z., Fan, Z., Chen, J., Zhang, L., Shibasaki, R., Song, X., 2023. Multi-task weakly supervised learning for origin-destination travel time estimation. IEEE Transactions on Knowledge and Data Engineering , 1–14doi:10.1109/TKDE.2023.3236060.

[52] Wang, X., Jiao, S., Bryan, C., 2024. Defogger: A visual analysis approach for data exploration of sensitive data protected by differential privacy. IEEE Transactions on Visualization and Computer Graphics .

[53] Wei, Z., Ding, S., Xu, W., Fang, J., Liu, C., Wang, Y., 2024. Rea-fm: automated generation of natural-looking flow maps through river extraction algorithm. Cartography and Geographic Information Science 51, 761–781.

[54] Weng, D., Zheng, C., Deng, Z., Ma, M., Bao, J., Zheng, Y., Xu, M., Wu, Y., 2020. Towards better bus networks: A visual analytics approach. IEEE Transactions on Visualization and Computer Graphics 27, 817–827. doi:10.1109/TVCG.2020.3030458.

[55] qing Wu, H., Mao, J., Sun, W., Zheng, B., Zhang, H., Chen, Z., Wang, W., 2016. Probabilistic robust route recovery with spatio-temporal dynamics. Proc.SIGKDD doi:https://doi.org/10.1145/2939672.2939843.

[56] Wu, L., Wang, W., Jing, P., Chen, Y., Zhan, F., Shi, Y., Li, T., 2020. Travel mode choice and their impacts on environment—a literature review based on bibliometric and content analysis, 2000–2018. Journal of Cleaner Production 249, 119391. doi:https://doi.org/10.1016/j.jclepro.2019.119391.

[57] Wu, W., Xu, J., Zeng, H., Zheng, Y., Qu, H., Ni, B., Yuan, M., Ni, L.M., 2015. Telcovis: Visual exploration of co-occurrence in urban human mobility based on telco data. IEEE Transactions on Visualization and Computer Graphics 22, 935–944. doi:10.1109/TVCG.2015.2467194.

[58] Wu, W., Zheng, Y., Cao, N., Zeng, H., Ni, B., Qu, H., Ni, L.M., 2017. Mobiseg: Interactive region segmentation using heterogeneous mobility data, in: Proc.PacificVis, IEEE. pp. 91–100. doi:10.1109/PACIFICVIS.2017.8031583.

[59] Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., et al., 2008. Top 10 algorithms in data mining. Knowledge and information systems 14, 1–37.

[60] Zeng, H., Shu, X., Wang, Y., Wang, Y., Zhang, L., Pong, T.C., Qu, H., 2020. Emotioncues: Emotion-oriented visual summarization of classroom videos. IEEE Transactions on Visualization and Computer Graphics 27, 3168–3181. doi:10.1109/TVCG.2019.2963659.

[61] Zhang, W., Tan, S., Chen, S., Meng, L., Zhang, T., Zhu, R., Chen, W., 2022. Visual reasoning for uncertainty in spatio-temporal events of historical figures. IEEE Transactions on Visualization and Computer Graphics doi:10.1109/TVCG.2022.3146508.

[62] Zhao, X., Zhang, Y., Hu, Y., Wang, S., Li, Y., Qian, S., Yin, B., 2020. Interactive visual exploration of human mobility correlation based on smart card data. IEEE Transactions on Intelligent Transportation Systems 22, 4825–4837. doi:10.1109/TITS.2020.2983853.

[63] Zheng, Y., Wu, W., Chen, Y., Qu, H., Ni, L.M., 2016. Visual analytics in urban computing: An overview. IEEE Transactions on Big Data 2, 276–296.

[64] Zhong, S., Chen, J., Cai, M., 2024. A transport mode detection framework based on mobile phone signaling data combined with bus gps data. Mathematics (2227-7390) 12.