Real-Time Driving State Identification and Collision Risk Detection in Dump Trucks: A GPS Streaming Data Approach

Chengbin Xie[1], Ye Zheng[2], Yang Zhao[1], Libo Zhang[3], Leyang Zhao[1], Weixi Wang[1], Xiaoming Li[1,*]

[1] Research Institute for Smart Cities, School of Architecture and Urban Planning, Shenzhen University & State Key Laboratory of Subtropical Building and Urban Science & Guangdong–Hong Kong-Macau Joint Laboratory for Smart Cities, Shenzhen 518060, China - xiechengbin2022@email.szu.edu.cn, 1615540462@qq.com, zhaoleyang@szu.edu.cn, 359912469@qq.com
[2] Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, 315211, China – zhengye@nbu.edu.cn
[3] National Quality Inspection and Testing Center for Surveying and Mapping Products, Beijing 100830, China – 37856062@qq.com
* Correspondence - lixming@szu.edu.cn

Keywords: GPS Streaming Data of Dump Trucks; Real-Time Analysis; Distributed Computing.

Abstract

Real-time analysis and mining of vehicle GPS data is essential for effective traffic regulation. Currently, most vehicle analysis algorithms based on GPS data are designed for static datasets, with fewer algorithms addressing dynamic streaming GPS data. Moreover, the limited number of real-time algorithms primarily focus on public transportation vehicles such as taxis and buses. There remains a significant gap in the analysis of specialized vehicles like dump trucks, which fails to meet the regulatory needs for monitoring these vehicles. To address this, this paper proposes a method for identifying driving states and detecting collision risks for dump trucks based on real-time GPS stream data. First, by partitioning the data, the method enables the separate calculation and identification of various operational states of different vehicles, such as their location, speed, and direction. Second, we partition the data based on vehicle positions to detect potential collision risks among vehicles in nearby areas. Experimental results show that the data throughput reaches 25,000 and 66,000 records per second for each method, with a data skew rate controlled below 0.1, demonstrating the method's efficiency in real-time driving state recognition and collision risk detection for dump trucks.

1. Introduction

In recent years, with the continuous development of software and hardware technologies, an increasing number of sensors have been deployed in public infrastructure to collect real-time information on equipment operation (Li et al., 2020). The constantly generated data, containing both temporal and spatial information, form spatiotemporal streams (Brandt et al., 2018). These highly time-sensitive spatial streams provide a foundational data support for real-time monitoring and analysis across various industries and fields, enabling more timely and informed decision-making (Li et al., 2022; King and Osborn, 2023; Chen et al., 2023).

To effectively analyse spatial streaming data, fully unlock its potential, and achieve the aforementioned goals, many researchers have conducted extensive research in this area. The two main research directions focus on extending support for spatiotemporal data in streaming data processing platforms and improving algorithms for more efficient data processing. In the area of extending spatiotemporal data support on streaming platforms, many researchers have carried out foundational work using platforms such as Flink and Spark, including defining fundamental spatial data types like points, lines, and polygons, and constructing spatial indexes. Additionally, a series of basic streaming operations have been developed, including spatial queries and spatial joins, which have significantly improved the real-time computation efficiency of spatial streams (Yu et al., 2019; Shaikh et al., 2022). In terms of improving algorithms for efficient data processing, researchers have designed various streaming algorithms for different application scenarios. For instance, some researchers have developed methods to detect anomalies in bus trajectories and missed stops based on real-time GPS data (Zou et al., 2023). To address the widespread challenge of storing large volumes of GPS trajectory data, others have designed and implemented a trajectory compression method, which enables efficient compression of large-scale GPS data in parallel computing environments (Xiong et al., 2023). Furthermore, for the detection and management of hazards involving large numbers of commercial vehicles or ships, researchers have developed methods to monitor moving entities and identify complex patterns (Ntoulias et al., 2021).

However, vehicle analysis and monitoring have predominantly focused on static data analysis, aiming to identify driving patterns by analysing vehicle driving data (Jeon et al., 2019; Zhao et al., 2020). The real-time analysis primarily focuses on public transportation vehicles such as buses and taxis. (Agrawal et al., 2018; Li et al., 2019; Mestre et al., 2023). And the real-time analysis of such public transportation vehicles mainly focuses on whether their movement trajectories are reasonable, which cannot fully cover the functions required for the regulation of dump trucks. Therefore, there are still many gaps in the research on regulatory algorithms for dump trucks. This paper designs and implements an efficient real-time computation system for the massive stream of dump truck GPS data based on the Flink streaming platform. By analysing the real-time driving states of dump trucks, the system identifies potential violations such as speeding and unauthorized entry into restricted areas, and provides warnings in dangerous areas when trucks pass each other. Finally, this enables efficient and real-time monitoring of dump trucks.

2. Method

Figure 1 illustrates the overall framework of the proposed method. The data processing workflow is divided into two main stages: driving state recognition and collision risk detection. The data source consists of an unbounded stream of real-time generated GPS data records for dump trucks. Each GPS record in the stream primarily includes the vehicle identification code, vehicle location, and the timestamp of the GPS record. When

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-975-2025

GPS data from dump trucks is generated, it is continuously consumed in real time by the computing cluster as a GPS data stream. To efficiently process and analyse this data, a pre-processing step is first conducted to convert the data into a format suitable for analysis and to filter out erroneous data. Subsequently, the data is partitioned based on the unique vehicle identification code using a hash partitioning strategy. This ensures that all records with the same vehicle identification code are sent to the same parallel task. Next, according to the requirements for driving state and collision risk detection, each vehicle’s driving state is calculated in real time, and potential collision risks among multiple vehicles are assessed. Finally, the results are aggregated to produce the final output. The specific algorithmic details for driving state recognition and collision risk detecting are provided in Sections 2.1 and 2.2.

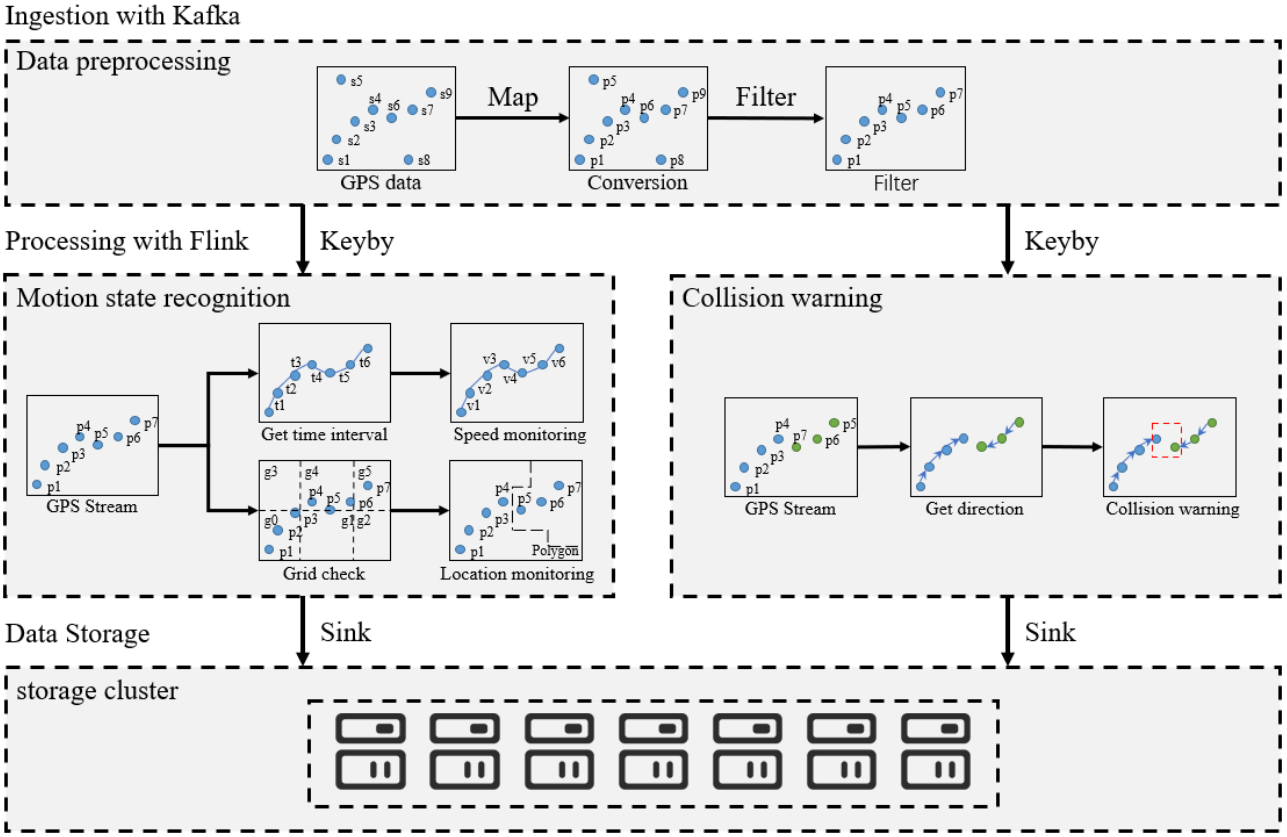

Figure 1. The overall framework of the methodology.

2.1 Driving States Identification

2.1.1 Speed Anomaly Detection Algorithm. The specific steps for detecting speed anomaly are as follows: When a new record is consumed by the computing cluster, the data is first logically partitioned based on the unique vehicle identification code. In a distributed environment, the data is then divided and processed separately on different computing nodes. For each individual vehicle, separate identification operations are performed. When new record arrives, the first step is to assess the time difference between it and the previous record. For instance, if the time interval between two records exceeds one minute , it indicates that the correlation between these two records is weak, and the new record does not need to be processed; if the time interval is within one minute, it proceeds to the next step of the processing flow. the next step is to determine the distance between the two GPS locations and calculate the average speed of the dump truck to assess if it is speeding. Due to possible errors in GPS data, it is necessary to exclude clearly erroneous calculations during the average speed computation. We used Haversine formula to calculate the distance. The Haversine formula is presented as follows:

$$a = sin^2(2\Delta\varphi) + cos(\varphi 1) \cdot cos(\varphi 2) \cdot sin^2(2\Delta\lambda) \quad (1)$$

$$d = 2R \cdot \arcsin(\sqrt{a}) \quad (2)$$

Where φ1 and φ2 are the latitudes of the two points
Δφ is the difference in latitude between the two points
Δλ is the difference in longitude between the two points
R is the radius of the Earth
d is the distance between the two points

2.1.2 Location Anomaly Detection Algorithm. The specific steps for location anomaly are as follows: First, based on the temporal characteristics of the restricted areas for construction trucks, the GPS timestamps are checked to determine if they fall within the restricted time periods, using this temporal attribute to filter the data. Second, restricted zones can be categorized into two main types: administrative area restrictions and road restrictions. Administrative area restrictions are typically larger in scope, covering extensive areas with multiple vector endpoints, while road restrictions are smaller in size and require higher precision. According to the characteristics of these two types of restricted areas and to ensure both the effectiveness of the algorithm's recognition and computational efficiency, different precision recognition methods are used. Specifically, an index set of basic spatial units that intersect with the vector restricted areas is obtained in advance. The basic spatial units

are continuous, non-overlapping square regions that cover the study area, similar to a fishing net. The index numbers start from the southwestern corner of the study area and increase from west to east. Upon reaching the eastern boundary, the indexing moves to the next row and continues to increase from west to east. In this study, the size of each basic spatial unit is 500m × 500m. During the evaluation process, for administrative area restrictions, determining whether the basic spatial unit containing the GPS data is included in the above index set. In contrast, for road restriction areas, a vector-based method will be used to evaluate the spatial relationship between the points and the restricted zones. By utilizing time, location, and the different characteristics of restricted areas for data filtering, this multi-layered screening mechanism significantly reduces the number of vector evaluations, thereby greatly enhancing the operational speed of the cluster. Figure 2 illustrates the methods for detecting location anomalies in different types of restricted areas.

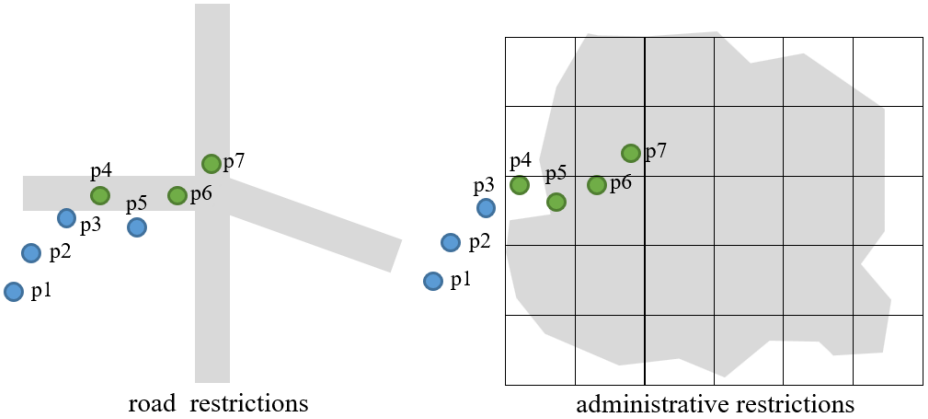

Figure 2. The methods for detecting location anomalies in different types of restricted areas.

2.2 Collision Risk Detection

The specific algorithmic steps for the collision risk detection algorithm are as follows: data is logically partitioned based on the unique vehicle identification code to calculate the driving direction of each vehicle separately. Vehicle information and driving direction data are then further partitioned according to the basic spatial units they located in. Vehicles moving in opposite directions within the same basic spatial units are compared using a dual-stream traversal comparison method. By calculating the distances between these vehicles, potential collision risks can be identified by determining whether the distances are less than 50 meters. Figure 3 presents the collision risk detection algorithm.

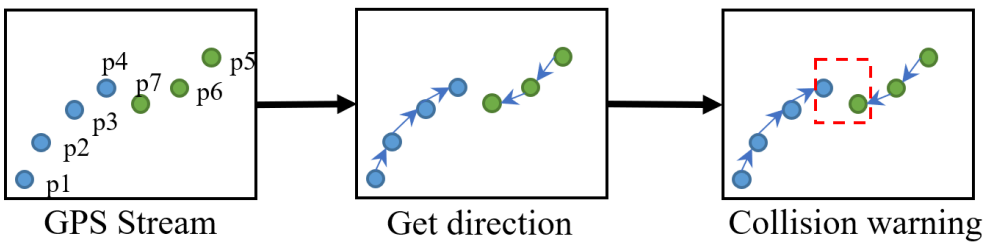

Figure 3. The collision risk detection algorithm.

3. Experiments

3.1 Test Datasets

The study area is defined as the minimum bounding rectangle of Shenzhen, ranging from 113.751647°E to 114.622924°E longitude, and from 22.446379°N to 22.855425°N latitude. The test dataset consists of dump truck GPS data within this area, totalling 53,385,954 records and approximately 3.36 GB in size. The data covers a three-day period from May 2 to May 4, 2020, and it primarily includes the vehicle's longitude and latitude information, vehicle ID, and the recorded timestamp. Table 1 presents the detailed attribute information of the data. Figure 4 shows the study area.

Attribute Name	Data Structure Type	Specific Meaning
TRACK_ID	NUMBER	Point id
VEHICLE_ID	NUMBER	Vehicle id
LNG	NUMBER	Longitude
LAT	NUMBER	Latitude
GPS_TIME	DATE	GPS Recording Time

Table 1. Data Attribute Table.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-975-2025 | © Author(s) 2025. CC BY 4.0 License.

Figure 4. Study area.

3.2 Experimental result and Algorithm Performance Analysis

3.2.1 Analysis of Speed Anomaly Detection Results. From a temporal perspective, the data volumes on May 2 and May 3 reached 50,564 and 56,534 records, respectively, while May 4 showed a sharp decrease to 11,341 records. This substantial difference likely stems from the fact that May 2 and May 3 fell on the weekend, whereas May 4 was a Monday. On weekends, particularly in the early morning or late-night hours, traffic volume often drops significantly. In these less congested conditions, dump truck drivers may be more inclined to exceed speed limits to enhance transportation efficiency, resulting in the observed high rate of speeding incidents, with weekend speeding accounting for 90.4% of cases. Table 2 provides speeding information according to time.

	Total	May 2nd	May 3rd	May 4th
Frequency	118,442	50,564	56,534	11,341
Percentage	100	42.7	47.7	9.6

Table 2. Speeding information according to time.

In terms of the locations where speeding occurred, most instances took place on highways, expressways, and main city roads. This aligns with the objective conditions typically conducive to speeding. Figure 5 illustrates the locations where speeding incidents occurred in Yantian District. The left depicts the situation on weekends, while the right represents the scenario on weekdays.

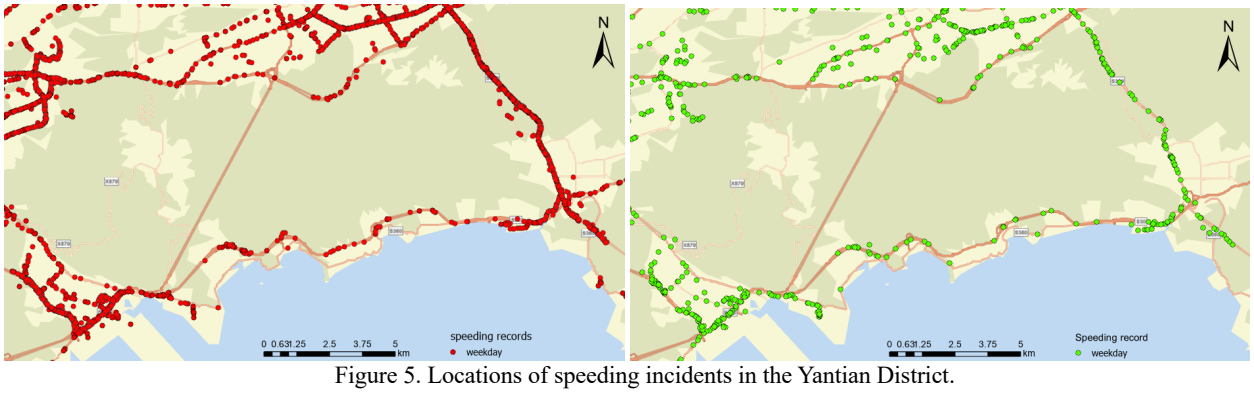

Figure 5. Locations of speeding incidents in the Yantian District.

3.2.2 Analysis of Location Anomaly Detection Results. After comprehensively considering the relationship between the accuracy of the algorithm and computational efficiency, this study opted to combine basic spatial units that intersect with restricted areas and vector-based assessments for determining whether the record is in the restricted zones. The experiments get a total of 1,414,689 abnormal data records. Among these, 57,976 records were located within basic spatial units but outside the polygon restricted areas, accounting for approximately 4%. In comparison, if all data were evaluated using vector-based methods, the processing time for a single data tuple would increase by about six times. Conversely, if solely basic spatial units were used for assessment, the number of records located within these units but outside the restricted areas would rise to 297,084, approximately 21%. Therefore, the other two approaches either had a higher misjudgement rate or

struggled to maintain operational efficiency. Consequently, not relying entirely on vector-based assessments remains a more suitable approach. Table 3 provides comparison of efficiency of location anomaly detection algorithms.

	Vector only	Spatial Unit only	Both
processing time for a single data tuple (ms)	0.135	0.007	0.029
Accuracy (%)	100	79	96

Table 3. Efficiency of location anomaly detection algorithm.

3.2.3 Analysis of Collision Risk Detection Results. From the analysis of collision risk results, the distribution of collision risk results was spread throughout the study area. The identification results are particularly meaningful in locations characterized by rugged mountain roads, with many sharp turns and poor visibility conditions. Providing potential collision risk alerts to drivers of large vehicles, such as dump trucks, in these areas can significantly reduce the possibility of accidents and enhance driving safety. Figure 6 presents the collision risk analysis results for certain provincial roads in Dapeng New District.

Figure 6. Collision risk points for provincial roads.

3.2.4 Analysis of Algorithm Performance. In the performance evaluation system of real-time computing systems, throughput and data skewness are critical metrics. Throughput refers to the ability to process data records per unit of time and serves as an important benchmark for measuring the system's processing efficiency and capacity. Data skewness indicates the uniformity of data distribution across different computing nodes in a parallel computing model. By assessing these two metrics, we can more objectively evaluate the processing effectiveness of real-time computing tasks, especially when dealing with large-scale, high-concurrency data streams.

The overall throughput is calculated by dividing the total amount of data consumed by the distributed computing cluster by the duration of the program's execution. The single-node throughput is determined by dividing the amount of data consumed by a single Slot in the distributed cluster by the program's runtime. Data skewness is represented by the coefficient of variation of the data consumed by each Slot. Specifically, it is the standard deviation of the data consumed by each computational node in the cluster divided by its own average. The detailed calculation formulas and values are presented in Table 4.

	Overall throughput (records/s)	Maximum single-node throughput(records/s)	Minimum single-node throughput (records/second)	Data skewness
Formula	throughput = (data volume)/time			std(data volume) / avg(data volume)
Driving states identification algorithm	25120	2855	2125	0.074
Collision risk Detection Algorithm	66164	5023	3562	0.093

Table 4. Performance evaluation metrics.

Overall, both algorithms achieved a throughput of over ten thousand records per second, which can effectively support the real-time analysis of large-scale GPS streams. Additionally, the data skewness was maintained below 0.1, indicating a relatively even distribution of data across the computing nodes, thereby decreasing the computational power loss caused by the "bottleneck effect."

https://doi.org/10.5194/isprs-annals-X-G-2025-975-2025

4. Conclusions

We achieved real-time analysis of the driving states of dump trucks based on GPS data and developed a collision risk detecting method, evaluating the results and performance using real data. As a contribution, we established a real-time vehicle status analysis algorithm suitable for streaming real-time computation, based on distributed stream computing technology and incorporating the principles of streaming real-time processing and classical algorithms. During the experimental process, we used a multi-layer filtering mechanism and a method that combines basic spatial units with vector-based judgments replaced only basic spatial units or vector-based judgments, effectively enhancing the efficiency of the analysis.

However, there is still room for improvement in the proposed algorithm. In the process of identifying restricted areas using uniform-sized basic spatial units instead of vectors, only units of the same size were used. While adjusting the size of these basic spatial units can directly influence the precision of the analysis, it may also impact the efficiency of the algorithm to some extent. Utilizing variable-sized or multi-level spatial units could improve this issue.

Acknowledgements

This work has been partly supported by the National Natural Science Foundation of China (Grant No.42471441, 42201449), Shenzhen Science and Technology Program (Shenzhen Key Laboratory of Digital Twin Technologies for Cities) (Grant Number ZDSYS20210623101800001) and Zhejiang Provincial Natural Science Foundation of China (Grant No. LQ23D010005)

References

Agrawal, S., Sonbhadra, S. K., Agarwal, S., 2018. Favour prediction of Taxi services using real-time visualization. In 2018 *International Conference on Advances in Computing, Communications and Informatics (ICACCI)* 2276-2282.

Brandt, T., Grawunder, M., 2018. GeoStreams: A survey. *ACM Computing Surveys (CSUR)* 51(3): 1-37.

Chen, Z., Cong, G., Aref, W. G., 2020. STAR: A distributed stream warehouse system for spatial data. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* 2761-2764.

Jeon, W., Zemouche, A., Rajamani, R., 2019. Tracking of vehicle motion on highways and urban roads using a nonlinear observer. *IEEE/ASME transactions on mechatronics* 24(2): 644-655.

King, L., Osborn, W., 2023. Ensemble Methods for Spatial Data Stream Classification. *Procedia Computer Science* 224: 155-162.

Li, C., Liu, Y., Zhang, H., 2019. Analysis of taxi track data based on spark platform. *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)* 2357-2361.

Li, W., Liu, Y., Wang, S., 2022. Real-time GIS Programming and Geocomputation. *Geographic Information Science & Technology Body of Knowledge* 2022(Q1).

Li, W., 2020. GeoAI: Where machine learning and big data converge in GIScience. *Journal of Spatial Information Science* 2020 (20): 71-77.

Mestre, D. G., Nóbrega, T. P., Araújo, T. B., et al., 2023. Efficient Map-Matching Parallelization over Bus Trajectories Using Spark. *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web* 238-245.

Ntoulias, E., Alevizos, E., Artikis, A., et al., 2021. Online trajectory analysis with scalable event recognition. *EDBT/ICDT Workshops*.

Shaikh, S., Kitagawa, H., Matono, A., et al., 2022. GeoFlink: An Efficient and Scalable Spatial Data Stream Management System. *IEEE Access* 10: 24909-35

Xiong, W., Wang, X,. Li, H., 2023. Efficient Large-scale GPS trajectory compression on spark: a pipeline-based approach. *Electronics* 12(17): 3569.

Yu, J., Zhang, Z., Sarwat, M., 2019. Spatial data management in apache spark: the geospark perspective and beyond. *GeoInformatica* 23: 37-78.

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

Zou, Q., Xiong, W., Wang, X., et al., 2023. Research on Real-Time Anomaly Detection Method of Bus Trajectory Based on Flink. *Electronics* 12(18): 3897.