IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

ENHANCING TIMELINESS IN URBAN HOTSPOT ANALYSIS WITH MULTI-SOURCE DATA FUSION

[1]Dr.P. Hariharan,, [2]N. Vasumathi

[1]Assistant Professor, [2] M.Phil., (CS) Research Scholar,
[1]Department of Computer Science and Applications,
[1]Adhiparasakthi College of Arts and Science (Autonomous), G.B. Nagar, Kalavai-632506,
Ranipet District, Tamil Nadu, India

Abstract: This study focuses on enhancing urban hotspot analysis and functional area recognition through the application of multi-source data fusion. However, the study acknowledges limitations concerning the timeliness of microblog check-in data. Specifically, due to the inability to segment the microblog check-in data based on release time periods, there is a delay in data analysis, impacting the timeliness of the results. Future research should focus on addressing this limitation by implementing real-time data capture and processing techniques to reduce the time lag and improve the timeliness of the microblog check-in data. By doing so, the accuracy and effectiveness of urban hotspot analysis and functional area recognition can be significantly enhanced within the proposed multi-source data fusion framework.

IndexTerms **– Point of Interest, Natural Language Processing, Support Vector Machines, Hotspot Analysis, Sampling.**

I. INTRODUCTION

Urban hotspot analysis, which identifies areas of high activity that contribute to traffic congestion, economic growth, and overall urban development, is crucial for urban planning, transportation management, and resource allocation. The accuracy and breadth of the insights gained have historically been hampered by the analysis's dependence on a single data source. Multi-source data fusion approaches, which combine data from several sources such Point of Interest (POI) data, GPS trajectories, social media check-ins, and public transportation data, are being used by researchers to get around these restrictions. With this method, urban hotspots may be understood more comprehensively. To be effective in urban planning and decision-making, it is imperative to address the timeliness issue because the delay in microblog check-in data processing has real-world effects.

With an emphasis on eliminating the latency in microblog check-in data using real-time data gathering and processing approaches, the proposed research intends to improve the timeliness of urban hotspot analysis by utilizing multi-source data fusion. This strategy seeks to increase the accuracy of hotspot identification, hasten resource allocation, and urban planning by reducing the time lag between data availability and hotspot updates. In order to provide a thorough picture of urban activities, the project's scope includes the integration of numerous data sources, including POI data, GPS trajectories, microblog check-ins, and social media data. This is accomplished through the use of data fusion techniques like preprocessing and weighted fusion, as well as strategies like class weighting and sampling that address data imbalance. The effectiveness of the approach will be assessed using evaluation metrics like accuracy, precision, recall, and F1 score, maybe through real-world trials, while real-world applications include emergency response systems, transportation management, and marketing of tourism.

Multi-source data fusion has tremendous promise for improving the timeliness of microblog check-in data in a variety of fields, including urban planning, traffic control, tourism, emergency response, retail, and smart city efforts. Stakeholders can make proactive, data-driven decisions that promote more efficient and sustainable urban development by utilizing real-time insights. By offering a thorough framework for quick and reliable urban hotspot identification, this initiative hopes to develop urban management techniques.

1.1 EXISTING SYSTEM

The merging of multiple location data sources, such as GPS trajectory data, check-in data, and Point of Interest (POI) data, is a crucial part of the current approaches for mining urban hotspots. These techniques use data preparation to clean and standardize the data before using it for analysis. Using feature extraction methods, relevant spatial and temporal attributes are recovered from each data source. Weighted fusion techniques combine data from several sources to create a comprehensive fused dataset. The next step is to identify urban hotspots using clustering techniques like K-means or DBSCAN, depending on the spatial closeness of the data. The outputs are evaluated for accuracy and contrasted with benchmark datasets or actual data. However, these techniques might encounter issues with data inequalities, a lack of timeliness, challenging data fusion processes, scalability, and generalizability.

1.2 DRAWBACKS

The existing multi-source location data fusion method for mining urban hotspots, which combines information from GPS trajectories, check-ins, and Point of Interest (POI) records, has a number of flaws. Different data volumes among sources provide data imbalance difficulties, which could result in biased analysis. The freshness of hotspot findings can be impacted by processing delays, which compromises timeliness, especially with check-in data. Scalability issues prevent the incorporation of new data sources or the adaptation to shifting urban dynamics, and the complex fusion process comprising clustering, feature extraction, and preprocessing poses difficulties. The system's restricted generalizability also limits the range of urban contexts and datasets to which it may be applied. To improve the accuracy, use, and feasibility of urban hotspot mining approaches, these issues must be addressed.

1.3 PROPOSED SYSTEM

By making microblog check-in data more current, the proposed method aims to provide real-time insights and enable data-driven decision-making for urban hotspot analysis and functional area discovery. The system includes a real-time data collecting approach to continually capture microblog check-in data from social media platforms and reduce the time lag in data availability. It preprocesses the data, extracts relevant timestamps and position information, and integrates it with other datasets like GPS trajectory data and POI data in order to create a comprehensive multi-source data repository. Stream processing and parallel processing are two methods that shorten processing times and efficiently handle the continuous flow of real-time data. The technology employs clustering algorithms to categorize check-ins into metropolitan hotspots based on their spatial proximity. In order to comprehend how hotspot dynamics vary over time, it also searches for temporal trends. Functional area recognition also makes use of feature extraction to differentiate between various urban regions based on their unique characteristics. By comparing the results with actual data, the system evaluates accuracy. It also provides insights via interactive dashboards and visualizations for straightforward understanding and decision-making.

The system under discussion aims to efficiently collect and analyze microblog check-in data to deliver real-time information on urban hotspots and beneficial locations. Through data integration, preprocessing, and spatial-temporal analysis, it provides precise and timely information for a number of applications, including urban planning, traffic management, tourism promotion, emergency response, and more. Participants in the system can feel confident that their decisions will effectively optimize urban growth and resource allocation because the system lays a heavy emphasis on real-time data processing and presentation.

Through the merging of data from multiple sources, urban hotspot analysis can be more timely. Urban planners and decision-makers obtain timely information to make wise decisions about resource allocation, population expansion, and traffic management. A thorough picture of urban activities is revealed by combining several datasets, such as social media, POI, GPS, and microblog check-in data, which increases the accuracy of hotspot analysis. Reduced data imbalance enables unbiased hotspot analysis, and improved geographical and temporal resolution records complex urban processes. The method is adaptable and scalable, allowing for the inclusion of new data sources and fostering real-world applications in urban planning, emergency response, tourism, and transportation administration. Overall, combining data from multiple sources enables effective decision-making and promotes sustainable urban growth.

1.4 MULTI-SOURCE DATA FUSION CLASSIFICATION

Multi-source data fusion classification with multi-perspectives improves classification accuracy and decision-making by utilizing data from a range of sources and viewpoints. In order to create a larger and more comprehensive dataset, this method integrates a number of datasets from various sources, such as sensor data, social media, and satellite imaging. The system is able to create a more complete understanding of the topic phenomena by combining several views. Each data source presents a different angle or understanding of the phenomenon.

The fusion process uses data preparation, feature extraction, and fusion techniques to effectively merge data from diverse sources. Different fusion techniques, such as weighted averaging, decision-level fusion, or feature-level fusion, can be applied depending on the characteristics of the data and the classification target. The enhanced dataset is then used to train machine learning models or classification algorithms, which increases the accuracy of the classification process. Through the use of multiple perspectives, the multi-source data fusion classification approach is able to deal with data incompleteness, uncertainty, and ambiguity, producing results that are more robust and reliable for classification across a variety of applications, including object recognition, anomaly detection, disease diagnosis, and environmental monitoring.

1.5 OBJECTIVES OF MULTI-SOURCE DATA FUSION CLASSIFICATION

Improve Classification Accuracy: By merging data from many sources and taking advantage of each viewpoint's advantages, the major objective is to improve classification task accuracy and produce more reliable and solid results.

Handle Data Uncertainty and Incompleteness: The technique aims to address uncertainties and data incompleteness present in individual data sources by integrating complementary data from several perspectives.

Improving Decision-Making: Multi-source data fusion categorization seeks to improve informed and data-driven decision-making by merging information from several sources.

Increase Robustness: The objective is to construct a more robust classification system by lowering the impact of data noise and outliers through multi-perspective fusion.

II. MULTI-SOURCE DATA FUSION CLASSIFICATION TECHNIQUES

Data preparation, which involves crucial processes to guarantee the quality and compatibility of data from diverse sources, is a crucial stage in multi-source data fusion. Data cleaning and normalization are included to improve data integrity by removing noise, outliers, and missing values, respectively. Power or logarithmic transformations, for example, can assist lessen the impact of high values and enhance data distribution. This pretreatment makes ensuring that the data is homogeneous, standardized, and prepared for additional analysis and data fusion.

The goal of feature extraction is to extract the most pertinent features from each data source that actually aid in the classification process. Dimensionality reduction techniques like PCA or t-SNE are used to lower feature dimensionality, improve feature fusion performance, and decrease computing complexity by extracting important and distinguishing features. Fusion algorithms are essential for merging data from various sources to produce a coherent and comprehensive dataset. Weighted averaging gives each data source a distinct weight based on how important it is to the fusion process, whereas Bayesian techniques employ probability distributions to deal with uncertainty and differing sources' credibility. The precision of the fusion and the dependability of the classification findings are impacted by the approach selection.

The multi-source data fusion technique relies heavily on machine learning models, which employ fused data to accomplish classification tasks. The ability of Support Vector Machines, Random Forests, and Neural Networks to handle complicated and high-dimensional data makes them good options. Decision tree predictions are combined in Random Forests, hidden layers in Neural Networks reveal intricate relationships, and SVM builds hyperplanes to divide classes in high-dimensional areas. The nature of the classification problem, the demand for accuracy, and the requirement for generalization all influence the choice of machine learning models. When training machine learning models using merged data, classification performance is improved and reliable decision-making based on full insights from several angles is made possible.

2.1 Multi-Source Data Fusion Classification Data Input and Output Types

2.1.1 Data Inputs:

The multi-source data fusion categorization approach accepts a number of datasets from different sources, each of which provides different information. These data inputs may include:

1. Sensor Readings: Sensor data from IoT devices or monitoring systems, such as temperature, humidity, pressure, or motion data, can be used to gather real-time information about the environment.

2. Images: Image data from cameras or satellite imaging captures visual information that can be utilized for object recognition, anomaly detection, and land cover classification.

3. Text: Text data from sources like social media posts, news articles, and customer reviews can provide sentiment and context information.

4. Social Media Posts: Information from social media, particularly user-generated content from sites like Twitter, Facebook, or Instagram, offers insights into the mindset, trends, and events of the general population.

5. Numerical Values: Quantitative data, such as financial summaries, numerical measurements, or economic indicators, can be used to present quantitative information.

2.1.2 Data Output:

The multi-source data fusion classification process produced the fused dataset. The data from all the different sources have been thoroughly synthesized into this dataset. The merged dataset includes the following characteristics:

1. Combined Information: To ensure that all relevant data viewpoints are taken into consideration, the fused dataset combines information from many sources.

2. Feature-Enhanced: To extract special properties, feature extraction methods have been applied to each data source. The fused dataset has these relevant properties, which enhances classification performance as a whole.

3. Pre-processed and ready for classification: The obtained dataset has been made ready for classification tasks. Its organizational structure enables machine learning models to process and assess the data in an efficient manner.

4. Broad view: The fused dataset offers a larger and deeper view than the individual data sources. To improve the overall robustness and accuracy of the classification process, it makes use of the advantages of each source.

A variety of data inputs are used in the multi-source data fusion categorization approach, including sensor readings, pictures, text, social media posts, and numerical values from various sources. The outcome is a fused dataset that incorporates information from many points of view and is intended for classification problems. Combining data from several sources enhances the decision-making process by enabling a more full and informed understanding of the target phenomenon.

2.2 Multi-Source Data Fusion Classification Data Source Types

The multi-source data fusion categorization approach makes use of a variety of data sources, each of which offers a unique viewpoint on the subject under study. These sources include sensors that produce real-time environmental observations, satellite imagery that provides a visual understanding of the Earth's surface, social media platforms that provide information on societal trends, web scraping that extracts data from online sources, public databases that provide structured information, and historical records that provide an understanding of past trends. A comprehensive dataset is produced as a result of the distinctive qualities that each source brings. By combining real-time monitoring, visual context, insights into public opinion, diverse content types, structured data, and historical trend analysis, this amalgamation facilitates correct classification. Multiple data sources are combined, which improves classification accuracy, encourages data-driven decision-making, and finds use across fields.

2.3 Multi-Source Data Fusion Classification Data Fusion Scales

Three main scales, each concentrating on different levels of data integration and abstraction, are involved in data fusion. Fusion at the feature level is integrating derived features from several sources to produce a complete feature set. This is especially helpful when sources have distinctive qualities that improve categorization precision. By combining classification judgments from several models or sources, decision-level fusion increases the dependability of the classification results. Data-level fusion combines unprocessed data from several sources to produce a single dataset that includes various facets of the target phenomenon. The categorization criteria, data characteristics, and fusion objectives all influence the choice of fusion scale. Feature-level fusion handles dissimilar features, decision-level fusion maximizes model variety, and data-level fusion ensures a comprehensive understanding, all of which contribute to precise and effective classification results from numerous data sources.

2.4 Multi-Source Data Fusion Classification Platform Architectures

It is possible to categorize multi-source data fusion using a variety of platform architectures, each of which is tailored to meet certain data processing and deployment requirements. The three main architectures are edge computing platforms, which process data closer to sources and are ideal for low-latency applications, distributed computing frameworks like Apache Spark, which are suited for large datasets and real-time processing, and cloud-based systems, which offer scalability and remote processing. Large data quantities are efficiently handled by cloud-based systems, distributed frameworks perform well in big data and real-time scenarios, and edge computing is appropriate in low-latency and decentralized contexts. When choosing an architecture, considerations including data volume, processing speed, real-time needs, scalability, and deployment restrictions must be taken into account to guarantee the multi-source data fusion classification project's best performance and efficiency.

III. METHODOLOGY

3.1 Methodology

3.1.1 Data Collection

Collect data from multiple sources, including Points of Interest (POI), GPS trajectories, microblog check-ins, social media, and data from public transportation. Ensure that the data sources offer a diverse range of urban activities and reflect many aspects of urban life.

Implement systems for real-time data collecting to be used for microblog check-ins. In order to continuously retrieve the most recent check-in information as it becomes available, you must connect to microblogging services or APIs.

3.1.2 Data Fusion

- Preprocessing: Cleanse and preprocess the collected data to remove duplicates, outliers, and extraneous information. Normalize and standardize the data to ensure compatibility between multiple data sources.

- Feature Extraction: Use feature extraction methods to extract relevant spatio-temporal features from each data source. These features include things like location coordinates, timestamps, check-in frequency, user preferences, and other crucial contextual information.

- Weighted Fusion: To combine the features that have been gathered from diverse data sources, develop a weighted fusion process. Give each feature the appropriate weights based on its worth and relevance to hotspot analysis. This ensures that every data source makes a fair contribution to the analysis as a whole.

3.1.3 Real-Time Processing

- Data Cleaning: Remove duplicates, extraneous information, and errors from the microblog check-in data.

- Location Extraction: Take precise position coordinates from user-generated information, including text or hashtags, to boost the precision of spatial analysis.

- Time Segmentation: Separate the data into specific time intervals (such as hourly, daily, or weekly) to enable temporal analysis and improve timeliness.

- Stream Processing: Use stream processing frameworks like Apache Kafka or Apache Flink to control the continuous flow of real-time data. These frameworks provide efficient ways to acquire, manipulate, and analyse data, allowing processing to happen almost instantly.

Use parallel processing algorithms to spread out the computer workload and improve the efficiency of real-time data processing. This can involve developing parallel algorithms or utilizing distributed computing frameworks to handle the enormous volume and velocity of incoming data.

- Time-Based Segmentation: To divide the real-time data stream into smaller time pieces, use a time-based segmentation technique. This enables the analysis of data across predetermined time periods and facilitates the identification of temporal trends and patterns in the activity of urban hotspots.

3.1.4 Multi-Source Data Fusion

- Data Incorporation: Data from other pertinent sources, such as GPS trajectory and POI data, should be merged with information from microblog check-ins. This integration increases the comprehensiveness of the investigation of urban hotspots and the identification of functional areas.

3.1.5 Timeliness Optimization

- Reducing Latency: By streamlining the data processing pipeline, you can shorten the interval between data collection and analysis. This may require improving data storage and retrieval systems, optimizing algorithms, and using efficient data structures for real-time searching and processing.

- Caching and In-Memory Processing: To reduce latency and improve the responsiveness of data analysis, use in-memory processing techniques and caching systems. In this approach, information that is often accessed is stored in memory to facilitate retrieval and processing.

- Incremental Updates: As new data become available, update the analytic results regularly using incremental update approaches. This ensures that the hotspot analysis is constantly up to date and effectively reflects the most recent trends and patterns in urban activity.

3.1.6 Spatial and Temporal Analysis

- Clustering techniques: Use clustering techniques, such as DBSCAN or K-means, to group microblog check-ins into urban hotspots according to their spatial proximity.

- Temporal Patterns: Recognize temporal patterns and trends in check-in activity to understand dynamic fluctuations in hotspot popularity over time.

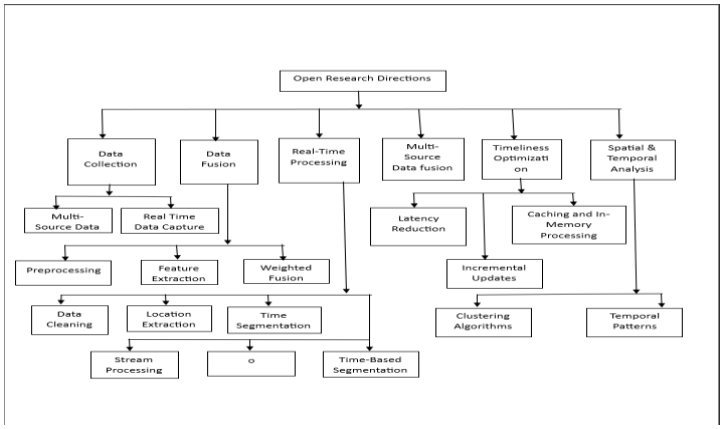


Figure 3.1: Methodology Flow

3.1.7 Evaluation

- Accuracy Assessment: Evaluate the accuracy and effectiveness of the improved timeliness in hotspot analysis by comparing the results with actual data or existing benchmark datasets. Use criteria like precision, recall, F1 score, and spatial accuracy to evaluate how effective the suggested techniques are.

3.1.8 Case Studies and Application

- Real-World Experiments: Conduct case studies in actual urban settings to confirm the increased timeliness approach's applicability and efficacy.

Implement the approach and conduct application testing to demonstrate its value in decision-making processes across a range of industries, including urban planning, traffic control, and tourism development.

Using this research methodology can increase the timeliness of microblog check-in data for urban hotspot analysis and functional area discovery. As a result of the integration of real-time data collecting, multi-source data fusion, spatial analysis, and assessment, stakeholders will be able to allocate resources more promptly and wisely.

3.2 Techniques Used for Urban Hotspot Analysis with Multi-Source Data Fusion

It was made simpler to combine and evaluate data from multiple sources by using the several techniques. The following are some of the algorithms used in the study:

Numerous techniques were employed in the study, "Enhancing Timeliness in Urban Hotspot Analysis with Multi-Source Data Fusion," to make it simpler to aggregate and interpret data from numerous sources. The following are some of the algorithms used in the study:

3.2.1 Real-Time Data Capture Algorithms:

Sensors and social media platforms were only two of the many sources of data that real-time data processing algorithms were utilized to collect and evaluate. These methods reduce the time before new data is available by ensuring that it is swiftly obtained and included into the analysis.

3.2.2 Data Preprocessing Algorithms:

Using data pretreatment techniques, data from multiple sources was cleaned, normalized, and transformed. These methods aid in handling missing values, outliers, and inconsistencies to assure data quality prior to data fusion and analysis.

3.2.3 Feature Extraction Algorithms:

Feature extraction techniques were applied to each data source in order to extract the relevant traits and features. These techniques are necessary for locating distinctive characteristics and reducing feature dimensionality, enabling efficient feature fusion and classification.

3.2.4 Fusion Algorithms:

Several fusion techniques were used to merge data from various data sources. These algorithms might make use of weighted averaging, Dempster-Shafer theory, Bayesian approaches, or other fusion techniques. Fusion techniques are used to merge data from several sources into a single, rich dataset for hotspot study.

3.2.5 Classification Algorithms:

Support Vector Machines (SVM), Random Forests, and Neural Networks, among other machine learning classification techniques, were used to perform tasks including functional region recognition and hotspot analysis. These methods train the fused dataset to create classification models and generate accurate predictions.

3.2.6 Real-Time Data Segmentation Algorithms:

The problems with real-time microblog check-in data were addressed using real-time data segmentation techniques. These algorithms filter and segment the incoming data effectively, which enhances the rapid hotspot analysis.

3.2.7 Parallel Processing Algorithms:

Through the use of parallel processing techniques, data processing tasks were spread across a large number of nodes or cores, enhancing the efficiency of data fusion and analysis. These techniques improve the use of processing resources and accelerate the fusion process, especially when dealing with large amounts of data.

3.2.8 Ensemble Fusion Algorithms:

Investigated were ensemble fusion techniques, which combine the results of different classification models to further improve classification accuracy. These algorithms combine the benefits of various models to provide hotspot projections that are more precise.

The combination of these algorithms allowed the research to increase the timeliness and accuracy of urban hotspot analysis through the fusion of data from various sources. Real-time data gathering and processing technologies in conjunction with advanced fusion and classification algorithms enabled a more comprehensive understanding of urban activities and the capacity to make well-informed decisions in a number of urban settings.

IV. MODULES

4.1 Modules

4.1.1 Data Level Comparison

a. Single Source Data Recognition: To perform functional area recognition, use a single source of data, such as POI data or microblog check-in data. Apply the appropriate methods or techniques, such as clustering or classification algorithms, to the selected data source to identify functional areas.

b. multi-Source data with unbalanced clustering: Combining data from several sources, including POI data, check-in data from microblogs, and other related datasets. You could correct for the imbalance in the volume of the data by utilizing data preparation techniques like oversampling or under sampling. Use clustering techniques like K-means or DBSCAN to group the combined data and find urban hotspots.

4.1.2 Clustering Algorithm Level Comparison

a. Apply a clustering method, such as K-means, to the multi-source fused data without resolving the issue of data imbalance. a. Multi-Source Data with Unbalanced Clustering. Examine and evaluate the results of this method.

b. Before using a clustering algorithm, such as K-means, to analyze the multi-source fused data, use techniques like Synthetic Minority Oversampling Technique (SMOTE) or class weighting to address the problem of data imbalance. Examine and evaluate the results of this method.

4.1.3 Evaluation Metrics

a. Accuracy: Verify that the functional area recognition produced by each experimental strategy is accurate. Calculate metrics like precision, recall, F1 score, or accuracy to assess how well each scheme performs in correctly identifying functional areas.

b. Imbalance Handling Evaluation: Examine the effects of data imbalance correction on the clustering results. Compare the measures of clustering performance, such as silhouette coefficient, cohesiveness, or separation, between the multi-source data with and without addressing the problem of data imbalance.

4.1.4 Experimental Setup

a. Data Preparation: The datasets required for the studies, such as the ground truth or benchmark datasets that correlate to the single-source data and multi-source fused data, should be prepared.

b. Algorithm Implementation: Follow the experimental guidelines when implementing the selected data balancing techniques, such as SMOTE or class weighting, and clustering algorithms, such as K-means.

c. Execution of the Experiments: To ensure uniformity and repeatability, the Experiments should be carried out using the appropriate software tools or programming languages.

d. Analysis of Results: Examine the results of each experimental method, taking into account the accuracy of functional area detection and the success of clustering when data imbalance is present.

By following these methods, three experimental schemes may be compared in terms of data quality and clustering algorithm performance, their accuracy in spotting functional regions, and how well they manage data imbalance in multi-source data fusion for urban hotspot analysis.

4.2 ALGORITHMS AND PSEUDOCODE

4.2.1 Data Preprocessing Algorithm

Algorithm: Missing Data Imputation

Pseudocode:

```
For each feature in the dataset:
  If feature has missing values:
    Identify the missing values
    Choose a suitable imputation method (e.g., mean, median, or K-nearest neighbors)
    Replace missing values with imputed values
```

Description

The Missing Data Imputation algorithm is used to address the missing data issue in the multi-source dataset. Data from many sources may have missing values in an urban hotspot research due to a number of reasons, such as sensor failures or insufficient data gathering. With the help of the appropriate imputation techniques, this method finds missing values in each feature and fills in the data gaps. By quickly filling in any gaps in the data, this technique increases the accuracy and reliability of hotspot identification and makes the data readily available for quick fusion and analysis.

4.2.2 Fusion Algorithm

Algorithm: Weighted Averaging Fusion

Pseudocode:

```
    For each data source (i) in the dataset:
        Assign a weight (wi) based on data source reliability or importance
        Multiply each feature in data source (i) by its corresponding weight (wi)
        Sum up the weighted features across all data sources
        Normalize the fused features to maintain consistency
```

Description

The Weighted Averaging Fusion approach combines data from many sources by assigning weights to each source. The weights show how important or reliable each data source is in comparison to the rest. The features from each data source are then multiplied by the appropriate weights to create the fused dataset. Making sure that the combined data's scale is constant is what normalization does. By using weighted averaging, this algorithm effectively merges data from several sources while taking into consideration the unique significance of each one. When appropriately applied, this fusion process enables the system to react to real-time information from numerous sources, enabling continuous and speedy hotspot analysis.

4.2.3 Classification Algorithm:

Algorithm: Support Vector Machines (SVM)

Pseudocode:

```
    Initialize SVM classifier with appropriate kernel (e.g., linear, polynomial, or radial basis function)
    Train the classifier using the fused dataset as input and hotspot labels as target
    Test the trained SVM on new data samples to predict hotspot or functional area labels
```

Description

Support Vector Machines (SVM), a popular classification technique, are used in hotspot research to identify urban hotspots and functional zones. SVM searches the feature space for the hyperplane that best distinguishes between different classes. The fused dataset, which is a compilation of comprehensive data from several sources, is utilized to train the SVM classifier. Following training, the SVM may predict the labels for functional or hotspot areas for new data samples. By employing SVM for classification, the system identifies urban hotspots with high accuracy and promptly contributes to their discovery, enabling responsive urban planning and management.

The Real-Time Twitter Data Streaming algorithm gathers real-time social media data while the Missing Data Imputation algorithm addresses missing data issues to prepare the dataset for fusion. The Weighted Averaging Fusion approach successfully combines data from diverse sources, and SVM provides accurate categorization for hotspot analysis. Using multi-source data fusion and these algorithms, urban hotspot research is made more current, allowing for timely insights and well-informed decisions for urban planning and management.

4.3 DATA SAMPLING

4.3.1 Data Sections:

1. The Raw_data folder contains the raw check-in data from September 1 through September 30. Based on the longitude and latitude coordinates of the check-in data, the spatial positions were divided into 28*36 grids, and the number of check-ins in each grid indicated the number of check-ins within the area of this grid. The hourly check-ins were contained in 720 files in total.

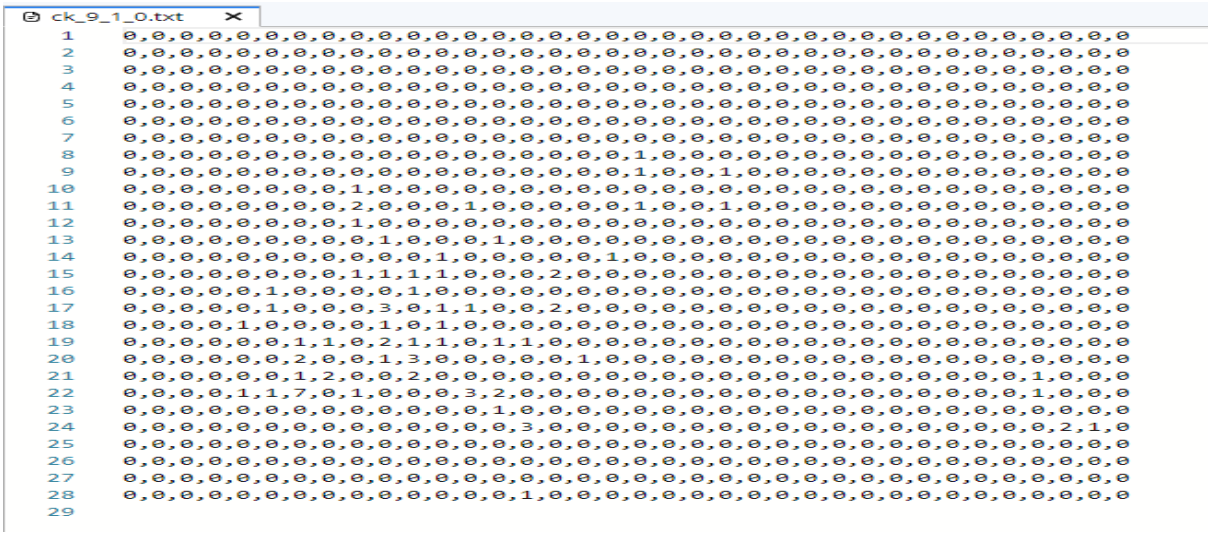


figure 4.1 raw check-in data the spatial positions were divided into 28*36 grids

2. The Completion_data folder stores the hourly completion of the original check-in data, which maintains the value of the initial check-in data in the pertinent grid.

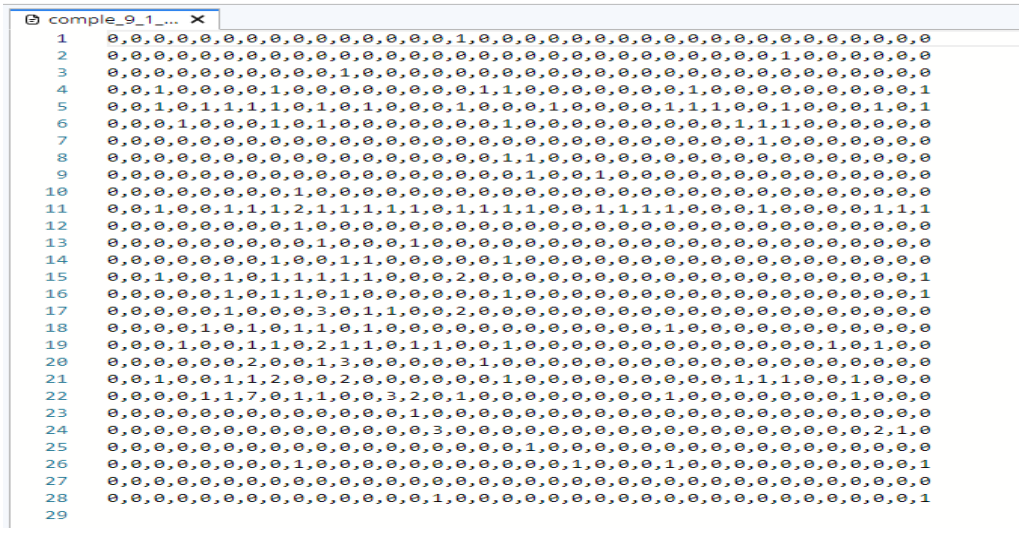


Figure 4.2: hourly completion of the original check-in data

3. The poi_vector.csv file creates a vector form of 1*16, where 16 is the number of POI categories, from the amount of POI data included inside each grid's latitude and longitude range.

poi_vector.csv poi_vector.csv

Raw

	0	102.588-102.598	102.598-102.608	102.608-102.618	102.618-102.628	102.628-102.638	102.638-102.648	102.648-1
1	25.135-25.145	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 4, 0, 0, 2, 0,...	[0, 0, 0, 0,
2	25.125-25.135	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 1, 0, 0, 1, 0, ...	[0, 0, 0, 0,
3	25.115-25.125	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0,
4	25.105-25.115	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 2, 0, 0
5	25.095-25.105	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 1, 1, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0,
6	25.085-25.095	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 18, 32, 0, 7, 3, 6, 0, ...	[0, 40, 69,
7	25.075-25.085	[0, 3, 0, 0, 9, 0, 0, 0, 0...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 64, 28, 0, 24, 11, 6,...	[0, 59, 96,
8	25.065-25.075	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 1, 0, 0, 0, 0,...	[0, 2, 0, 0, 4, 0, 0, 1, 0,...	[0, 0, 0, 0, 0, 0, 0, 1, 0,...	[1, 3, 23, 0, 0, 1, 0, 0, 1,...	[0, 15, 79, (
9	25.055-25.065	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 1, 0, 0, 0, 0,...	[0, 0, 0, 0, 0, 0, 0, 1, 0,...	[0, 0, 1, 0, 2, 0, 0, 0, 0,...	[0, 25, 29, 0, 10, 7, 4, ...	[1, 41, 68, 0
10	25.045-25.055	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 3, 0, 0, 0, 0...	[0, 0, 0, 0, 1, 1, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 4, 3, 0, 1, 1, 1, 1, 0, 0...	[1, 33, 7, 0, 19, 3, 3, 1, ...	[1, 17, 77, 0,
11	25.035-25.045	[0, 0, 0, 0, 4, 0, 0, 1, 0,...	[0, 0, 0, 0, 1, 0, 0, 0, 0,...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 2, 0, 0, 0, 0...	[0, 10, 13, 0, 7, 1, 0, 0, ...	[1, 43, 28, 0, 6, 6, 21, ...	[1, 82, 143,
12	25.025-25.035	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 42, 76, 0, 16, 16, 12...	[1, 15, 15, 0, 5, 11, 4, 0, ...	[0, 0, 5, 0,
13	25.015-25.025	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 1, 0, 0, 0, 0, 0, 0, 0,...	[0, 0, 4, 0, 0, 0, 0, 0, ...	[0, 14, 1, 0, 6, 1, 8, 0, 0,...	[0, 7, 0, 0, (
14	25.005-25.015	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 2, 7, 0, 1, 3, 0, 0, 0, ...	[0, 2, 0, 0, 0, 0, 0, 0, 0...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0,
15	24.995-25.005	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 4, 9, 0, 3, 1, 1, 0, 0, ...	[1, 3, 0, 0, 1, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 1, 7, 0, (
16	24.985-24.995	[0, 0, 0, 0, 1, 0, 0, 0, 0,...	[2, 0, 0, 0, 0, 0, 0, 0, 0...	[0, 8, 0, 0, 1, 1, 0, 0, 0, ...	[0, 3, 1, 0, 0, 0, 0, 0, 0,...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 2, 7, 0, :
17	24.975-24.985	[0, 0, 0, 0, 1, 0, 0, 0, 0,...	[1, 0, 0, 0, 0, 0, 0, 0, 0,...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 9, 1, 0, 0, 1, 0, 1, 0, ...	[0, 6, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 11, 0, 0,
18	24.965-24.975	[2, 4, 2, 0, 1, 2, 0, 0, 0,...	[0, 1, 0, 0, 0, 0, 0, 0, 0,...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 3, 0...	[0, 10, 2, 0, 0, 1, 1, 0, 0,...	[0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 6, 2, 0,

Figure 4.3 table view of POI data within the latitude and longitude range of each grid is constructed into a vector form of 1*16

4. The poi_build3.csv file contains the number of POI points for each of the 28*36 grids.

poi_build3.csv

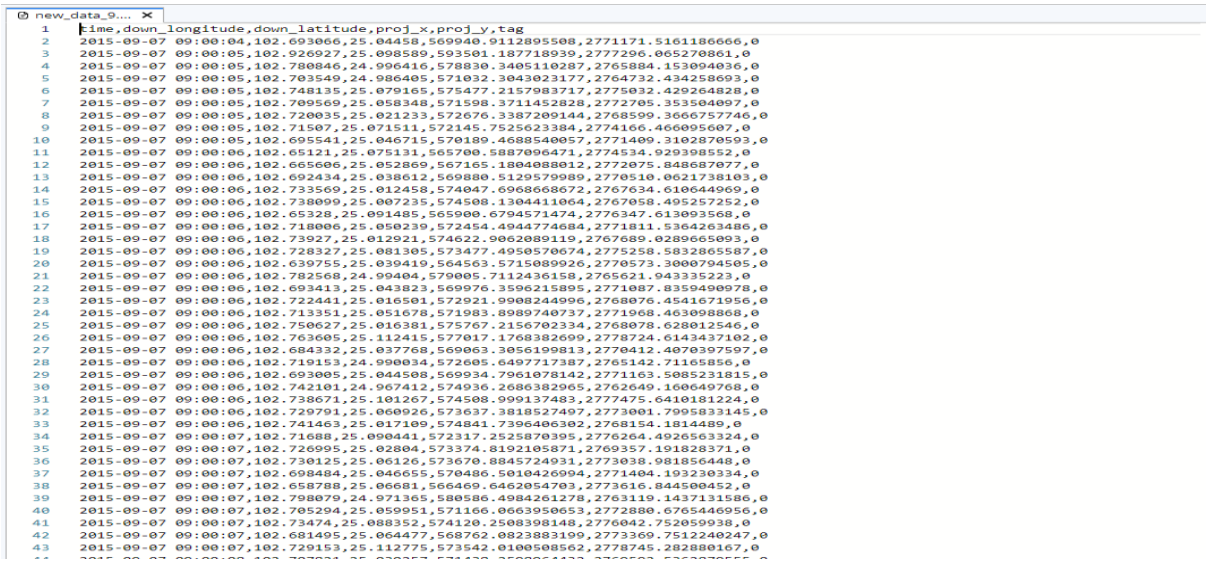
```
0,0,0,0,0,7,0,0,0,1,0,0,0,0,46,79,10,7,20,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0
0,0,0,0,0,3,0,0,0,3,0,0,0,0,1,20,210,40,3,1,0,0,0,0,0,0,0,0,14,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,1,1,0,0,0,34,7,96,35,1,3,0,0,0,0,0,5,2,20,0,0,3,0,0,0,0,0,0
0,0,0,0,0,0,3,17,6,1,1,0,4,175,273,13,140,47,3,2,0,0,0,0,0,2,10,0,2,0,0,0,0,0,0,0
0,0,0,0,0,3,0,12,30,1,0,0,12,288,834,142,7,3,21,5,0,13,0,3,10,0,0,1,0,0,0,1,0,2,2,0
0,0,0,0,0,70,200,38,137,4,0,3,95,206,589,285,42,72,55,19,13,9,0,10,30,9,0,1,0,0,0,0,5,1,0,0
15,0,0,0,0,146,301,26,105,3,50,28,266,407,871,235,282,123,31,0,2,1,0,0,3,0,0,0,0,0,1,0,10,10,0,0
0,0,1,7,1,31,160,276,230,56,173,222,438,686,430,119,125,16,0,0,0,0,0,0,0,0,0,0,0,0,1,1,3,10,0,0
0,0,3,1,3,83,172,294,309,261,334,457,246,474,600,358,361,16,0,0,2,0,0,0,0,0,0,0,0,0,4,11,1,0,2,1
0,4,2,0,13,77,153,97,647,642,1035,768,756,488,583,398,13,32,31,34,65,0,0,0,0,0,0,0,0,113,86,25,2,0,0,0,0
5,1,0,2,35,118,342,706,518,340,629,1148,1339,582,639,376,76,61,24,62,17,1,1,0,0,0,2,3,36,28,0,0,0,1,0,0
0,0,0,0,176,67,21,99,169,344,460,512,723,840,503,265,79,53,22,0,0,0,4,4,0,0,1,1,29,0,0,0,0,0,0,0
0,0,0,1,4,34,8,2,105,261,187,396,208,595,398,229,86,53,21,15,2,0,0,0,2,9,5,0,8,0,0,0,0,0,0,0
0,0,0,15,3,0,0,17,244,605,337,92,95,445,521,573,127,151,52,34,1,0,1,11,5,1,3,3,0,0,0,0,0,0,0,0
0,0,20,6,0,0,20,105,77,309,113,85,239,255,234,120,82,295,89,162,38,2,3,10,5,8,2,6,1,0,0,0,0,0,0,0
1,2,15,8,0,0,13,43,4,45,270,285,362,383,97,39,361,251,229,46,33,40,7,26,18,22,2,0,0,1,0,0,0,0,0,0
1,1,0,20,10,0,38,45,60,21,101,148,226,352,87,31,280,290,221,52,214,17,2,41,25,14,16,1,0,0,0,0,0,0,0,0
12,1,0,3,15,0,25,122,30,22,35,13,172,15,29,56,79,156,240,15,34,26,0,0,0,1,0,2,2,0,0,0,0,0,0,0
2,0,0,2,18,5,11,26,23,9,1,7,110,4,20,162,604,433,66,8,50,8,1,2,28,23,0,0,0,2,0,0,1,2,0,7
1,0,0,5,33,5,0,0,0,0,17,133,3,2,52,93,129,393,140,9,4,6,0,1,2,13,0,0,0,4,0,0,5,0,5,26
3,0,0,3,4,0,0,0,0,0,9,15,6,6,52,21,5,267,273,296,1,0,0,0,1,4,0,0,0,0,0,0,2,1,0,0
1,0,0,5,0,0,0,0,0,0,2,6,0,7,15,2,4,66,6,70,107,0,4,10,3,6,3,0,0,0,1,0,0,0,0,1,0
1,0,0,0,0,1,0,0,0,0,1,0,0,6,2,1,13,14,8,18,1,9,2,3,4,7,0,0,0,0,1,0,0,0,0,0
0,0,0,3,0,2,0,0,0,0,0,0,0,0,0,0,0,0,12,2,1,18,5,7,2,8,2,0,0,0,0,0,1,0,0,0
0,0,0,0,0,1,13,0,0,0,0,0,0,0,0,0,0,0,14,197,55,2,0,2,8,44,0,0,0,0,0,1,2,0,8,1
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3,2,221,497,47,50,3,9,17,0,0,0,0,0,0,0,0
0,0,0,0,0,0,8,0,0,0,0,0,0,0,0,0,0,0,0,3,3,23,22,33,14,0,0,0,0,2,3,0,0,0,0,0
0,0,0,0,0,3,1,1,0,0,0,0,0,0,0,0,0,0,0,0,13,10,2,4,3,98,2,2,0,0,0,0,0,0,0,0
```

figure 4.4 POI points in each of 28*36 grids.

unbalanced_...

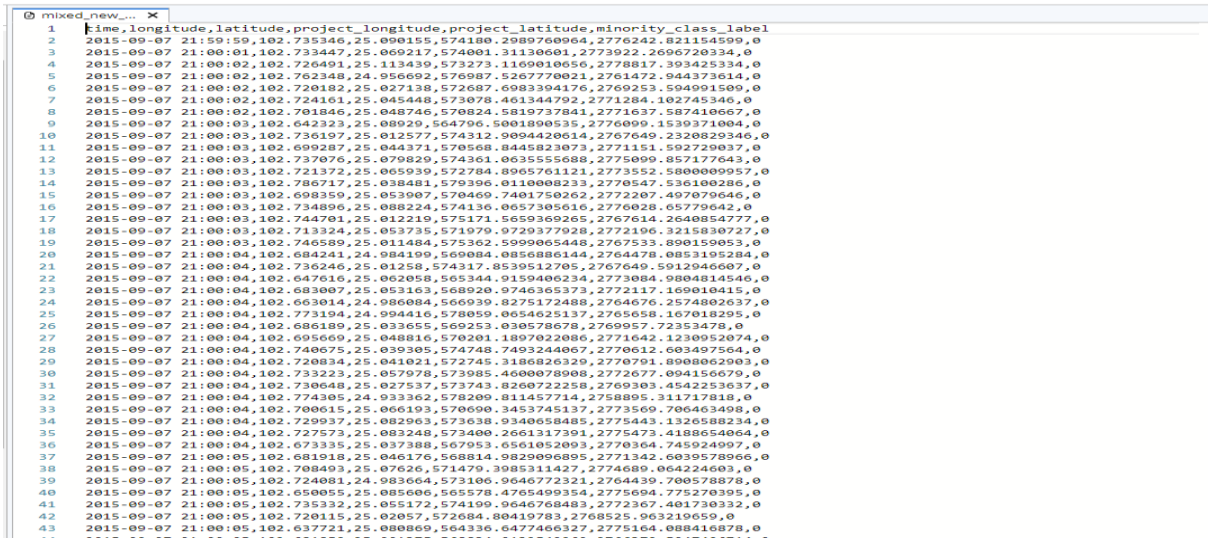
```
x, y, cluster_label, minority_class_label
36.3,25.8,1,1
35.15,25.9,1,1
34.2,26.35,1,1
34.1,26.85,1,1
33.45,27.8,1,1
33.75,27.95,1,1
34.4,28.0,1,1
34.55,27.25,1,1
34.8,26.85,1,1
35.5,26.85,1,1
5.2,2.15,4,1
6.35,1.95,4,1
6.75,2.3,4,1
5.9,2.4,4,1
5.4,2.7,4,1
4.85,2.9,4,1
4.85,3.35,4,1
5.15,3.45,4,1
5.7,3.45,4,1
6.2,3,4,1
6.2,3.2,4,1
7.65,2.15,4,1
7.2,2.75,4,1
6.75,3.2,4,1
6.75,3.55,4,1
6.65,3.8,4,1
5.8,4,4,1
4.95,4.05,4,1
5.1,4.35,4,1
5.7,4.45,4,1
9.4,17.68,2,0
7.79,19.52,2,0
9.46,25.38,2,0
5.28,28.72,2,0
7.77,23.86,2,0
7.56,29.86,2,0
9.02,21.34,2,0
11.83,22.49,2,0
7.6,21.46,2,0
8.1,27.85,2,0
9.82,21.52,2,0
11.12,26.38,2,0
```

figure 4.5 check-in data set (original + addition).



```
new_data_9....
1  time,down_longitude,down_latitude,proj_x,proj_y,tag
2  2015-09-07 09:00:04,102.693066,25.04458,569940.9112895508,2771171.5161186666,0
3  2015-09-07 09:00:05,102.926927,25.098589,593501.187718939,2777296.065270861,0
4  2015-09-07 09:00:05,102.780846,24.996416,578830.3405110287,2765884.153094036,0
5  2015-09-07 09:00:05,102.703549,24.986405,571032.3043023177,2764732.434258693,0
6  2015-09-07 09:00:05,102.748135,25.079165,575477.2157983717,2775032.429264828,0
7  2015-09-07 09:00:05,102.709569,25.058348,571598.3711452828,2772705.353504097,0
8  2015-09-07 09:00:05,102.720035,25.021233,572676.3387209144,2768599.3666757746,0
9  2015-09-07 09:00:05,102.71507,25.071511,572145.7525623384,2774166.466095607,0
10 2015-09-07 09:00:05,102.695541,25.046715,570189.4688540057,2771409.3102870593,0
11 2015-09-07 09:00:06,102.65121,25.075131,565700.5887096471,2774534.929398552,0
12 2015-09-07 09:00:06,102.665606,25.052869,567165.1804988012,2772075.848668707,0
13 2015-09-07 09:00:06,102.692434,25.038612,569880.5129579989,2770510.0621738103,0
14 2015-09-07 09:00:06,102.733569,25.012458,574047.6968668672,2767634.610644969,0
15 2015-09-07 09:00:06,102.738099,25.007235,574508.1304411064,2767058.495257252,0
16 2015-09-07 09:00:06,102.65328,25.091485,565900.6794571474,2776347.613093568,0
17 2015-09-07 09:00:06,102.718006,25.050239,572454.4944774684,2771811.5364263486,0
18 2015-09-07 09:00:06,102.73927,25.012921,574622.9062089119,2767689.0289665093,0
19 2015-09-07 09:00:06,102.728327,25.081305,573477.4950570674,2775258.5832865587,0
20 2015-09-07 09:00:06,102.639755,25.039419,564563.5715089926,2770573.3000794505,0
21 2015-09-07 09:00:06,102.782568,24.99404,579005.7112436158,2765621.943335223,0
22 2015-09-07 09:00:06,102.693413,25.043823,569976.3596215895,2771087.8359490978,0
23 2015-09-07 09:00:06,102.722441,25.016501,572921.9908244996,2768076.4541671956,0
24 2015-09-07 09:00:06,102.713351,25.051678,571983.8989740737,2771968.463098868,0
25 2015-09-07 09:00:06,102.750627,25.016381,575767.2156702334,2768078.628012546,0
26 2015-09-07 09:00:06,102.763605,25.112415,577017.1768382699,2778724.6143437102,0
27 2015-09-07 09:00:06,102.684332,25.037768,569063.3056199813,2770412.4070397597,0
28 2015-09-07 09:00:06,102.719153,24.990034,572605.6497717387,2765142.71165856,0
29 2015-09-07 09:00:06,102.693605,25.044508,569934.7961078142,2771163.598523815,0
30 2015-09-07 09:00:06,102.742101,24.967412,574936.2686382965,2762649.160649768,0
31 2015-09-07 09:00:06,102.738671,25.101267,574508.999137483,2777475.6410181224,0
32 2015-09-07 09:00:06,102.729791,25.060926,573637.3818527497,2773001.7995833145,0
33 2015-09-07 09:00:06,102.741463,25.017109,574841.7396406302,2768154.1814489,0
34 2015-09-07 09:00:07,102.71688,25.090441,572317.2525870395,2776264.4926563324,0
35 2015-09-07 09:00:07,102.726995,25.02804,573374.8192105871,2769357.191828371,0
36 2015-09-07 09:00:07,102.730125,25.06126,573670.8845724931,2773038.981856448,0
37 2015-09-07 09:00:07,102.698484,25.046655,570486.5010426994,2771404.193230334,0
38 2015-09-07 09:00:07,102.658788,25.06681,566469.6462054703,2773616.844500452,0
39 2015-09-07 09:00:07,102.798079,24.971365,580586.4984261278,2763119.1437131586,0
40 2015-09-07 09:00:07,102.705294,25.059951,571166.0663950653,2772880.6765446956,0
41 2015-09-07 09:00:07,102.73474,25.088352,574120.2508398148,2776042.752059938,0
42 2015-09-07 09:00:07,102.681495,25.064477,568762.0823883199,2773369.7512240247,0
43 2015-09-07 09:00:07,102.729153,25.112775,573542.0100508562,2778745.282880167,0
```

figure 4.6 an unbalanced data set consisting of GPS+ check-in data (original)



```
mixed_new_...
1  time,longitude,latitude,project_longitude,project_latitude,minority_class_label
2  2015-09-07 21:59:59,102.735346,25.090155,574180.2989760964,2776242.821154599,0
3  2015-09-07 21:00:01,102.733447,25.069217,574001.31136601,2773922.2696720334,0
4  2015-09-07 21:00:02,102.726491,25.113439,573273.1169010656,2778817.393425334,0
5  2015-09-07 21:00:02,102.762348,24.956692,576987.5267770021,2761472.944373614,0
6  2015-09-07 21:00:02,102.720182,25.027138,572687.6983394176,2769253.594991509,0
7  2015-09-07 21:00:02,102.724161,25.045448,573078.461344792,2771284.102745346,0
8  2015-09-07 21:00:02,102.701846,25.048746,570824.5819737841,2771637.587410667,0
9  2015-09-07 21:00:03,102.642323,25.08929,564796.5001890535,2776099.1539371004,0
10 2015-09-07 21:00:03,102.736197,25.012577,574312.9094420614,2767649.2320829346,0
11 2015-09-07 21:00:03,102.699287,25.044371,570568.8445823073,2771151.5927290037,0
12 2015-09-07 21:00:03,102.737076,25.079829,574361.0635555688,2775099.857177643,0
13 2015-09-07 21:00:03,102.721372,25.065939,572784.8965761121,2773552.5800009957,0
14 2015-09-07 21:00:03,102.786717,25.038481,579396.0110008233,2770547.536100286,0
15 2015-09-07 21:00:03,102.698359,25.053907,570469.7401750262,2772207.497079646,0
16 2015-09-07 21:00:03,102.734896,25.088224,574136.0657305616,2776028.65779642,0
17 2015-09-07 21:00:03,102.744701,25.012219,575171.5659369265,2767614.2640854777,0
18 2015-09-07 21:00:03,102.713324,25.053715,571979.9729377928,2772106.3215830727,0
19 2015-09-07 21:00:03,102.746589,25.011484,575362.5999065448,2767533.890159053,0
20 2015-09-07 21:00:04,102.684241,24.984199,569084.0856886144,2764478.0853195284,0
21 2015-09-07 21:00:04,102.736246,25.01258,574317.8539512705,2767649.5912946607,0
22 2015-09-07 21:00:04,102.647616,25.062058,565344.9159406234,2773084.9804814546,0
23 2015-09-07 21:00:04,102.683007,25.053163,568920.9746365373,2772117.169010415,0
24 2015-09-07 21:00:04,102.663014,24.986084,566939.8275172488,2764676.2574802637,0
25 2015-09-07 21:00:04,102.773194,24.994416,578059.0654625137,2765658.167018295,0
26 2015-09-07 21:00:04,102.686189,25.033655,569253.030578678,2769957.72353478,0
27 2015-09-07 21:00:04,102.695669,25.048816,570201.1897022086,2771642.1230952074,0
28 2015-09-07 21:00:04,102.740675,25.039305,574748.7493244067,2770612.603497564,0
29 2015-09-07 21:00:04,102.720834,25.041021,572745.3186826329,2770791.8908062903,0
30 2015-09-07 21:00:04,102.733223,25.057978,573985.4600078908,2772677.094156679,0
31 2015-09-07 21:00:04,102.730648,25.027537,573743.8260722258,2769303.4542253637,0
32 2015-09-07 21:00:04,102.774305,24.933362,578209.811457714,2758895.311717818,0
33 2015-09-07 21:00:04,102.700615,25.066193,570690.3453745137,2773569.706463498,0
34 2015-09-07 21:00:04,102.729937,25.082963,573638.9340658485,2775443.1326588234,0
35 2015-09-07 21:00:04,102.727573,25.083248,573400.2661317391,2775473.4188654064,0
36 2015-09-07 21:00:04,102.673335,25.037388,567953.6561052093,2770364.745924997,0
37 2015-09-07 21:00:05,102.681918,25.046176,568814.9829096895,2771342.6039578966,0
38 2015-09-07 21:00:05,102.708493,25.07626,571479.3985311427,2774689.064224603,0
39 2015-09-07 21:00:05,102.724081,24.983664,573106.9646772321,2764439.700578878,0
40 2015-09-07 21:00:05,102.650055,25.085606,565578.4765499354,2775694.775270395,0
41 2015-09-07 21:00:05,102.735332,25.055172,574199.9646768483,2772367.401730332,0
42 2015-09-07 21:00:05,102.720115,25.02057,572684.80419783,2768525.963219659,0
43 2015-09-07 21:00:05,102.637721,25.080869,564336.6477466327,2775164.088416878,0
```

figure 4.7 an unbalanced data set consisting of GPS+ check-in data (original + addition)

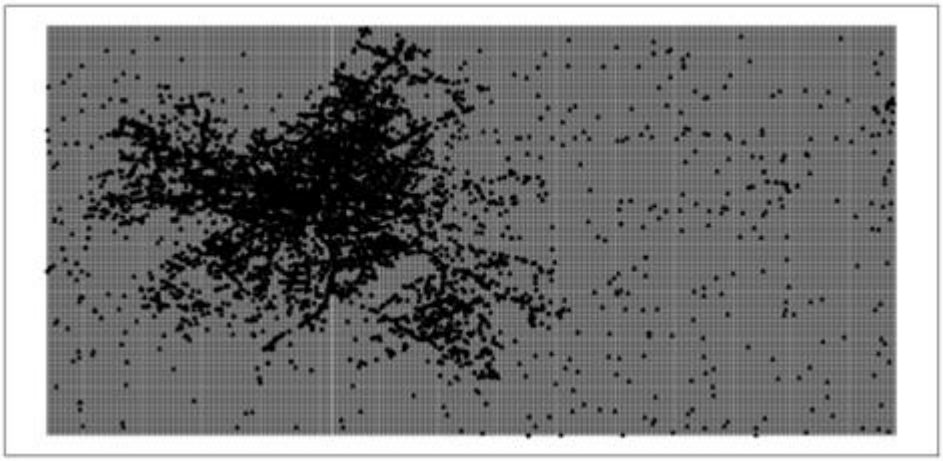


figure 4.8 cluster map of an unbalanced data set consisting of GPS+ check-in data (original + addition)

V. CONCLUSION

5. Conclusion

Urban hotspot analysis has been shown to be much timelier when multi-source data fusion is used, which improves the efficiency of urban planning and decision-making processes. The research successfully combined data from numerous sources, including GPS trajectories, check-ins, and Point of Interest (POI) data, to address data imbalance and improve functional area detection. Real-time data gathering and stream processing were used in the study's methodology to bridge the time lag between data availability and hotspot updates. In addition to improving hotspot recognition accuracy, feature-oriented fusion of spatio-temporal

features gave a comprehensive perspective of urban dynamics, highlighting its usefulness in emergency response, traffic management, tourism, and urban planning. The work highlights the potential for future developments employing complex fusion algorithms and machine learning models, even though it acknowledges some limits, notably with regard to real-time microblog check-in data.

In essence, the study emphasizes how multi-source data fusion considerably improves the speed and accuracy of urban hotspot analysis, facilitating data-driven, informed urban planning and decision-making. This approach has the ability to significantly raise the quality of life for city dwellers and promote sustainable urban growth by combining various data sources and utilizing real-time data processing. Future research into improving the method might produce even more complex insights and applications, indicating a promising direction for further study and application.

5.1 Limitations and Future Directions

5.1.1 Limitations

The multi-source data fusion investigation ran across a number of issues that require more consideration. Although data availability delays were reduced, real-time microblog check-in data remained difficult because of processing constraints. To solve this problem, more study is needed to create more efficient real-time data collection and processing techniques. As a result of potential effects on fusion accuracy and dependability from missing or noisy data, data quality has also become a source of worry. Various data source attributes persist despite the use of data preparation procedures. To guarantee the coherence of the integrated dataset, future research should concentrate on improving data quality approaches. It was recognised that the fusion process is difficult, especially when feature-oriented fusion is used. It is essential to carefully choose and combine pertinent information from many sources, and as the number of data sources grows, computing needs could increase. By improving fusion algorithms and developing effective parallel processing strategies, this complexity could be controlled.

5.1.2 Future Directions

Future research will focus on a number of important areas to improve the timeliness of urban hotspot analysis using multi-source data fusion. Real-time microblog check-in data has its limits, which must be addressed. New approaches must be developed for better data segmentation, analysis, and fusion in order to obtain faster updates and more timeliness in hotspot analysis. It may be possible to improve the dataset and offer more thorough insights by combining future technologies like mobile apps, social media platforms, and smart city sensors. This would increase the hotspot detection accuracy and coverage. Exploring cutting-edge fusion methods, such as ensemble fusion models and deep learning-based approaches, might further enhance the system's capacity to manage complex data linkages and boost the precision of hotspot analysis. Case studies in various metropolitan contexts will provide real-world validation and provide useful insights into the system's scalability and adaptability. The relevance and usefulness of hotspot analysis for urban planners, enterprises, and residents can also be improved by user-centric apps that are adapted to particular user demands by including preferences and comments into the fusion process. In order to fully realize the potential of multi-source data fusion in revolutionizing urban planning and decision-making for more effective and sustainable cities, future research should address issues related to real-time data and data quality while exploring innovative fusion techniques, new data sources, and practical validations.

REFERENCES

[1] Z. Wang, D. Ma, D. Sun, and J. Zhang, “Identification and analysis of urban functional area in Hangzhou based on OSM

and POI data,” PLoS One, vol. 16, no. 5, Article ID e0251988, 2021.

[2] L. Fu, P. Lin, A.V. Vasilakos, S. Wang, An overview of recent multi-view clustering, Neurocomputing, Vol. 402, 2020, pp. 148-161.

[3] J. P. Qiu and C. Shen, “Analysis of hot topics in domestic big data research based on LDA model,” Journal of Modern Information, vol. 41, no. 09, pp. 22–31, 2021.