⁎Traffic Congestion Prediction in Beijing Using Historical GPS Trajectories (2007–2012)

A Machine Learning-Based Approach with Spatio-Temporal Visualization and Route-Level Forecasting

†Runar Kenzhekeyev
Computer Science and Technology
Beijing Institute of Technology
Beijing, China
runarkenz@gmail.com

†Kazhybek Asset
Computer Science and Technology
Beijing Institute of Technology
Beijing, China
kazhybekpsg@gmail.com

ABSTRACT

This project presents a traffic congestion prediction system developed using the GeoLife GPS trajectory dataset collected in Beijing between 2007 and 2012. The dataset comprises over 17,000 user-recorded trajectories and 18 million GPS points, capturing detailed spatio-temporal movement patterns. The system preprocesses the data, labels congestion hotspots based on spatial and temporal density, and trains a supervised machine learning model (Random Forest) to predict congestion. Users can interact with the system by inputting natural language addresses to predict congestion on specific routes and visualize results on a map. Various visualizations, including yearly trends, hourly activity, and congestion heatmaps, are generated to support exploratory analysis. This tool highlights the potential of historical trajectory data in intelligent urban traffic management and smart mobility applications.

Keywords: Traffic Congestion Prediction, GeoLife Dataset, GPS Trajectories, Spatio-Temporal Data, Machine Learning, Route Forecasting, OpenRouteService, Urban Mobility, Data Visualization, Smart Transportation

1. Introduction

Urban traffic congestion is a critical issue faced by modern cities, leading to increased travel time, pollution, and economic losses. Predicting traffic congestion in advance can play a vital role in optimizing route planning, improving transportation systems, and enabling smarter city infrastructure. With the growing availability of spatio-temporal data from GPS-enabled devices, machine learning techniques offer new opportunities to analyze urban mobility patterns at scale.

This project aims to develop an intelligent congestion prediction system using historical trajectory data from Beijing, collected as part of the GeoLife project by Microsoft Research Asia. The GeoLife dataset contains over 17,000 GPS trajectories, recorded by 182 users between April 2007 and August 2012, encompassing more than 1.2 million kilometers of travel. The trajectories include diverse daily activities such as commuting, shopping, sightseeing, and leisure, making the dataset a rich resource for mobility analysis.

The core objective of the project is to transform raw trajectory data into actionable insights on traffic congestion. To achieve this, the system performs the following steps:

- Preprocessing raw GPS points into a structured dataset.
- Identifying congestion hotspots using spatio-temporal density-based labeling.
- Training a supervised learning model (Random Forest) to predict congestion using features such as location, time, and day of the week.
- Enabling users to enter natural language routes (e.g., place names or addresses) and receive congestion predictions along the path.
- Visualizing both the route and congestion status using interactive maps and statistical diagrams.

This user-friendly tool integrates machine learning, geospatial analysis, and mapping services to support dynamic and intuitive route planning. The methodology and outcomes presented in this report contribute to the field of intelligent transportation systems, with potential applications in urban planning, traffic management, and mobility services.

2. Data Source and Preprocessing

The dataset used consists of GPS trajectories of Beijing taxis. Each point contains latitude, longitude, and timestamp values. The data is stored in a .pkl file and processed as follows:

- **Date conversion**: Excel-format timestamps were converted to Python datetime objects.
- **Data enrichment**: Each point was annotated with the hour, day, month, and year.
- **Filtering**: Trajectories were limited to the Beijing region and to dates from April 2007 to August 2012.

Data Overview

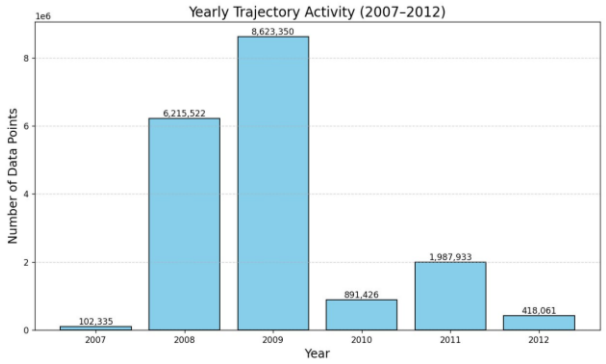
- Total points loaded: 18238630
- Points after filtering: 18238627

3. Exploratory Data Analysis

The system generates several visuals to understand spatio-temporal traffic dynamics.

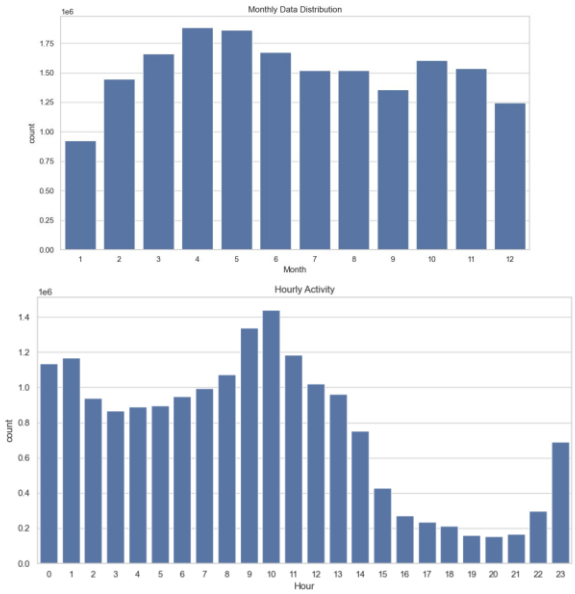
3.1 Yearly Trajectory Activity

Displays number of GPS points per year from 2007 to 2012.



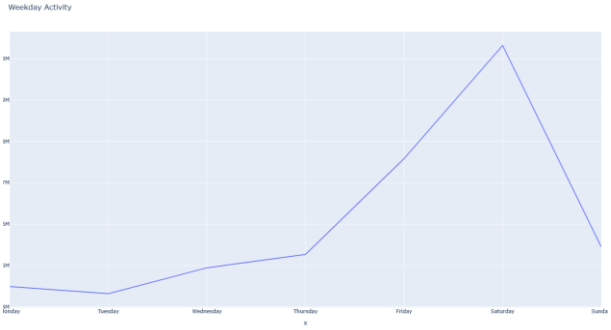
3.2 Monthly and Hourly Distributions

Indicates which months and hours have the highest traffic activity.



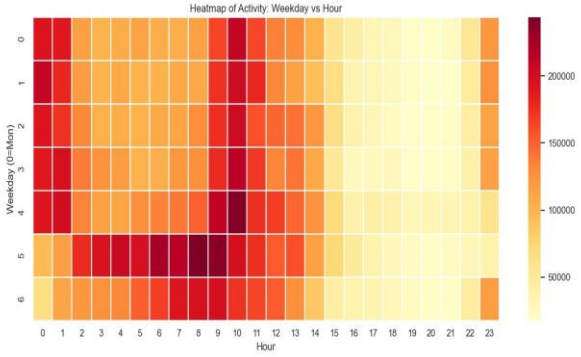
3.3 Weekday Activity

Analyzes activity across different days of the week.



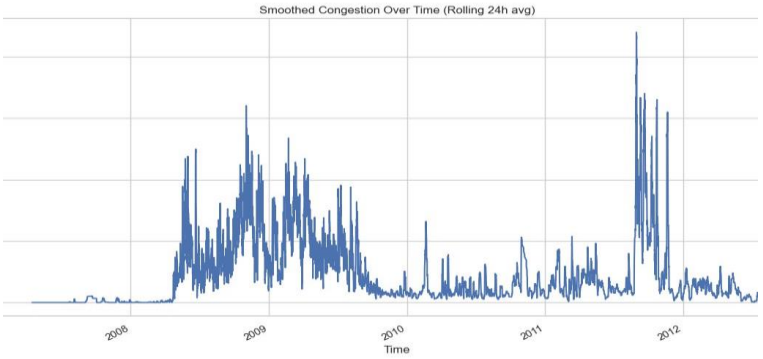
3.4 Heatmap: Hour vs Weekday

Helps in identifying congestion trends based on both weekday and hour.



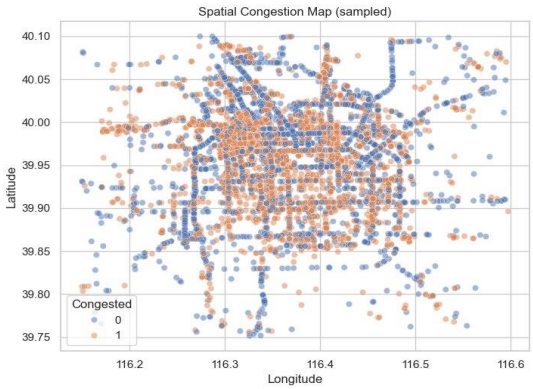
3.5 Congestion by Hour

Distinguishes between congested and uncongested points for each hour.



3.6 Spatial Congestion Map

A scatterplot showing congested and free-flow points in Beijing.

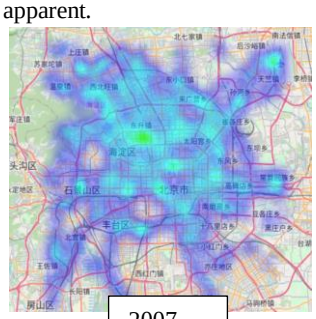


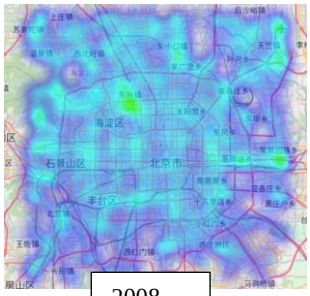
3.7 Annual and Seasonal Heatmap Visualization

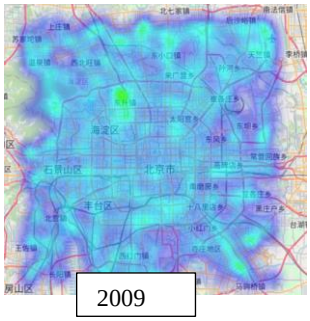
To gain deeper insights into the evolution of traffic congestion in Beijing over time, annual and seasonal heatmaps were generated using the spatial distribution of GPS points labeled as congested. These heatmaps help identify not only high-density urban hotspots but also observe how congestion patterns shift throughout different years and seasons.

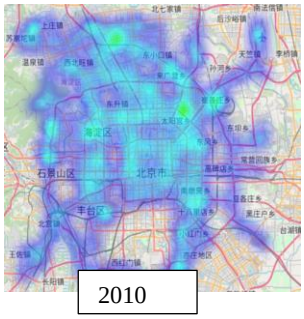
Annual Heatmaps (2007–2012):

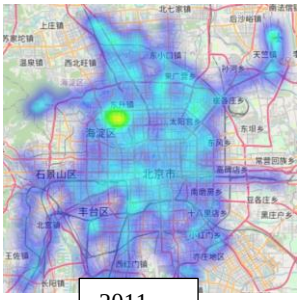
Each year's congestion data was visualized as a heatmap of Beijing, highlighting the areas with the highest frequency of congestion points. These yearly maps allow analysts to compare the expansion or reduction of traffic hotspots across six years. As the number of GPS-enabled taxis increased and road infrastructure evolved, visible changes in congestion distribution became apparent.

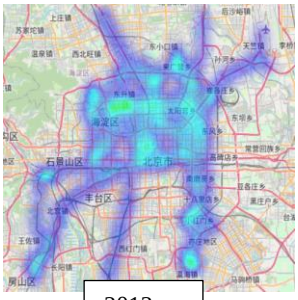








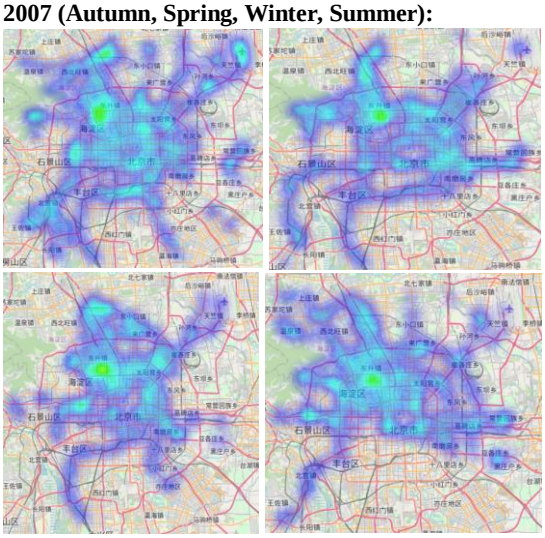




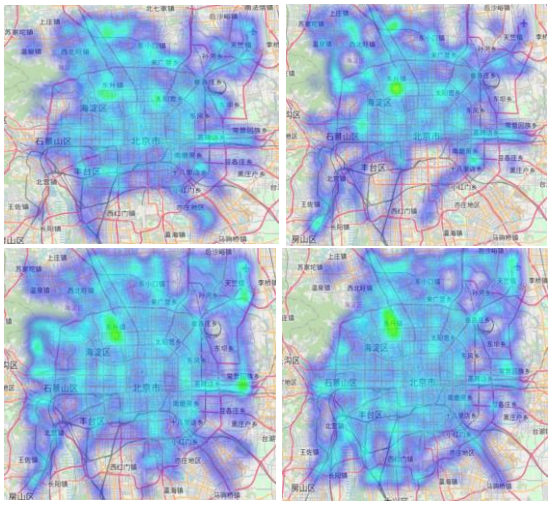
- **Example Observations:**
  - In early years (2007–2008), congestion is mainly clustered around central business districts.
  - By 2011–2012, suburban congestion appears more frequently, reflecting urban sprawl.

Seasonal Heatmaps per Year:

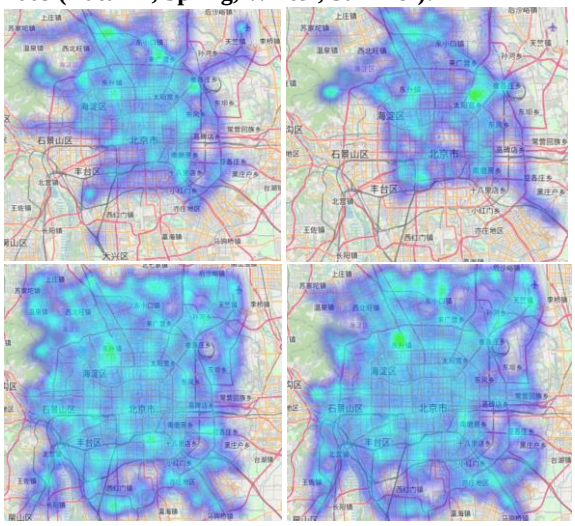
2007 (Autumn, Spring, Winter, Summer):



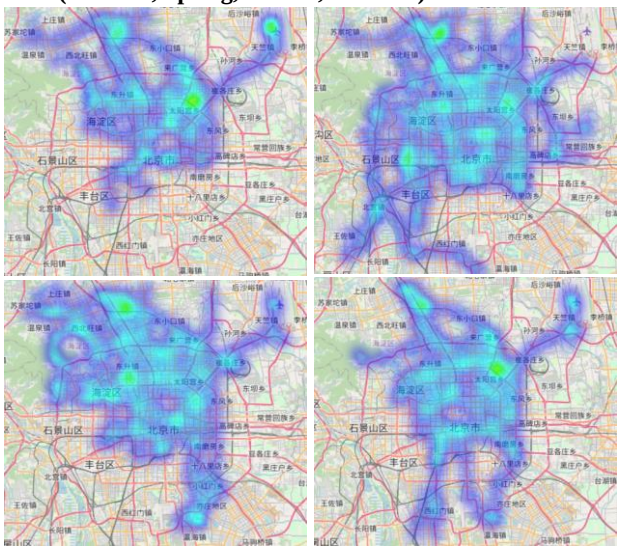
2008 (Autumn, Spring, Winter, Summer):



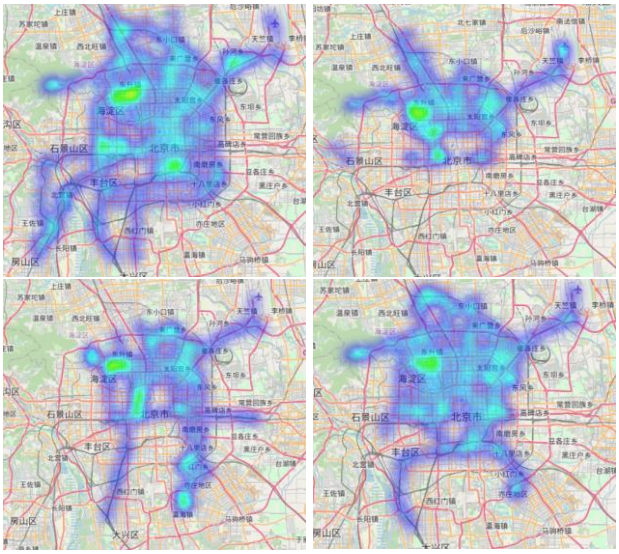
2009 (Autumn, Spring, Winter, Summer):



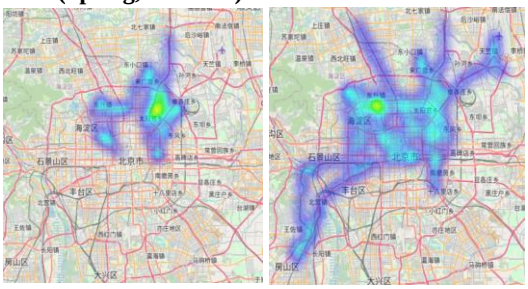
2010 (Autumn, Spring, Winter, Summer):



2011 (Autumn, Spring, Winter, Summer):



2012 (Spring, Summer):



For a more fine-grained temporal analysis, each year was further divided into four seasons (Spring: Mar–May, Summer: Jun–Aug, Autumn: Sep–Nov, Winter: Dec–Feb). Seasonal heatmaps reveal the influence of external factors such as tourism, school vacations, and weather on traffic flow.

- **Spring:** Moderate congestion, often increasing toward May.
- **Summer:** Elevated congestion in touristic zones, with variability due to holidays.
- **Autumn:** Relatively balanced flow; congestion often linked to work and school routines.
- **Winter:** Central congestion dominates, but reduced activity is noticeable during extreme cold or during major holidays (e.g., Chinese New Year).

Visualization Method:

- All heatmaps were generated using a consistent grid size (latitude and longitude rounded to 0.01).
- Each cell's congestion intensity is color-coded (e.g., yellow to red) based on the number of congested points.
- Heatmaps are overlaid on Beijing's city map for geospatial context.

These visualizations serve as an important foundation for understanding spatio-temporal congestion dynamics, supporting better model training and traffic forecasting decisions.

4. Congestion Labeling

Each point was assigned to a grid cell (based on latitude/longitude rounded to 0.01). If more than 100 points appeared in a grid-hour block, it was marked as congested (label=1). This binary label serves as the target for model training.

(Insert sample table showing grid cell aggregation)

5. Model Training

A Random Forest Classifier was trained using the following features:

- Latitude
- Longitude
- Hour
- Month
- Weekday Number

Training Statistics:

- Unique training samples: 18238627
- Accuracy on test set: 87%
- Model saved to: congestion_model.pkl

```python
def train_and_save_model(df, model_path="congestion_model.pkl"):
    print("Preparing training data...")
    df['WeekdayNum'] = df['Timestamp'].dt.weekday
    features = ['Latitude', 'Longitude', 'Hour', 'Month', 'WeekdayNum']
    target = 'Congested'
    df_unique = df[features + [target]].drop_duplicates()
    print(f"Total unique points for training: {len(df_unique)}")
    df_sampled = df_unique.sample(n=min(18238627, len(df_unique)), random_state=42)
    print(f"Using {len(df_sampled)} points for training")
    X = df_sampled[features]
    y = df_sampled[target]
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
    model = RandomForestClassifier(n_estimators=100, n_jobs=-1, random_state=42)
    print("Training model...")
    for _ in tqdm(range(1)):
        model.fit(X_train, y_train)
    acc = model.score(X_test, y_test)
    joblib.dump(model, model_path)
    print(f"Model saved to {model_path}. Accuracy: {acc:.2%}")
    return model

def load_or_train_model(df, model_path="congestion_model.pkl"):
    if os.path.exists(model_path):
        print(f"Loading model from {model_path}...")
        return joblib.load(model_path)
    else:
        return train_and_save_model(df, model_path)
```


Submitted to Beijing Institute of Technology – Department of Computer Science and Technology.

6. Congestion Prediction from Address Inputs

The user can input two place names (e.g., "Beijing Institute of Technology" to "Forbidden City") and a date-time (e.g., 2025-06-21 09:00). The system performs:

1. Geocoding using Nominatim (OpenStreetMap)
2. Route generation using OpenRouteService
3. Congestion prediction for each point along the route
4. Visual rendering with color-coded markers

Example Input:

Enter route as: Place1, Place2, YYYY-MM-DD HH:MM

Example:

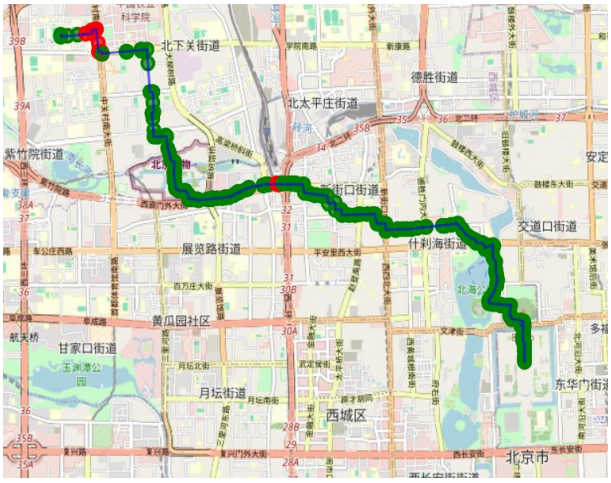
> Beijing Institute of Technology, Forbidden City, 2025-06-21 09:00

Beijing Institute of Technology -> Coordinates: 39.9578214, 116.3098236715735

Forbidden City -> Coordinates: 39.91727565, 116.39076940577283

Output Map:

- Green: Clear segments
- Red: Congested segments



7. Conclusion

This tool successfully demonstrates how historical GPS trajectory data can be leveraged for traffic congestion prediction. By integrating address geocoding, map routing, and machine learning, it provides a practical and user-friendly interface for traffic forecasting.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to the School of Computer Science and Technology at Beijing Institute of Technology for providing the academic environment and support necessary to carry out this research. We are also thankful to Microsoft Research Asia for making the GeoLife GPS Trajectory Dataset publicly available, which served as the foundation for our

traffic congestion prediction study. Their contributions to open research data are invaluable to the academic community.

REFERENCES

[1] Gao, H., Yang, Y., Huang, L., Wang, Y., Jia, B., Yang, F., & Zhu, Z. (2018). Trajectory Data-Driven Pattern Recognition of Congestion Propagation in Road Networks. In *Lecture notes in computer science* (pp. 199–211). https://doi.org/10.1007/978-3-030-05054-2_15

[2] Zheng, Y. (2015). Trajectory data mining. *ACM Transactions on Intelligent Systems and Technology*, *6*(3), 1–41. https://doi.org/10.1145/2743025

[3] Oleś, A. K. (2024). *openrouteservice: An "openrouteservice" API Client* [Dataset]. https://doi.org/10.32614/cran.package.openrouteservice

[4] Zheng, Y., Fu, H., Xie, X., Ma, W., & Li, Q. (2011). Geolife GPS trajectory dataset - User Guide. *IEEE Data Engineering Bulletin*. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/User20Guide-1.2.pdf