Chapter 3: Methodology

Sun Qi
MCST-1043

3.1 Overview

This chapter describes in detail the research methods used in the Johor Bahru traffic flow analysis project. When analyzing the traffic flow of Johor Bahru, Malaysia, it is necessary to collect a variety of data from multiple channels based on the specific characteristics of the city, and conduct an overall analysis based on traffic patterns and different regions. In this process, several steps are required, and the methods used in each step are different. The following will briefly describe the methods involved in each step.

3.2 Methods for clarifying the range of selected data

Before collecting data, the selected scope needs to include important traffic arteries and important transportation hubs in Johor Bahru, such as bridges, viaducts, highways, intersections leading to large shopping malls, tourist attractions, municipal departments, etc. At the same time, it is necessary to determine the time range, such as the whole day or commuting time, the transportation modes involved, such as private cars, buses, motorcycles, bicycles, etc. Finally, weather conditions need to be included.

3.3 Data Collection Methods

Data analysis requires rigor, so the source of data is particularly important. Wrong data will have wrong consequences on the final result.

To collect traffic flow data in Johor Bahru, the following methods can be used

- Manual counting is performed through monitoring records on traffic lights. The advantage of this method is that the recorded data is more accurate and reliable, but it is time-consuming and labor-intensive, and it will waste human resources.
- Automatic counting is performed through cameras installed at key intersections. This method solves the problem of manual counting,

which is labor-intensive and records quickly, but the corresponding problem is not accurate enough

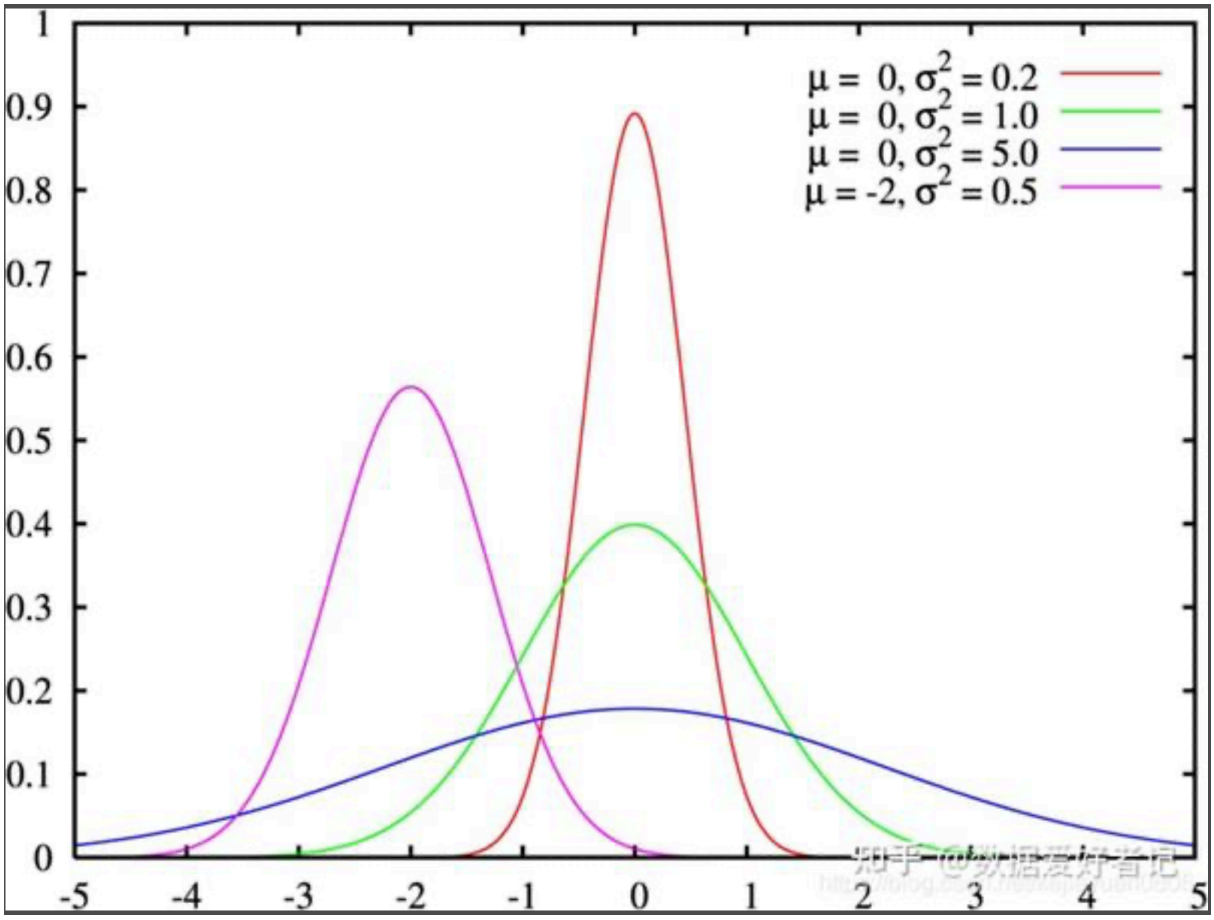
- Counting traffic at intersections through drones or counting vehicles
- Collecting data through third-party data platforms (such as Google Maps, Waze platforms)
- Directly collecting data from government agencies and traffic management departments

3.4 Data preprocessing methods

The collected complex information needs to be summarized and sorted first to extract useful information.

- Data anomalies

Outliers in data refer to values that are beyond or below the normal range. There are several common methods for dealing with outliers, using simple statistical methods, the 3σ principle, normal distribution, etc.



- Handling missing values in data

In the process of data collection, it is inevitable to encounter missing values. The common method for missing values is to delete them directly, use the mean, median, and median in statistics to supplement the missing values, or mark the missing values and find the corresponding values in other data to supplement them.

- Remove duplicate values

When collecting multiple data sets, there is a high probability of encountering duplicate data. To remove duplicate data, you can use Power BI to process the data to remove duplicate values and use Python scripts to process the data.

- Unified data format

In order to make the data beautiful and neat, the data format needs to be processed. Unifying the units, time format, and font size are all good methods.

3.5 Data analysis methods

3.5.1 Statistical analysis

The processed data can be analyzed using statistical methods, such as calculating the variance, mean value, and peak flow of the data. At the same time, the flow of different time periods can be analyzed for comparison.

3.5.2 Data visualization

Another common solution is data visualization, which integrates data into various pie charts, bar charts, and line charts for people to observe and record.

3.6 Model selection and construction method

For the collected traffic flow data, various models can be constructed and classified according to the available scenarios.

- Time series model

This model is suitable for analyzing the trend of traffic flow in the future.The following methods can be used: ARIMA (autoregressive integrated moving average model), SARIMA (seasonal ARIMA),

Exponential Smoothing (exponential smoothing method), LSTM (long short-term memory network), GRU (gated recurrent unit).

- Specific classification model

This model is suitable for judging whether the traffic is congested. The methods that can be used are: logistic regression, decision tree, support vector machine (SVM), gradient boosting model (XGBoost/ LightGBM/CatBoost).

- Clustering classification model

This model is suitable for discovering traffic patterns. The methods that can be used are K-Means, DBSCAN (density clustering), and Gaussian mixture model (GMM).

3.7 Subsequent summary and collation methods

Problems encountered should be recorded to prevent the possibility of recurrence next time

3.8 References

- Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). Time Series Analysis: Forecasting and Control. Wiley.
- Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice. OTexts.
- Wu, Y., & Tan, H. (2016). “Short-term traffic flow prediction with LSTM neural network.” 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC).
- Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
- Chen, T., & Guestrin, C. (2016). “XGBoost: A scalable tree boosting system.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Zhang, Z., Wang, J., & Sun, H. (2020). “Traffic flow prediction using machine learning and comparative analysis.” Transportation Research Part C: Emerging Technologies.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). “A density-based algorithm for discovering clusters in large spatial databases with noise.” Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD).

- Huang, L., & Zhang, Y. (2018). “Clustering traffic patterns for urban road networks.” Transportation Research Record: Journal of the Transportation Research Board.
- Jain, A. K. (2010). “Data clustering: 50 years beyond K–means.” Pattern Recognition Letters.
- Barceló, J. (2010). Fundamentals of Traffic Simulation. Springer.
- Krajzewicz, D., Erdmann, J., Behrisch, M., & Bieker, L. (2012). “Recent development and applications of SUMO – Simulation of Urban Mobility.” International Journal On Advances in Systems and Measurements.
- Nagel, K., & Flötteröd, G. (2012). “Agent–based traffic assignment: Going from trips to behavioral travelers.” Proceedings of the 91st Annual Meeting of the Transportation Research Board.