3st Assignment COMPX523-25A

June 11, 2025

Contents

1 Abstract

2 Introduction

Increasing urbanisation around the world has resulted in an increasing demand for public transport services, particularly in developing countries. However, financial and infrastructure constraints have made it difficult for governments to match this growing demand (Motta, Da Silva, & SANTOS, 2013). The public transport crowding that has resulted from this causes increased stress, increased unreliability of public transport services, and reduced productivity for passengers (Tirachini, Hensher, & Rose, 2013).

Intelligent Public Transport Systems (IPTS) is a nascent field that attempts to address issues with public transport using machine learning (ML) (Zear, Singh, & Singh, 2016). Accurate prediction of public transport volumes is one way that ML can be leveraged to improve public transport and reduce crowding. If passenger volumes are known ahead of time, it may be possible for public transport agencies to optimise the availability of buses and trains in order to match demand and thus reduce crowding. Passengers can also directly benefit from foreknowledge of the demand at a given station or bus route, as they can plan their travel to avoid crowding (di Torrepadula, Napolitano, Di Martino, & Mazzocca, 2024).

The data stream mining paradigm is designed for large quantities of data which arrive continuously, in a setting where memory and computational power might be constrained (Bifet, Gavalda, Holmes, & Pfahringer, 2023). Public transport data fits this paradigm, thus the field of data stream mining provides us with a set of ML algorithms that are appropriate for deployment in this context.

The setting of this study is Salvador, the fifth largest city in Brazil. Compared to other metropolitan areas in Brazil, there is a higher reliance on public transport in Salvador due to a broadly lower socioeconomic status. The majority of public transport in Salvador is by bus, although some passengers take the Salvador Metro ("BRICS Cities: Facts & Analysis 2016", 2016). The Salvador Urban Network Transportation (SUNT) dataset contains high granularity information about passenger volumes at bus stops and stations across Salvador, between March and May 2024 (dos Santos Ferreira et al., 2025).

The aim of this study is to investigate data stream ML models by which the load of passengers on buses in Salvador can be estimated. Such models, if deployed, could be used by agencies or passengers to accurately estimate the moment-by-moment demand at each of these bus stops. These models could also be inspected using explainable AI techniques to gain insights into what factors influence public transport demand in

Salvador.

3 Background

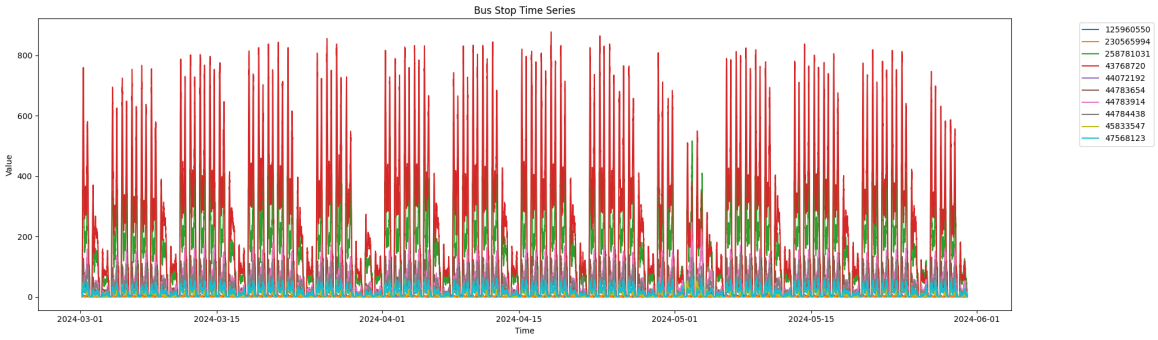
This study analyzes time series data provided by public transport datasets from 10 stations in a large Brazilian city. The datasets include boarding_03-05_2024.csv, landing_03-05_2024.csv, and loader_03-05_2024.csv, which contain accumulated data at 5-minute and 30-minute intervals. These files respectively represent boardings, landings, and loader (bus occupancy), as shown in the table 1. The time range spans from 05:00:00 on 1 March, 2024, to 00:55:00 on 31 May, 2024.

Table 1: Data Source

Data	**Source**
Boarding	Data mining techniques from Automatic Vehicle Location (AVL), Automatic Fare Collection (AFC), and General Transit Feed Specification (GTFS) systems
Alighting	Trip chaining method
Bus occupancy	Ratio of boardings and alightings per vehicle

This study uses the bus occupancy of 5-minute time series data (Figure 1) and finds that some bus stops experience heavy passenger flows during certain periods, while others often remain sparsely used.

Figure 1: Bus Occupancy



The reasons for the large passenger flows at certain stations shown in Table 2:

Table 2: Bus Stop

Stop ID	**Location**	**Observations**
44783654	In front of the Federal University of Bahia (UFBA)	High volume of students and university staff
43768720	Lapa Station	One of the city's main public transport terminals
230565994	Itapuã Lighthouse Beach	Passenger flow varies throughout the day; higher on weekends and holidays
125960550	Arena Fonte Nova surroundings	Demand influenced by sports and cultural events
45833547	Near Manoel Barradas Stadium	Sharp increases in demand during game days
44784438	Near the Ferry Boat terminal	Primarily serves intercity and cross-bay travelers
47568123	Near a major shopping mall	High commercial traffic throughout the day
44072192	Close to Castro Alves Theater	Passenger flow increases during performance and event hours
258781031	Salvador Bus Station	Central hub for urban and intercity bus routes
44783914	Lacerda Elevator tourist area	Located in a busy commercial and tourist district

The surge in passenger volume is caused by some objective patterns. For example, during an annual event held near a specific stop, such as a sports competition or cultural celebration, a large number of people tend to take buses to that location. As a result, the bus routes near the stop become unusually busy and may even lead to traffic congestion. Passengers who are unaware of the situation might further worsen the congestion.

4 Proposal

To address the problem of congestion caused by high passenger flow, we propose a station recommendation system, which analyzes the busyness of transit stops and provides real-time suggestions for alternative stations. In this way, passengers can avoid overcrowded stops, thereby reducing congestion and improving travel efficiency.

By integrating the CapyMOA online learning algorithm, the system is able to adapt dynamically to changes in passenger flow and offer station recommendations.

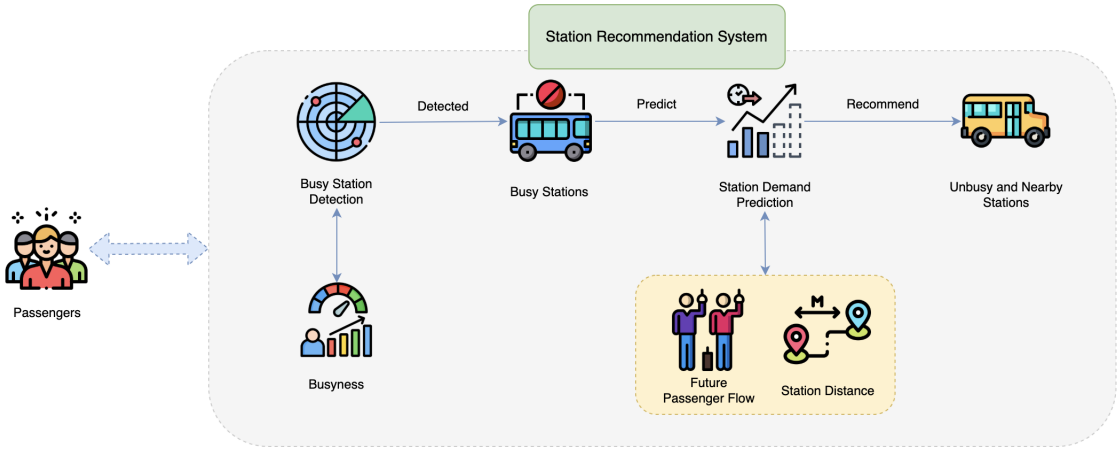
4.1 System Design

The station recommendation system consists of two core tasks: First, identifying busy stations by using 5-minute aggregated data to detect stations with high passenger flow in real time. Second, recommending alternative stations that are nearby and have lower traffic, in order to balance the transportation load.

4.2 System Achitecture

Passengers can use the station recommendation system to check whether a stop is crowded. If the queried station is busy, the system predicts the station with the lowest passenger flow at the current time and calculates the nearest stop to recommend. This helps passengers avoid congested stops, as shown in the Figure 2.

Figure 2: Achitecture



5 Experiments

5.1 Data Preparation

This study analyzes both 5-minute and 30-minute aggregated data. The 5-minute data is more suitable for monitoring real-time passenger flow. Based on the two main tasks of the system, we propose building two separate models: the Busy Station Detection (BSD) Model and the Station Demand Prediction (SDP) Model.

5.1.1 Features and Label Design

The BSD task is defined as a classification problem. The goal is to generate a set of features for each station at a specific time, which will be used for model training. The designed features include:

1. Boarding, Landing, and Loader: These are the number of boardings, landings, and onboard passengers at the current station and time. If the data is missing, the value is set to 0.

2. Hour, Day of Week, and Is Weekend: Time-related features, including the hour of the day, day of the week, and whether it is a weekend.

3. Nearby Flow: The total number of onboard passengers at the three nearest stations at the current time.

The label is binary classification, indicating whether a station is busy (0 means not busy, 1 means busy). According to local regulations and vehicle types in Salvador, Brazil, the maximum passenger capacity of public buses depends on the model:

1. Standard single-unit buses: about 70 to 90 people (including seated and standing passengers).

2. Articulated buses (articulado): about 120 to 150 people (longer buses with more capacity).

Therefore, we set an absolute minimum passenger threshold of 100 people to exclude stations with very low traffic and reduce noise.

We use the Z-score standardization method. Assuming the data follows a normal distribution, a Z-score greater than 1.5 corresponds to approximately the 93.3rd percentile, meaning the station's passenger flow is higher than 93.3% of all stations.

For each station's passenger flow x_i (total passengers), the Z-score is calculated as:

$$z_i = \frac{x_i - \mu}{\sigma + \epsilon} \tag{1}$$

Where:

1. μ: Mean passenger flow of all stations;

2. σ: Standard deviation of all station passenger flows;

3. ϵ: A very small value (e.g., 1×10^{-6}) to avoid division by zero.

Based on this analysis, we define the set of busy stations S as those that meet both of the following conditions:

$$S = \{i \mid z_i > 1.5 \land x_i > 100\} \tag{2}$$

1. $z_i > 1.5$: The passenger flow is significantly higher than the average (Z-score threshold);

2. $x_i > 100$: The absolute number of passengers exceeds the minimum threshold

The SDP Model is a regression learner. It uses the same features as the BSD Model, but the label is the passenger flow in the next time period, which is of type 'numeric'. For example, given a time period (timestamps), the target is to attain the passenger flow in the next time period, that is:

$$target\_timestamp = timestamp + 1 \tag{3}$$

5.1.2 Calculate Distance

To recommend alternative stations when one becomes busy, we need to find replacement stations within 1000 meters. To do this, we calculate the distance between two stations. Although the Haversine formula is more accurate for geographic distance, we simplify the calculation using the Euclidean distance formula.

Let:

1. lat_1, lon_1: Latitude and longitude of station 1;
2. lat_2, lon_2: Latitude and longitude of station 2;

The simplified Euclidean distance d (in degrees) is:

$$d = \sqrt{(lat_1 - lat_2)^2 + (lon_1 - lon_2)^2} \tag{4}$$

Since the result is in degrees, we need to convert it into meters. Near the equator:

$$1^\circ \approx 111{,}000 meters \tag{5}$$

So the final distance in meters:

$$D = d \times 111{,}000 \tag{6}$$

5.1.3 Recommend Alternative Stations

To predict and recommend nearby stations with lower future passenger flow, the system follows these steps:

1. Calculate the distance between the current busy station and all other stations, and sort them from nearest to farthest.

2. Select the top five nearest stations, and for each station:

   (a) Extract its features.

   (b) Predict the station's future passenger flow using the SDP Model, and record the result.

   (c) Compute the distance between the busy station and the candidate station.

3. Generate a set of recommended stations, where each station includes the following information:

   (a) Station name

   (b) Predicted passenger flow

   (c) Distance from the busy stationdistance

   (d) A final score, calculated as:

$$score = \frac{predicted\_flow + 1}{distance} \tag{7}$$

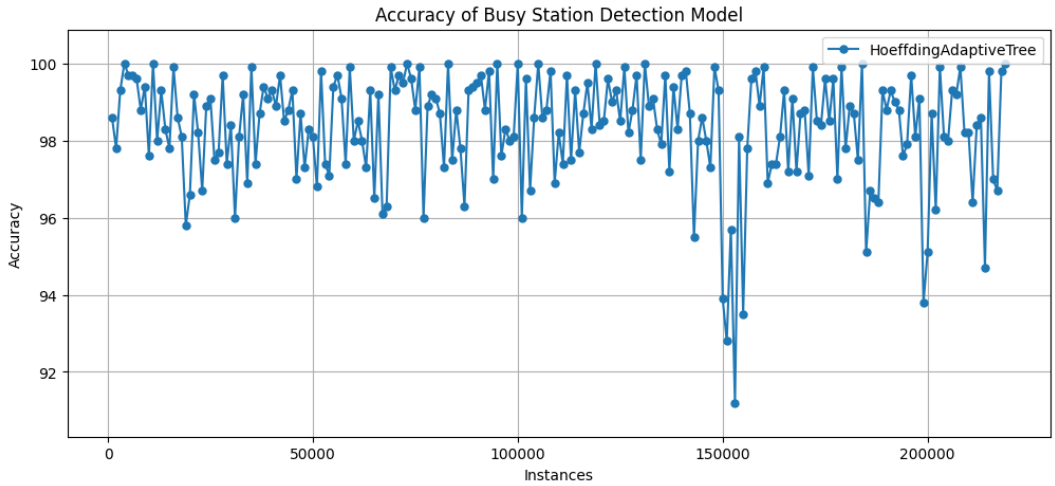
   The lower the score, the better the recommendation. A lower score indicates closer distance and lower predicted flow.

4. Finally, sort all candidate stations in ascending order by their score, and return the top two recommended stations.

5.2 Results

The BSD Model uses Hoeffding Adaptive Tree (HAT). HAT is an adaptive incremental decision tree algorithm designed for data stream environments. It is especially suitable for non-static data that changes over time. When classifying station status (busy or not busy) in real time, HAT can quickly adapt to sudden changes in passenger flow patterns.

Figure 3: Accuracy of Busy Station Detection Model



The accuracy in most windows ranges between 96% and 100% shown in Figure 3, with only a few windows experiencing brief drops, but overall fluctuations are minor. There are a few dips, such as around the 150,000 and 180,000 sample marks, where accuracy drops significantly, reaching as low as 91%. These declines may correspond to concept drift in the data distribution, but the model quickly recovers afterward, indicating that HAT has strong adaptability.

The SDP Model uses FIMT-DD (Fast Incremental Model Tree with Drift Detection). FIMT-DD is an incremental regression tree algorithm for predicting continuous numeric values. When predicting future passenger demand (e.g., the next 5-minute period), FIMT-DD can capture both trend changes and unexpected fluctuations effectively.

Figure 4: Adjusted R2 of Station Demand Prediction Model

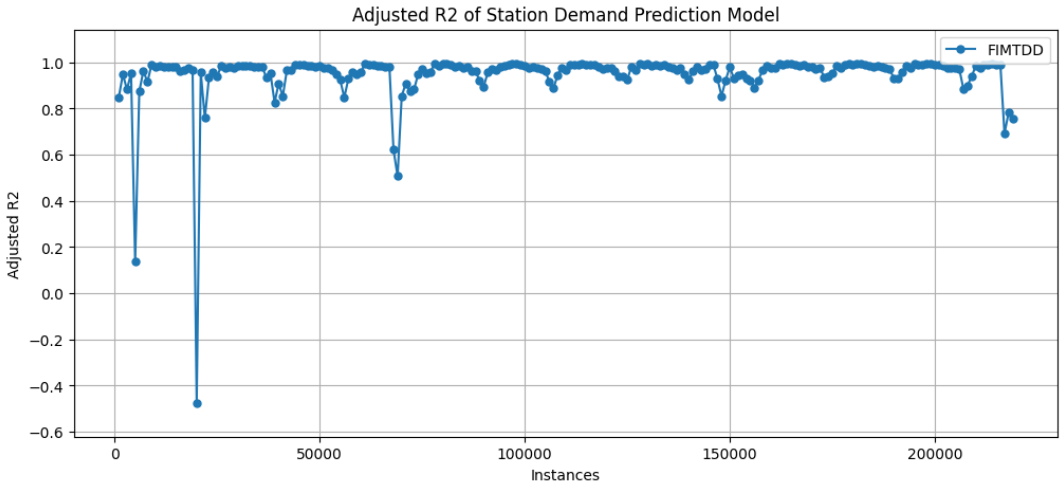
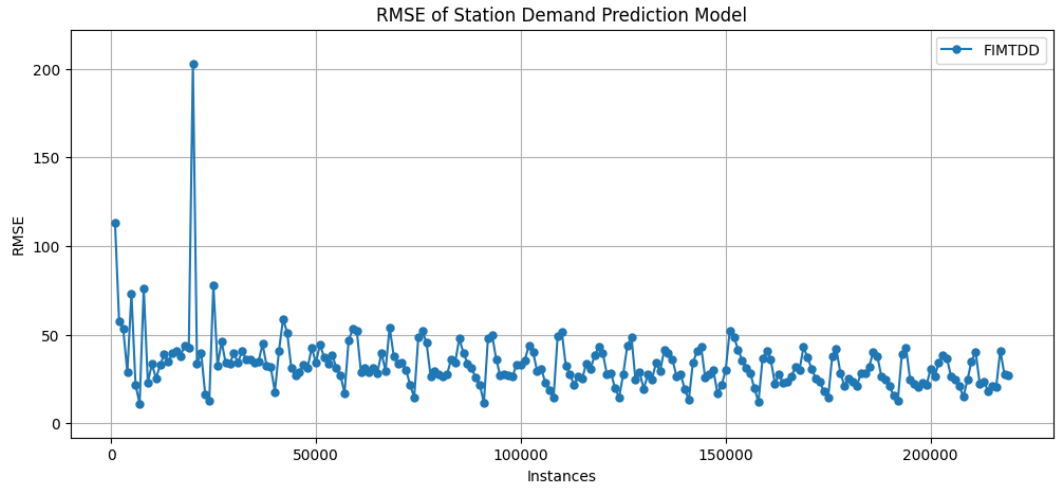


Figure 5: RMSE of Station Demand Prediction Model



The Adjusted R^2 remains above 0.9 most of the time, even approaching 1, which demonstrates high prediction accuracy shown in Figure 4. At a few time points (such as around 10,000, 65,000, and 210,000), Adjusted R^2 drops sharply for a short period, with the lowest point reaching -0.5. However, the model quickly regains a high level of fit, reflecting FIMTDD's strong adaptability and robustness. Except for a few points (such as the first RMSE peak at about 210), RMSE remains in the 20–60 range for most of the time shown in Figure 5. The overall low RMSE indicates that the prediction results have small fluctuations and controllable errors.

5.2.1 Run real-time station recommendation system simulation

We simulate a real-time station recommendation system. First, a random time is selected during working hours between 7:00 AM and 7:00 PM. Then, the BSD model is used to predict the busyness level of all stations.

If any busy stations are detected, the system will search for alternative stations within 1000 meters that are closer in distance and have lower predicted passenger flow.

If such less crowded stations exist, the system will recommend at least two stations to passengers.

A sample output of the simulation is shown below:

Timestamp	**Log Details**
2024-05-29 11:50:00	• Detected 1 busy station: `Avenida Vale Do Tororo 327 Salvador - Bahia 40050 Brasil` • Recommended alternative: `Avenida Vasco da Gama 4274 - Brotas Salvador - BA Brasil - 44784438` • Distance: 843m
2024-05-29 11:55:00	• Detected 1 busy station: `Avenida Vale Do Tororo 327 Salvador - Bahia 40050 Brasil` • Recommended alternative: `Avenida Vasco da Gama 4274 - Brotas Salvador - BA Brasil - 44784438` • Distance: 843m

2024-05-29 12:00:00	• Detected 1 busy station: `Avenida Vale Do Tororo 327 Salvador - Bahia 40050 Brasil` • Recommended alternative: `Avenida Vasco da Gama 4274 - Brotas Salvador - BA Brasil - 44784438` • Distance: 843m

Table 4: Simulation Results Summary

Timestamp	**Busy Stations Count**	**Recommendations Count**
2024-05-29 11:50	1	1
2024-05-29 11:55	1	1
2024-05-29 12:00	1	1
2024-05-29 12:05	1	1
2024-05-29 12:10	1	1
2024-05-29 12:15	1	1
2024-05-29 12:20	1	1
2024-05-29 12:25	1	1
2024-05-29 12:30	1	1
2024-05-29 12:35	1	1
2024-05-29 12:40	1	1
2024-05-29 12:45	1	1
2024-05-29 12:50	1	1
2024-05-29 12:55	1	1
2024-05-29 13:00	1	1
2024-05-29 13:05	1	1
2024-05-29 13:10	1	1
2024-05-29 13:15	1	1
2024-05-29 13:20	1	1
2024-05-29 13:25	1	1
2024-05-29 13:30	1	1
2024-05-29 13:35	1	1
2024-05-29 13:40	1	1
2024-05-29 13:45	1	1

6 Discussion

By using the BSD and SDP models to recommend alternative stations, this system aims to help solve urban traffic congestion problems. In our simulation experiments using 5-minute interval data, the results were promising — the system could effectively detect busy stations and recommend alternative stations.

However, in cases of sudden changes in passenger flow, such as during public holidays when the number of passengers increases sharply, further analysis and optimization of the system are needed to improve its response to such abnormal situations.

7 Conclusion

This study explores dynamic station recommendation using online learning algorithms to analyze bus station data. The proposed method takes full advantage of CapyMOA's streaming data processing capabilities, enabling real-time adaptation to changing passenger flow and avoiding the delay caused by batch learning.

According to the experimental results:

1. The cumulative accuracy of the busy station detection model reached 98.330
2. The cumulative adjusted R^2 of the station demand prediction model was 0.974.

Both the BSD and SDP models achieved high overall accuracy with low fluctuation. Even when data changed suddenly, the models could quickly recover and adapt.

8 CRediT Statement

References

Bifet, A., Gavalda, R., Holmes, G., & Pfahringer, B. (2023). *Machine learning for data streams: with practical examples in moa*. MIT press.

Brics cities: Facts & analysis 2016. (2016).

di Torrepadula, F. R., Napolitano, E. V., Di Martino, S., & Mazzocca, N. (2024). Machine learning for public transportation demand prediction: A systematic literature review. *Engineering Applications of Artificial Intelligence*, *137*, 109166.

dos Santos Ferreira, M. V., de Souza, M. C., Rios, T. N., da Costa Fernandes, I. F., Andrade, D. O., Gama, J., ... Rios, R. (2025).

Salvador urban network transportation (sunt). Mendeley Data. Retrieved from https://data.mendeley.com/datasets/85fdtx3kr5/1 doi: 10.17632/85fdtx3kr5.1

Motta, R. A., Da Silva, P. C. M., & SANTOS, M. P. D. S. (2013). Crisis of public transport by bus in developing countries: a case study from brazil. *International journal of sustainable development and planning*, *8*(3), 348–361.

Tirachini, A., Hensher, D. A., & Rose, J. M. (2013). Crowding in public transport systems: effects on users, operation and implications for the estimation of demand. *Transportation research part A: policy and practice*, *53*, 36–52.

Zear, A., Singh, P. K., & Singh, Y. (2016). Intelligent transport system: A progressive review.