Bus Stop Attractiveness Scoring Using Public Transport Ridership Data

Ege Altıok
Graduate School of Natural and Applied Sciences
Dokuz Eylül University

I. Introduction

Public transport is an essential component of daily life, which is why continuous efforts are made to make it more sustainable, accessible, and affordable. Despite these efforts, many cities struggle with people preferring private to public transportation, causing a variety of problems.

Recent advancements in data collection, especially with Automatic Fare Collection (AFC) systems and Smart Card Data (SCD), have been a key factor to analyse and understand passenger behaviour to optimise public transportation. Smart card transactions reveal when and where passengers board or alight buses, offering insights into travel patterns and network dynamics.

This study aims to develop a **Stop Attractiveness Scoring (AS)** framework using origin–destination (OD) ridership records, encompassing multiple lines. Unlike optimisation-based works that focus on improving route planning and network efficiency, this approach prioritises human-oriented understanding of service accessibility by analysing how and why people choose specific stops. By doing so, it aims to provide personalised, data-driven suggestions for passengers to help them make more informed travel decisions and improve their public transport experience.

II. Literature Survey

Earlier studies on public transport optimisation [2] emphasised algorithmic efficiency in route planning through Dijkstra-based methods and transfer-minimisation strategies. While these approaches were effective in improving network performance, they largely treated passengers as static objects rather than active decision-makers. Most existing studies therefore focused on route-level optimisation or network-wide accessibility, overlooking how individual stops influence passengers' travel choices.

More recent work, such as Luo et al. [3], introduced joint passenger flow prediction models to understand multimodal travel interactions. This study showed the potential of using large-scale Automatic Fare Collection (AFC) data for uncovering hidden mobility patterns.

However, the literature remains limited when it comes to analysing stop attractiveness from a passenger perspective. This study fills this research gap by analysing passenger preferences through Smart Card Data, focusing on three key metrics: boarding frequency, effective destination diversity, and attraction-weighted access scores. These metrics provide a quantitative framework for understanding stop attractiveness from a passenger-centric perspective. Ultimately, the focus shifts from *"How to make a route more optimal?"* to *"What makes each stop attractive to passengers?"*, offering transparency into the decision-making factors that shape public transit usage.

III. Methodology

A. Data Source

This study uses the hourly ridership origin–destination (OD) records from the Bay Area Rapid Transit (BART) system in California, United States [4]. Each record represents the number of passengers travelling from a *source* station to a *destination* station within a specific hour. The dataset is provided in a comma-separated values (CSV) format with the following fields:

Variable	Description	Value
date	service date	YYYY-MM-DD
hour	hour of day	integer (0 – 23)
source	source station	4-letter code
destination	destination station	4-letter code
passengers	number of passengers	integer ...

The station reference table reports the selected attributes used to represent each transit station in the analysis. It includes a unique station identifier, descriptive naming fields, and geographic coordinates, which together support consistent station matching across datasets and enable spatial computations.

Variable	Description	Value
code	station code (unique)	2-letter uid
name	station name	up to 100 chars
abbreviation	station abbreviation	4-letter code
latitude	station latitude	float
longitude	station longitude	float

B. Data Preparation

The data preparation methodology consists of three principal stages that transform the raw origin–destination (OD) records into an analysable format.

- **Filtering and Cleaning**: The dataset is first cleaned by removing invalid entries, such as records where the source and destination stations are identical.

- **Station-Level Aggregation**: Records are grouped by source station to compute station-level indicators, including total boardings and destination-choice dispersion (effective destination diversity).
- **Day-Level Aggregation**: The station-level indicators are aggregated by date to facilitate day-level and time-series analyses.

C. Attractiveness Scoring Framework

For each station S_i, the **Attractiveness Score (AS)** is computed as a weighted aggregation of **three** metrics derived from the aggregated OD records:

$$AS_i = w_1 \cdot Board_i + w_2 \cdot EffDst_i + w_3 \cdot Access_i$$

Where:

- $Board_i$: number of boardings at station S_i, normalised to $[0, 1]$:
$$B_i = \sum_{\text{source}=S_i} passengers$$
$$Board_i = \frac{B_i - \min(B)}{\max(B) - \min(B)}$$
- $EffDst_i$: **effective destination diversity**, which shows how widely passengers distribute their destination choices (i.e. many meaningful options rather than a long tail of rarely used destinations). First, define aggregated OD flow $F_{ij} = \sum passengers$ for trips from S_i to S_j, and the destination choice probability
$$p_{ij} = \frac{F_{ij}}{\sum_j F_{ij}} \quad \text{for } \sum_j F_{ij} > 0$$
Then compute Shannon entropy H_i [1] and its "effective number of destinations" (Hill number of order 1):
$$H_i = -\sum_j p_{ij} \ln(p_{ij}), \qquad D_i = \exp(H_i)$$
Finally normalise to $[0, 1]$:
$$EffDst_i = \frac{D_i - \min(D)}{\max(D) - \min(D)}$$
- $Access_i$: an **attraction-weighted option index** that values access to highly demanded destinations. Let destination attractiveness be the inbound volume:
$$A_j = \sum_{\text{destination}=S_j} passengers$$
Let $dist_{ij}$ be the geographic distance between stations S_i and S_j (computed from station coordinates), and let the distance-decay function be:
$$f(dist_{ij}) = \frac{1}{1 + dist_{ij}}$$
Then:
$$Acc_i = \sum_{j \neq i} A_j \cdot f(dist_{ij}),$$
$$Access_i = \frac{Acc_i - \min(Acc)}{\max(Acc) - \min(Acc)}$$
- w_1, w_2, w_3: weight coefficients to be adjusted.

Note: `TransferFrequency` is not considered in this setting because the OD-hour aggregated format does not provide enough information to reliably identify individual transfer events or sequential legs of a single passenger journey.

IV. References

[1] L. Jost, "Entropy and diversity," *Oikos*, vol. 113, no. 2, pp. 363–375, 2006. DOI: https://doi.org/10.1111/j.2006.0030-1299.14714.x [Online]. Available: https://nsojournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2006.0030-1299.14714.x

[2] A. Bozyigit, G. Alankuş, and E. Nasibov, "A public transport route recommender regarding the least transfer," *Sigma Journal of Engineering and Natural Sciences*, vol. 9, pp. 437–446, 2018.

[3] D. Luo, J. Wang, W. Lu, W. Ding, X. Yan, and H. Yang, "Learning dynamic interaction for multimodal transportation systems: A joint passenger flow prediction approach based on spatiotemporal hypergraph convolution networks," *Transportation Research Part C: Emerging Technologies*, vol. 178, p. 105 257, 2025, ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2025.105257

[4] *Bay area rapid transport ridership data*, https://www.bart.gov/about/reports/ridership, Accessed: 2025-12-28.