MATH 421 - Homework 4
Modeling Proposals for NYC Subway System

Ducheng Tan

October 29, 2025

These are only current extension and ideas I have in mind for now, methodology may vary in the final paper

Idea 1: Modeling the Distribution of Subway Headways

Application Field: This project lies in the field of transportation reliability and applied probability. It focuses on understanding the time gaps between consecutive subway trains, called *headways*, which directly affect passenger waiting times and system efficiency. Modeling these headways statistically helps evaluate service consistency and detect irregularities such as train bunching.
Problem Statement: *What probability distribution best describes the observed headways on a selected NYC subway line, and how do these distributions change across times of day?* By identifying the underlying distribution, we can quantify how predictable the service is and assess deviations from scheduled performance.
Variables and Parameters:

- T_i: observed arrival time of the i-th train at a fixed station.
- $X_i = T_{i+1} - T_i$: observed headway (minutes).
- Parameters: mean headway $\mu = E[X]$, variance $\sigma^2 = Var(X)$.

Neglected Variables: We initially ignore passenger demand, day-of-week variation, and weather conditions, since our primary focus is the statistical distribution of train intervals rather than causal factors.Focus on causal factors may add server complexity to our model.
Data Collection: We would collect real-time arrival data from the MTA's GTFS Realtime API or historical records available on NYC Open Data. The data will be filtered by station and direction, and headways will be computed from consecutive arrival timestamps.
Mathematical Approach: We will fit candidate probability models (Exponential, Gamma, Lognormal) to the empirical headway data using maximum likelihood estimation and compare fits using AIC or KS tests. The analysis will involve basic statistical modeling and parameter inference.

Idea 2: Simple Optimization of Train Scheduling Frequency

Application Field: This project lies in the area of transportation operations and optimization. The balance between service quality (short waiting times) and operating costs (number of trains) is a fundamental question in public transit planning.
Problem Statement: *What is the optimal train frequency (headway H) that minimizes the combined cost of passenger waiting and operational expenses for a given subway line?*
Variables and Parameters:

- H: train headway (minutes, decision variable).
- $C(H)$: total cost function.

- c_1: cost per minute of passenger waiting.
- c_2: operating cost per train per hour.
- $E[W] = \frac{H}{2}$: expected waiting time assuming random passenger arrivals.

Neglected Variables: We ignore factors such as train capacity, maintenance constraints, and stochastic disruptions. These can be incorporated later but are not essential for an initial steady-state optimization.
Data Collection: Data on mean headways and service frequency will be taken from MTA's static GTFS feed (schedules). Average passenger loads or estimated ridership can be incorporated to set the relative weights of waiting cost (c_1) and operating cost (c_2).
Mathematical Approach: We model the total cost as

$$C(H) = c_1 \frac{H}{2} + c_2 \frac{1}{H},$$

and find the optimal headway H^* by minimizing $C(H)$ with respect to H. This provides a simple quantitative rule for balancing service frequency and cost.

Idea 3: Estimating the Probability of Missing a Subway Transfer

Application Field: This project belongs to transportation reliability and applied probability. Transfer timing between lines strongly affects total travel time, and quantifying the chance of missing a connection can improve scheduling and passenger experience.
Problem Statement: *Given two connecting lines at a transfer station, what is the probability that a passenger arriving on one line will miss the next train on the receiving line, given the variability in headways and walking time between platforms?*
Variables and Parameters:

- X: headway (interarrival time) of the receiving line.
- T_{walk}: walking time between platforms (seconds).
- $P_{\text{miss}} = P(X < T_{\text{walk}})$: probability of missing the transfer.

Neglected Variables: We neglect random dwell-time extensions due to crowding and assume T_{walk} is constant for simplicity. These factors can be introduced later as random variables to enhance realism.
Data Collection: Arrival timestamps for both lines at a chosen transfer station will be gathered from the MTA GTFS feed. Headways for the receiving line will be estimated empirically, and the walking time will be measured or approximated from station maps. The walking times are also set to be deterministic for each train transfer at a specific train station. For example, T_{walk} for transferring from the 6 train to the 7 train remains the same and we assume crowding and miscellaneous factors during the walk are ignored. However adding possible variations to the walk times can be extended in our future work as well.
Mathematical Approach: After fitting a distribution (e.g., Gamma) to the receiving line's headways, we compute

$$P_{\text{miss}} = F_X(T_{\text{walk}}),$$

where F_X is the fitted cumulative distribution function. This provides a probabilistic estimate of transfer reliability that can be validated against real data.

Citations

1. data.ny.gov
2. https://www.mta.info/open-data
3. GitHub-Current Papers referenced.