Project Proposal **Utsav Singh (62865753)**

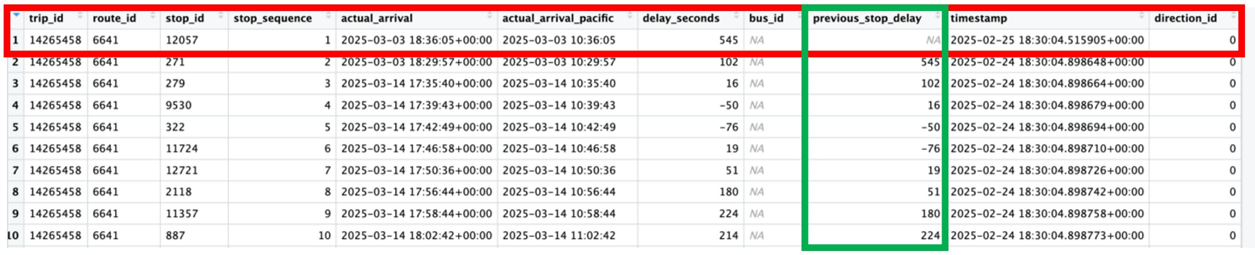
Modeling Delay Propagation in Vancouver Bus Transit Network

GitHub: https://github.com/Utsav02/TranslinkBayes

Introduction: Metro Vancouver's bus transit system, operated by TransLink, is one of the finest in the continent. However, like all other transit systems around the globe, it experiences unpredictable delays due to congestion, weather, and operational inefficiencies. Through this project, I am trying to apply the concepts learned in the course and explore how Bayesian inference can be applied to model delay propagation in a high-frequency bus network of Vancouver using real-time GTFS data from TransLink. The objective is to develop a probabilistic framework for delay prediction while capturing uncertainty, leveraging historical data and hierarchical Bayesian modeling.

Project Theme: My project is heavily influenced by a previous study conducted in Sweden taking a similar Bayesian approach (Rodriguez et al. (2022)). Based on the project proposal themes, this project aligns with the theme of Bayesian regression and time series modeling. It can also be considered a spatial model since the dataset does have longitudes and latitudes, however, my current approach is not using it but I plan to try incorporating it by the project report.

Dataset & Data Collection: I am using real-time and static GTFS datasets from TransLink. Static data comes from TransLink's website (https://www.translink.ca/about-us/doing-business-with-translink/app-developer-resources/gtfs/gtfs-data) and the Real-time data is obtained by a Python script running on my local computer every 5 mins that stores the data in an SQLite database. The project structure is outlined in detail in the ReadMe file on the GitHub Repository.



	trip_id	route_id	stop_id	stop_sequence	actual_arrival	actual_arrival_pacific	delay_seconds	bus_id	previous_stop_delay	timestamp	direction_id
1	14265458	6641	12057	1	2025-03-03 18:36:05+00:00	2025-03-03 10:36:05	545	NA	NA	2025-02-25 18:30:04.515905+00:00	0
2	14265458	6641	271	2	2025-03-03 18:29:57+00:00	2025-03-03 10:29:57	102	NA	545	2025-02-24 18:30:04.898648+00:00	0
3	14265458	6641	279	3	2025-03-14 17:35:40+00:00	2025-03-14 10:35:40	16	NA	102	2025-02-24 18:30:04.898664+00:00	0
4	14265458	6641	9530	4	2025-03-14 17:39:43+00:00	2025-03-14 10:39:43	-50	NA	16	2025-02-24 18:30:04.898679+00:00	0
5	14265458	6641	322	5	2025-03-14 17:42:49+00:00	2025-03-14 10:42:49	-76	NA	-50	2025-02-24 18:30:04.898694+00:00	0
6	14265458	6641	11724	6	2025-03-14 17:46:58+00:00	2025-03-14 10:46:58	19	NA	-76	2025-02-24 18:30:04.898710+00:00	0
7	14265458	6641	12721	7	2025-03-14 17:50:36+00:00	2025-03-14 10:50:36	51	NA	19	2025-02-24 18:30:04.898726+00:00	0
8	14265458	6641	2118	8	2025-03-14 17:56:44+00:00	2025-03-14 10:56:44	180	NA	51	2025-02-24 18:30:04.898742+00:00	0
9	14265458	6641	11357	9	2025-03-14 17:58:44+00:00	2025-03-14 10:58:44	224	NA	180	2025-02-24 18:30:04.898758+00:00	0
10	14265458	6641	887	10	2025-03-14 18:02:42+00:00	2025-03-14 11:02:42	214	NA	224	2025-02-24 18:30:04.898773+00:00	0

Table 1: Head of Stop Delays Table

All columns in red are retrieved through the API and a result of inner joins based on trip, route, and stop IDs from static data. The previous stop delay is created using lag based on trip and stop sequence. Considering this project has been approved by the professor but due to the requirement of the rubric, I can also use https://bustime.mta.info/wiki/Developers/ArchiveData as my dataset which uses New York Transit Data.

Proposed Methodology: The plan is to use hierarchical Bayesian regression to model stop-level delay propagation, leveraging a Student-t distribution as a prior to account for heavy-tailed transit delays (Rodriguez et al. (2022)). Using MCMC inference via RStan, I will estimate posterior delay parameters by incorporating previous stop delays to capture network-wide delay propagation. Optimistically, I also aim to explore spatial dependencies using latitude and longitude data from GTFS to refine the model. Lastly, another important goal is to develop a probabilistic delay prediction model that not only forecasts delay but also provides uncertainty estimates, offering transit agencies and commuters more interpretable and actionable insights. The question that still remains unanswered is if I wish to focus only on certain routes such as 99 or Rapid buses, or try to create a model for all buses which would be challenging since the basis of the model would be based on stops that are inherently different for each route. I plan to perform more EDA and explore this in order to have a concise final report that aligns with the project guidelines.