Two-Dimensional Inhomogeneous Markov Chains for Online Timeliness of Buses

Group 19: Rémi Dion, Alexander Peplowski, Léa Ricard

Université de Montréal

Introduction

- **Quality**: Real-time information provided to users can increase their perception of the quality of a public transport service.
- **Regularity**: For each route of a bus network, several trips are scheduled at regular intervals (i.e. these departure and arrival times are timetabled).
- **Problem**: Disruptions or variability in travel time can ripple down onto future stops and trips.

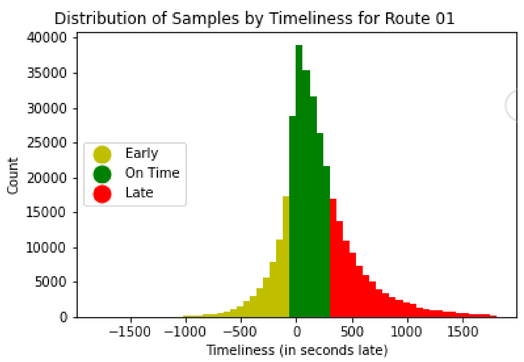
Bus delay prediction problem

Given knowledge on timeliness of past stops, or previous trips, we want to infer next stop or trip timeliness defined by US transit authorities [1] as:

- Early (> 1 min early)
- Late (> 5 min late)
- On Time (otherwise)

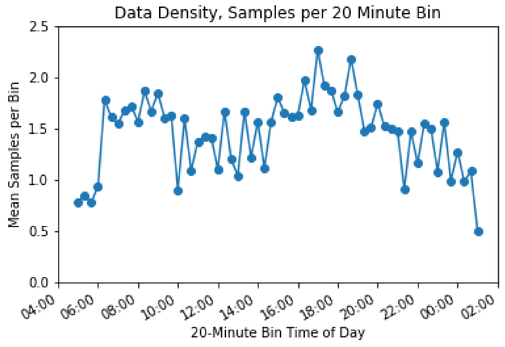
Dataset

Data source: MBTA dataset [2] provides per-stop scheduled and actual arrival times for every bus stop of any route for every day from Aug 2018 to Dec 2019.
Routes: A bus route is defined as a series of bus stops: We use a single bus route, route 01, which has the highest data availability for our experiments.



Challenges:

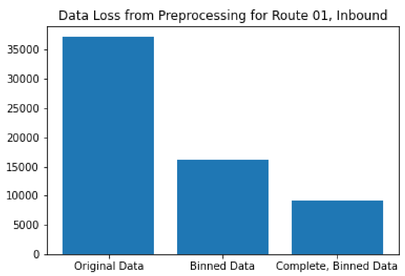
- Scheduled departure time are sometimes modified
- Some trips are not recorded
- Days are not identical (holidays, weekend)



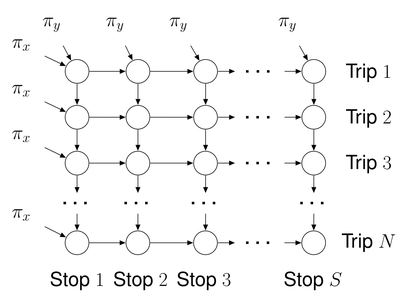
Data Preparation

Our graphical model approach will require a constant number of recorded trips per day with comparable departure times.
In order to uniquely identify a bus stop with a bus ID and a trip start time, some processing reduced the data availability:

- Stops on route: Trips with missing stops are disregarded
- Dimensionality of route start time: Time binning forces route start times to a fixed-dimension uniform distribution



These transformations produced a lattice structure where we can hypothesize dependence on neighboring nodes [3] as shown below in the Markov chain model below.



Models

Baseline model: predict most common class (mode predictor)
Markov Chain Model: horizontal (1D, per stop), vertical (1D, per time bin), lattice (2D, both)
This lattice model is referred to as Two-Dimensional Inhomogenous Finite Markov Chain [4] and is the one we implemented:

$$\begin{aligned} T_i &= p(X_{n,s} = t_i \mid X_{n,s-1} = t_j, X_{n-1,s} = t_k) \\ &= \alpha P_{ij} + (1-\alpha) Q_{ik} \end{aligned} \tag{1}$$

When $X_{n,s-1}$ is not available, we use priors π_x such that P_{ij} is replaced by π_x.
Same goes for Q_{ik} with π_y.

Experimental setup

The data set was divided in a **training set** (40%), a **validation set** (40%) and a **test set** (20%).
P^{prep} is learned on the training set and P^{slid} is learned on recent past trips (using sliding windows). Both sets of parameters are weighted using γ, such as

$$P = \gamma P^{prep} + (1-\gamma) P^{slid}$$

and equivalently for Q, π_y and π_x.

- **Data preparation**: Binning size: 20min & 30min
- **Time-binning (t)**
- **Alpha (α)**: Weight of P over Q
- **Stop Homogeneity** *(stop ID: s)*
  - Horizontally: P vs $P^{(\cdot,s)}$
  - Vertically: (Q, π_y) vs $(Q^{(\cdot,s)}, \pi_y^{(s)})$
- **Time Homogeneity** *(period of day: p)*
  - Horizontally: (P, π_x) vs $(P^{(p,\cdot)}, \pi_x^{(p)})$
  - Vertically: Q vs $Q^{(p,\cdot)}$
- **Sliding window size**: Recent past trips
- **Gamma**: Importance of recent trips vs train data

Results

Dataset	Model	Accuracy of next (%)			
		Trip/day *Training*	Stop *Training*	Trip/day *Test*	Stop *Test*
Route 01 inbound	Baseline			50.2	50.2
	1-d (x-axis)	53.6	**82.0**	55.3	**82.4**
	1-d (y-axis)	**57.0**	57.0	**57.3**	57.3
	1-d (z-axis)	**57.8**	57.8	**57.9**	57.9
	2-d (x and y)	54.5	81.8	55.5	82.2
	2-d (x and z)	54.8	81.7	55.4	82.2

* Next trip accuracy: y-axis; next day accuracy: z-axis

Conclusions

- Timeliness at a stop has a high dependence with the timeliness of the previous stop and a mild dependence with the timeliness of the previous trip.
- For both directions (horizontal and vertical), learning stop-varying transition matrices greatly improves the accuracy over the baseline, while learning in addition time-varying transition matrices mildly improves the accuracy.
- Inhomogeneous 1-dimensional Markov chains outperform inhomogeneous 2-dimensional Markov chains.
- 1-dimensional models are simpler, thus we recommend training an horizontal and a vertical Markov chain separately.

Credits

This work was inspired by the work of [4]. Credits of this team are three-fold:

- Cleaned and prepared the dataset
- Suggested a graphical model reflecting dependence both on the last stop and the last trip
- We implemented 1-dimensional and 2-dimensional inhomogeneous Markov chains

References

[1] Transit service reliability: Analyzing automatic vehicle location (avl) data for on-time performance and to identify conditions leading to service degradation.
Technical report, National Center for Transit Research (NCTR), 2016.

[2] Massachusetts Bay Transportation Authority.
Mbta bus arrival departure times 2020.

[3] Jérémy Roos, Gérald Gavin, and Stéphane Bonnevay.
A dynamic bayesian network approach to forecast short-term urban rail passenger flows with incomplete data.
Transportation Research Procedia, 26:53 – 61, 2017.
Emerging technologies and models for transport and mobility.

[4] Ettore Fornasini.
2d markov chains.
Mathematical Problems in Engineering, 140:101 – 127, 1990.

Université de Montréal