Inferring Passenger Boarding and Alighting Preference for the Marguerite Shuttle Bus System

Adrian Albert

Abstract

We analyze passenger count data from the Marguerite Shuttle system operating on the Stanford University campus. The aim is to infer the time-varying passenger alighting and boarding preference across stops for several important shuttle routes. We look at the problem from the perspective of filling in the missing entries of a transit route O-D matrix. A preliminary analysis indicates that there are certain trips (e.g., the Quarry Welch Rd - Med School Quarry) that are preferred by passengers.

1 Introduction and motivation

Stanford is undergoing a large effort for reducing the its traffic-associated environmental footprint. Taking part in this campus-wide initiative, the P&TS collaborates with the Information Systems Lab in the Electrical Engineering department to investigate possibilities for increasing energy efficiency of the Marguerite Shuttle system[1]. This can be achieved through bus allocation strategies for the various routes that better take into account passenger rideship preference.

Marguerite shuttles are equipped with devices that allow P&TS to accurately assess the number of passengers boarding and alighting at each stop in the Marguerite system (see Section 4). However it is not currently possible to determine how many of the passengers that board at one stop choose to alight at another (given) stop.

An estimate of passenger preference for the various stops in the system based on temporal factors (such as day of week or month of year) or on historical factors (special events, construction sites, etc.) may aid P&TS in e.g., planning the number of busses allocated for particular routes at different times. In addition, an assessment on the number of people riding the bus in between different stops could contribute to estimating and predicting spatial and temporal information of overcrowding on the Marguerite.

In this paper we focus our attention on one of the main lines in the Marguerite system, the *A Line*. This line runs on weekdays and has 36 stops, starting from (and ending at) the Palo Alto Transit Center, where the buses station for up to 10 minutes.

2 Problem formulation

Let us consider the case of one Marguerite shuttle completing one full trip (route), meaning it starts from a certain initial stop, visits other stops on the way, and ends at the same stop it started from. We are interested to infer the distribution across stops of the number of people who boarded at a certain stop and alighted at another (given) stop, for all stops along a given line, given the boarding and alighting passenger counts for each stop along that line (in the present paper, the *A Line*). We may assume that the bus completes an integer number of routes per day, and after each route the bus has no passengers left on.

Mathematically, this is equivalent to finding individual entries of a $m \times m$ square matrix (with m the total number of stops on the respective line) when only the row and column sums are known. The matrix N columns and rows represent the individual stops S_i, with $i \in \{1, 2, ..., m\}$, and N_{ij} = number of people boarding at stop i who alight at stop j. As it is immediately apparent, the row sum for row i is the total numberve of people b_i boarding at stop S_i, whereas the column sum for column j is the total number of people a_j alighting at stop S_j:

$$\sum_j N_{ij} = b_i \text{ and } \sum_i N_{ij} = a_j. \tag{1}$$

This problem is illustrated in Table 2. Above we assumed that no passengers board at the final stop ($b_m = 0$) or alight at the initial stop ($a_1 = 0$). While the a_i's and b_i's are observable, the N_{ij}'s are not and need to be inferred. By definition, $N_{ij} = 0$ for all $i \geq j$, since in this case no people have yet boarded at stop i who can alight at j. This is the well-known transit route *Origin-Destination (O-D)* problem.

[1] Stanford University's free public transportation service

Stop	S_1	S_2	S_3	...	S_{m-1}	S_m	Destination (alights)
S_1	0	$N_{1,2}$	$N_{1,3}$	...	$N_{1,m-1}$	$N_{1,m}$	a_1
S_2	0	0	$N_{2,3}$	...	$N_{2,m-1}$	$N_{2,m}$	a_2
S_3	0	0	0	...	$N_{3,m-1}$	$N_{3,m}$	a_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_{m-1}	0	0	0	...	0	$N_{m-1,m}$	a_{m-1}
S_m	0	0	0	...	0	0	a_m
Origin (boards)	b_1	b_2	b_3	...	b_{m-1}	b_m	

Tab. 1: Transit route Origin-Destination (OD) matrix problem. The alight counts a_i and board counts b_i are observable, while the transition counts N_{ij} ($i \in 1, \vdots m,\ j \in 1, \vdots m$ are not.

3 Approach

The problem of filling an *Origin-Destination matrix* is highly underspecified, whereby there are many unknown entries and only a few observations. For a given set of observed values of row and column sums there are arbitrarily many solutions, and solving for integer entries is NP-hard.

There are many approaches in the literature that attempt to fill in the matrix elements N_{ij} using linear algebra-based approaches, such as *the balancing method* proposed by [Lamond and Stuart, 1981]. There, a seed O-D matrix (with known entries obtained, e.g., from historical data) is transformed by iteratively multiplying each entry in a row with a certain number, such that the row sum matches the observed count. The disadvantage of such linear algebraic methods is that they do not have a natural way to incorporate prior information, and only look at a "snapshots" of the routes independently of eachother.

A dynamic approached is adopted by [Hazelton, 2008], who looks at inference for O-D matrices for the time-varying case. For a sequence of daily counts, they describe the day-to-day evolution of O-D matrices using a model of the day-to-day structural change in the matrix (e.g., weekday vs. week-end). The model parameters are learned using Bayesian inference on the time series data.

In the following we build upon a recent study by [Li, 2009], which we outline below.

3.1 A Markov Decision Model

[Li, 2009] break down the transit route O-D matrix filling problem into two parts: the estimation step and the reconstruction step.

In the **estimation step** we make the important observation that a passenger boarding at S_j can alight at S_i only if he is onboard at stop S_{i-1}. We assume that the transition probability for stop S_i will depend only on the passenger's status at a small number of stops $S_k, S_{k+1}, ..., S_{i-1}$, for some $k \in \{1, 2...i-1\}$. This motivates the use of a Markov chain model (MCM) to express transition probabilities.

For the sake of simplicity, in a first stage we look at a first order MCM. Defining the random variable ξ_i by $\xi_i = 1$, if a passenger is on board at stop i, and $\xi_i = 0$, if the passenger is not on board at stop i, the Markov transition probabilities become:

$$\begin{aligned} Pr\{\xi_i = 0 | \xi_{i-1} = 1\} &= q_i \\ Pr\{\xi_i = 1 | \xi_{i-1} = 1\} &= 1 - q_i, \end{aligned} \tag{2}$$

for $i = 2, \ldots, m-1$. Above, q_i is the probability that a passenger alights at stop i given that they are onboard at stop $i-1$. It is shown in [Li, 2009] that the transition preference probability distribution across stops is given by:

$$p_{i,i+1} = q_{i+1} \text{ and } p_{ij} = q_j \prod_{k=i+1}^{j-1} (1 - q_k)(j = i+2, \ldots, m). \tag{3}$$

In the **reconstruction step** we use the obtained preference probability distribution to calculate the entries of the sought OD matrix:

$$\widehat{N_{ij}} = p_{ij} b_i, (i = 1, \ldots, m-1; j = i+1, \ldots, m), \tag{4}$$

and $\widehat{N_{ij}} = 0$ if $i \geq j$.

3.2 Bayesian inference

In order to infer the Markov parameters q_j ($j = 1, \ldots, m$), [Li, 2009] propose that $q_j \sim \text{beta}(\alpha_j, \beta_j)$. Using this assumption and that $a_j \sim \text{Bin}(\sum_{k=1}^{j-1}(b_k - a_k), q_j)$ (since the transition process is Markov, see [Li, 2009]), one

Line_name	Date_time	Count_type	Count	Stop_Label	Stop_Id	Bus_Id
A Line	07/31/2008 19:54:28	boarding	2	Quarry HooverPv	101	8
A Line	07/31/2008 19:50:58	boarding	1	MedicalC Quarry	74	8
A Line	07/31/2008 19:47:32	alighting	54	PaloAlto TransC	66	8
A Line	07/31/2008 19:47:32	boarding	43	PaloAlto TransC	66	8
⋮						

Tab. 2: Marguerite data for the A Line (excerpt). For example, 54 people alighted at the Palo Alto Transit Center on the 31^{st} of July, 2008, 19:47:32, and 43 people boarded at that stop on bus 8.

arrives at the following estimate on q_j:

$$\hat{q_j} = (\alpha_j + a_j) / \left(\alpha_j + \beta_j + \sum_{k=1}^{j-1} (b_k - a_k) \right), \; j = 2, \ldots, m-1 \tag{5}$$

The Bayesian framework allows for incorporation of prior information, which is effectively the learning process through which the initial non-informative priors ($\alpha_j = \beta_j = 1$) can be updated to reflect the new information after solving the OD estimation problem for a training set of routes. We leave this aspect for a future study, and concentrate here on the simpler approach of calculating the aggregate estimate of the whole set of routes with non-informative priors.

4 Data processing and inspection

4.1 Data retrieval

The P&TS provided access to the Marguerite Shuttle data, which was in the form of several tens of SQL/ODBC tables stored remotely on the servers of a contractor company. A fair amount of interfacing with P&TS and the contractor was necessary in order to obtain the permissions for VPN access. We used commercial database and spreadsheet software (MS Access and Excel) for inspecting the myriad tables available, constructing queries on variables of interest across multiple tables and exporting results into a text format easily accessible to Matlab. An excerpt from a sample file obtained after processing is presented in Table 2.

4.2 Extracting route information

Extracting route information from the P&TS data proved to be particularly challenging, owing to the sheer volume and format complexity of the available information, and to "imperfections" in the data relative to the assumptions of the OD matrix filling problem. For example, counts were not available for certain stops or periods of time, there were multiple counts taken at the same time for certain stops and buses, or routes were recorded to start at stops other than the designated initial stop. Also, the P&TS tables centralized counts from all buses in service (averaging 8-10 at most times for the A Line only), which meant that for mining the data to extract routes we needed to keep track of the progress of every bus along the line.

Our inspection of the data suggests that counts at the initial (and thus end-route) stop are especially sensitive to error sources such as the ones outlined above. For example, at the final stop a bus stations for a longer period of time, during which several readings are taken (both alights and boards), which usually are large compared to the ones at other stops along the line. It may also happen that passengers choose to board a few stops before the end stop, and remain on the bus while it stations at the end stop, continuing their journey thereafter. In particular, effects like these lead the total passenger count deficit for some completed trips,

$$D \equiv \sum_{j=1}^{m} b_j - \sum_{j=1}^{m} a_j, \tag{6}$$

to take values $D \neq 0$, which contradicts a key assumption of the transit route OD problem.

We sequentially parse the pre-processed data to extract route information, storing information in a bus queue as we proceed (see Algorithm 12). Our implementation uses a state machine and a queue for keeping track of the stops visited by each bus, and of the routes completed, and thus is fairly robust to any change in the order of which the count readings are processed. When a route is signaled to have ended with a non-zero passenger deficit, we divide up the value of last count (either *board* or *alight*) between the last stop of the completing route and the first stop of the new route such that the last route had a total value of $D = 0$ (the *balancing* step in Algorithm 12).

We believe that, by far, the effect that plagues most our analysis is the non-zero value of the passenger deficit D for many of the routes. To obtain an intuition of how the deficit varies with time for the interval

Input: count readings (time series) as a matrix of the form in Table 2
Output: completed routes (time series) in the form given in Table 2

```
Initialize bus queue
foreach count reading do
    if current stop ≠ last stop in route then
        queue = push(queue, current stop info)
    end
    else
        completed route = pop(queue)
        balance last count between completed route and new route to satisfy Eq. (6)
        Set route to transit route OD format as in Table 2

            • assign time stamp to each route as average time of visits to stops

        Store completed route in routes array
    end
end
Sort routes chronologically according to time stamp
```

Algorithm 1: Algorithm for extracting route information from pre-processed counts data.

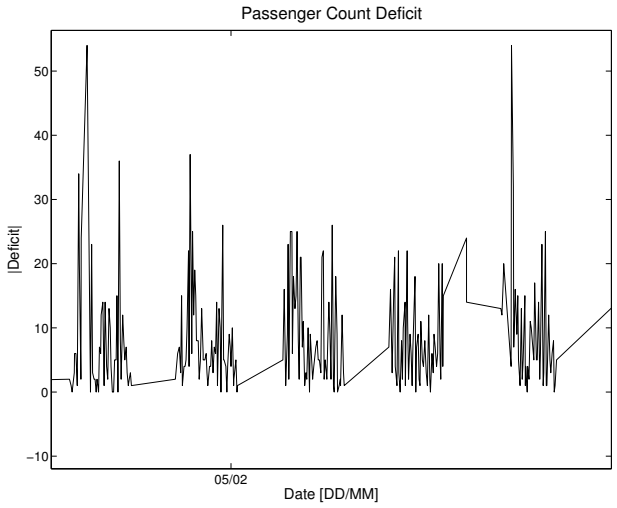

Fig. 1: Passenger count deficit D for the week of February 5^{th}, 2008. The deficit has two peaks around 8:00AM and 6:00PM, and is small around noon.

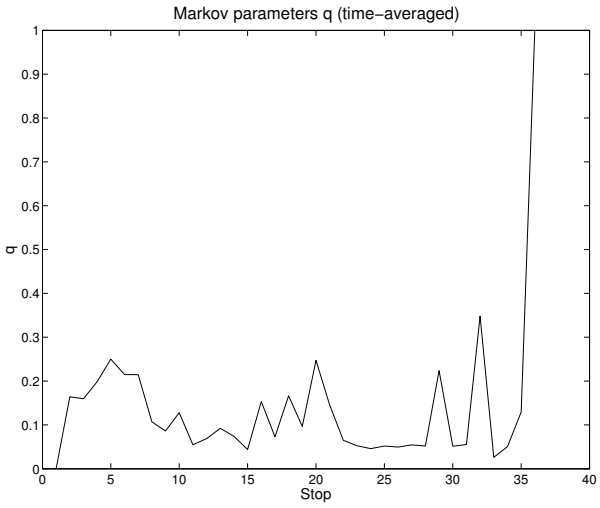

Fig. 2: Markov parameters q accross stops in the A Line (time-averaged). By construction, at the second-to-last stop $q_{36} = 1$ and at the initial stop $q_1 = 0$.

considered we created plots such as the one in Figure 1. Such a plot can be useful as a crude measure of the confidence in the estimate on the preference probability matrix $\mathbf{P}$ when considering variations on time scales such as hour-of-day.

5 Preliminary analysis

Implementing the approach presented in Section 3 is fairly straightforward: for each completed route obtained as described in Section 4 we estimate the Markov parameters q using Equation 5, which we use to infer the preference probability matrix $\mathbf{P}$. The entries of the OD matrix $\mathbf{N}$ are calculated using Equation 4. In a first analysis we use non-informative priors, and take the time-averages of quantities of interest for the considered interval (January-April 2008). During this interval there were 3236 complete trips on the A Line.

We plot the time-averaged Markov model parameters q_j obtained with non-informative priors $\alpha_j = \beta_j = 1$ in Figure 2. As the model requires (see Section 3), at the second-to-last stop $q_{36} = 1$ and at the initial stop $q_1 = 0$. On average, people who board around the 5-th stop (*Quarry Welch Rd*) will tend to remain on the bus for more than one stop with a probability of 20%, while people who board between stops 10 and 15 (*SerraM MainQuad* through *Serra S CampusW*) will only stay on the bus for after the next stop with a 10% probability.

In Figure 3 we illustrate the preference probability matrix estimate $\mathbf{P}$, computed with non-informative priors $\alpha_j = \beta_j = 1$, and time-averaged for January-April 2008. For this period, people who board at stop 5-8 (*Quarry Welch Rd* through *ViaOrtega Serra*) have a 35% preference probability of alighting at the next stop or staying

on the bus until stop 20 (*Olmsted Wellsly*). People boarding from stop 6 through 27 (*Quarry MedCent - Serra BurnhamPa*) have a similar preference probability of 35% of alighting at stop 29 (*Serra Mall Oval*) or stop 32 (*MedSchoo Quarry*). Also, people seem to have higher preference ($\geq 50\%$) to alight at the latter stop if they boarded at the former.

6 Conclusions and Future Work

Motivated by the problem of increasing energy efficiency for the Marguerite Shuttle by better allocation of buses, we analyzed passenger count data for the A Line from January through March 2008 to determine passenger preference probability of boarding and alighting accross stops. To solve the associated transit route Origin-Destination (OD) matrix filling problem, we used an approach based on a first-order Markov model an on Bayesian analysis. The bulk of our effort until now went into preprocessing the data and parsing it to extract complete OD trips. Interesting insights offered by a preliminary analysis using non-informative priors include, e.g., that passengers have high preference of alighting at the Medical School Quarry, if they boarded at the Quarry Welch Rd.

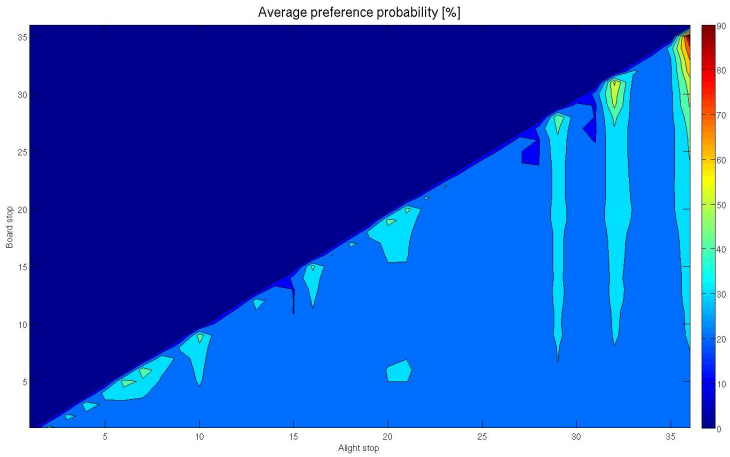

Fig. 3: Preference probability matrix with non-informative priors (time-averaged for January-April 2008).

We plan to continue working on this problem to extend the approach outlined in Section 3 as follows:

- Incorporate Bayesian learning: treat part of the data as "training set", then use learned parameters as priors for subsequent analysis on remaining data;
- Replace first order Markov model with a higher order model, as first order MCM tends to give preference to very short trips (one stop);
- Investigate temporal distribution of passenger preference probability by dividing up the data in time intervals of interest and analyzing each interval separately. For this we plan to incorporate the time-varying model of [Hazelton, 2008] with the (seemingly static) Markov chain model of passenger boarding/alighting probability proposed pursued in this paper.

Acknowledgements. The author would like to thank Prof. Balaji Prabhakar, Deepak Marugu, and Seewong Ou from the Electrical Engineering Department, and Catie Chang from the Computer Science Department at Stanford for fruitful discussions and suggestions. We are grateful to Ramses Madou and Angus Davol from the Stanford Parking & Transportation Services for providing access to the Marguerite Shuttle data and helpful hints about processing it.

References

[Hazelton, 2008] Hazelton, M. L. (2008). Statistical inference for time varying origin-destination matrices. *Transportation Research Part B: Methodological*, 42(6):542 – 552.

[Lamond and Stuart, 1981] Lamond, B. and Stuart, N. F. (1981). Bregman's balancing method. *Transportation Research B*, 15(4):239–248.

[Li, 2009] Li, B. (2009). Markov models for bayesian analysis about transit route origin-destination matrices. *Transportation Research Part B: Methodological*, 43(3):301–310.