USDOT Region V Regional University Transportation Center Final Report

NEXTRANS Project No. 168OSUY2.2

ESTIMATING TRANSIT ROUTE OD FLOWS IN THE PRESENCE OF MULTIPLE LATENT FLOW PATTERNS

By

Rabi G. Mishalani, Principal Investigator
Professor of Civil, Environmental, and Geodetic Engineering
The Ohio State University
mishalani@osu.edu

Mark R. McCord, co-Principal Investigator
Professor of Civil, Environmental, and Geodetic Engineering
The Ohio State University
mccord.2@osu.edu

and

Andrew Landgraf, Investigator
Research Scientist, Civil, Environmental, and Geodetic Engineering
The Ohio State University
landgraf.5@osu.edu

Report Submission Date: Feb. 13, 2017

ACKNOWLEDGMENTS AND DISCLAIMER

Partial funding for this research was provided by the NEXTRANS Center, Purdue University under Grant No. DTRT12-G-UTC05 of the U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology (OST-R), University Transportation Centers Program. Additional funding was provided by The Ohio State University (OSU) including the College of Engineering and Transportation and Traffic Management.

The authors are grateful to OSU's Transportation and Traffic Management for its support of the OSU Campus Transit Lab (CTL) and Sarah Blouch, Chris Kovitya, and Beth Kelley-Snoke for their efforts in supporting the development and continued maintenance of the CTL, which resulted in the availability of the data used in this study.

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION AND MOTIVATION

Route-level Origin-Destination (OD) flow matrices provide useful information for ridership forecasting, service planning (e.g., extending routes, splitting or combining routes, and introducing new routes), and control strategies development (e.g., short turning, expressing, and holding). Since directly observing OD flows via on-board surveys is time consuming and costly, many methodologies have been proposed to estimate route-level OD matrices from boarding and alighting counts (Ben-Akiva et al., 1985; Ben-Akiva, 1987; Kikuchi and Perincherry, 1992; Li and Cassidy, 2007; Li, 2009; Hazelton, 2010; Ji et al., 2014; Ji et al., 2015; Mishalani et al., 2017). Passenger boarding and alighting counts are relatively easier and less costly to collect than OD flow data. Moreover, many transit agencies are now collecting large quantities of boarding and alighting counts on a routine basis via Automatic Passenger Count (APC) technologies, thus, providing the opportunity to estimate up-to-date OD flows on an ongoing basis.

While methods for estimating route-level OD matrices from boarding and alighting counts are based on different assumptions and employ different estimation approaches, they all assume, explicitly or implicitly, that the passenger flow patterns reflect a single underlying probability OD flow matrix. However, across the day and even within a time-of-day (e.g., morning, mid-day, and evening periods), different travelers are engaged in a variety of trip purposes, such as work, personal business, education, and shopping. Such variations within a time-of-day period are becoming even more prevalent with the advent of more complex life and work needs and duties across varying household structures. Therefore, passengers' travel purposes, which are associated with different sets of origin-destination pairs and departure times within a time-of-day period would be more realistically represented by multiple underlying probability OD flows reflecting the variety of travel purposes. That is, it is conceivable that bus trips within a period could carry travelers that collectively exhibit different underlying probability OD flow patterns. It is possible to extend the formulations and methods proposed by Hazelton (2010) and Ji et al. (2015) to take into account the presence of multiple underlying OD flow matrices. However, the computational costs associated with doing so would render them infeasible. In contrast, the computationally efficient variational Bayes method (Mishalani et al., 2017) developed in a separate study offers the advantage of allowing for capturing the presence of multiple underlying OD flow matrices in a computationally feasible manner.

In this study, this computationally efficient variational Bayes method is extended to capture the presence of multiple underlying OD flow matrices. In addition, a data-inspired simulation based evaluation is conducted to assess the value of recognizing multiple underlying probability OD flow matrices. Moreover, a preliminary empirical study is conducted to investigate the potential presence of multiple underlying probability OD flow matrices on an operational bus route.

2. METHODOLOGY

The notation used in this report is the following:

y_l = volume OD flow matrix for the lth bus trip,
l = index representing bus trips,
L = number of bus trips,

n_l = total number of passengers on trip l,
α = probability OD flow matrix,
x_l = boarding and alighting counts for trip l,
q = approximate posterior distribution of the true OD flows,
k = index representing the latent underlying probability OD flow matrices for a given time-of-day period,
K = number of latent underlying probability OD flow matrices for a given time-of-day period,
π = mixing probabilities of the K latent underlying probability OD flow matrices, and
z_l = cluster assignment taking the value of the k^{th} latent matrix assigned to the l^{th} bus trip.

For the variational Bayes method developed by Mishalani et al. (2017), it is assumed that each bus trip volume OD flow matrix is the result of a multinomial trial, with known total number of passengers, n_l, and unknown probability OD flow matrix, α. These assumptions are similar to those adopted by Hazelton (2010) and Ji at al. (2015). Let y_l represent the volume OD flow matrix for the l^{th} bus trip. The OD flow volumes are not directly observed, but rather row- and column-wise summaries of them are provided in the boarding and alighting count data, which for a given trip l are represented by x_l.

In light of the presence of multiple latent probability OD flow matrices reflecting varied trip purposes across travelers within a time-of-day period as hypothesized and elaborated on in the introduction section, a model based clustering (Fraley and Raftery, 2002) formulation for estimating OD flows is developed by extending the variational Bayes method developed for the single OD flow matrix case. In this extended formulation, it is assumed that there are a specified number, K, of underlying probability OD flow matrices for a given time-of-day period and that each bus trip volume OD flow matrix within that period is a realization of one of these K probability OD flow matrices. It is further assumed that there are mixing probabilities associated with each of the clusters that determine how often each of the OD flow patterns occurs on average within a time-of-day period. Adding the clustering component to the model allows for the possibility that all bus trips within a time-of-day period may not be similar. Imposing a prior distribution on the K probability OD matrices and the mixing probabilities of the clusters, the goal is to estimate the posterior distribution of the time-of-day period level probability OD flow matrix. This matrix is defined as the average of the underlying K latent OD matrices, weighted by their respective mixing probabilities.

A graphical representation of the model is shown in Figure 1. The round cornered rectangular "plate" on the right-hand-side of the figure represents K independent latent probability OD flow matrices and the "plate" on the left-hand-side of the figure represents L independent realizations of bus trips. The circles represent the variables. The shaded circles represent observations, and the unshaded circles represent unobserved variables. The circles containing x_l and n_l are shaded because they represent the observed boarding and alighting data and the total number of passengers, respectively. The cluster assignments to trips, denoted by z_l taking the value of the k^{th} latent probability OD flow matrix assigned to the l^{th} bus trip, are assumed to be sampled from the mixing probabilities, π. Further, the unobserved bus trip OD volume matrices, y_l, each with total number of passengers, n_l, are generated from one of K possible probability OD flow matrices, α_k. Finally, the boarding and alighting data, x_l, are a direct result of the bus trip volume OD flow matrices by means of the row totals and column totals.

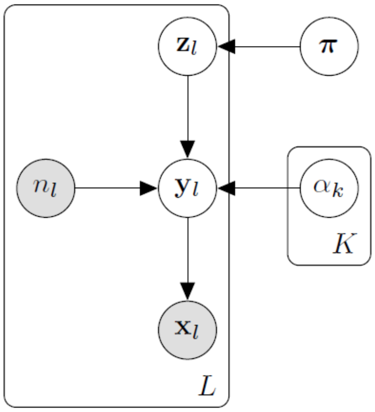

FIGURE 1: Graphical representation of trip clustering framework

There are several approaches that can be used to estimate probability OD flow matrices and their mixing probabilities. However, in light of the computational infeasibility of extending the methods developed by Hazelton (2010) and Ji et al. (2015), the variational Bayes method developed by Mishalani et al. (2017) is extended to approximate the posterior distribution in a computationally feasible manner. Variational methods (Jordan et al., 1999) approximate the posterior distribution of the true OD flows with another (variational) distribution, *q*, by assuming *q* belongs to a simpler class of distributions than the true posterior. If no restriction is put on *q*, the closest distribution will be the posterior itself. Naturally, this is not helpful and, therefore, simpler classes are used to find a closed form solution of the posterior distributions of the entities to be estimated. Ideally, the goal is to assume a realistic approximation that is not so complex as to make the computation infeasible. Once the class of distributions is specified, *q* is solved for such that the Kullback-Leibler (KL) divergence of *q* from the true posterior is minimized (Jordan et al., 1999). To estimate the probability OD flow matrices, it is assumed that the variational distribution of the OD trip volumes is independent of the variational distribution of the probability OD flow matrices, i.e., that *q* factorizes. Under this assumption, *q* can be estimated efficiently by iteratively updating the parameters of the variational distribution in a fashion reminiscent to the Expectation Maximization (EM) algorithm (Parisi, 1988). Since it is infeasible to determine the expectation of the bus trip volume OD flow matrices conditional on the boarding and alighting counts, the heuristic approximation developed by Ji (2011) and Ji et al. (2015) is used.

3. SIMULATION-BASED EVALUATION

To assess the value of recognizing multiple underlying probability OD flow matrices through the application of the proposed method in the presence of such matrices, APC data are simulated from an assumed known true set of underlying probability matrices, the variational method is applied to these APC data for different assumed number of clusters including one, and the resulting estimated time-of-day period probability OD flow estimates are compared to the true probability OD flows for each of the assumed number of clusters. Using simulated data has the advantage of allowing the true number of clusters, underlying period-level probability OD flow matrices, and mixing probabilities to be known.

To increase the realism of the simulation, all the variables needed for the simulation (including the probability OD flow matrices) were estimated from an actual bus route's APC data from The Ohio State University's Campus Transit Lab (CTL) (Campus Transit Lab, 2016). The CTL is a living lab based on Campus Area Bus Service (CABS), a transit service that serves approximately five million passengers annually on six routes. Specifically, APC data from 780 bus trips of OSU's Campus Loop South (CLS) route for the 8 to 10 AM period are used. For the purpose of the simulation, the underlying clusters of probability OD flow matrices are estimated from the APC data by the developed variational Bayes methodology for an assumed number of clusters. The estimated matrices are then assumed to represent the underlying true matrices in the simulation where realizations of error free APC data are generated.

Results of a representative example of the simulation experiment are presented here where the truth is assumed to consist of two clusters. The accuracies of the period-level probability OD flow estimated matrices (i.e., after aggregating across clusters) are plotted in Figure 2. The six subplots correspond to six sets of simulated data, where each subplot corresponds to different numbers of bus trips with APC data, where the data are all generated from the same two assumed underlying true probability OD flow matrices. For a given set of APC data, the period-level probability OD flow matrix is estimated for different numbers of assumed clusters, ranging from one up to five. The y-axis displays the value of the squared Hellinger distance (HD^2) metric, which is commonly used to compare two probability distributions, in this case comparing the true and estimated period-level OD flow probability matrices for each of the initializations. Specifically, the HD^2 metric is the sum of the squared difference between the square root of the estimated probabilities and the square root of the true probabilities (Yang et al., 2000). Lower values of HD^2 indicate more similar probability matrices and, therefore, that the estimated matrix is closer to the true overall time-of-day period-level probability OD flow matrix. As with almost all clustering problems, initialization plays an important role. For each simulated data set and assumed number of clusters, 20 random initializations reflecting different prior (seed) values for the K probability OD flow matrices and the mixing probabilities are used, assuming that each is uniformly distributed. Boxplots of the HD^2 values over the 20 different initializations are shown in Figure 2.

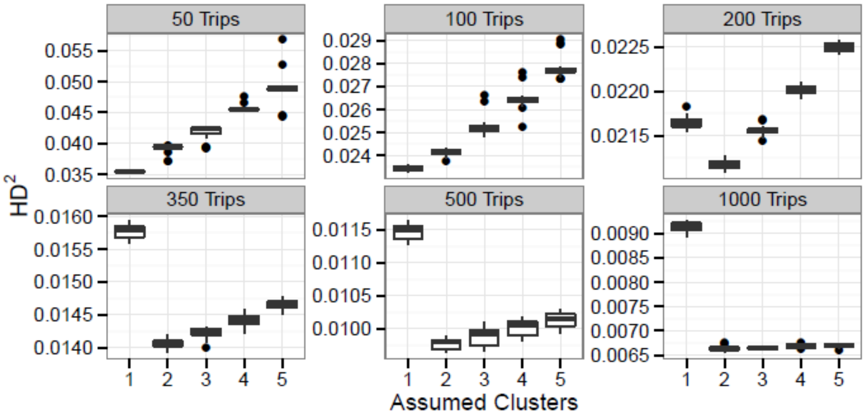

FIGURE 2: HD^2 of estimates with varying APC data sample sizes and assumed number of clusters in the presence of two underlying probability OD flow matrices for the CLS-AM route-period

As expected, the results indicate that as the number of bus trips increases, using the correct number of clusters – two in this case – gives the most accurate results. Also, as the number of bus trips with APC data increases, the improvement in accuracy of going from one to two clusters far outweighs the degradation in accuracy for having more than two clusters. The latter result indicates that assuming more clusters than the actual number of underlying latent probability OD flow matrices leads to little degradation in the quality of the estimates when the number of bus trips with APC data is sufficiently large. In summary, based on the simulation results, the proposed method is able to improve estimation accuracy when multiple underlying probability OD flow matrices are assumed to generate different passenger flow patterns within the same time-of-day period.

4. PRELIMINARY EMPIRICAL VALIDATION

The simulation study indicates the ability of the proposed variational Bayes method to improve estimation performance when multiple underlying OD flow matrices are assumed to exist. A preliminary empirical study has also been conducted to investigate the potential presence of multiple underlying probability OD flow matrices on an operational bus route and the nature of the flow estimates determined by proposed variational Bayes method. In this empirical investigation, APC data collected from 780 bus trips on the CLS route for the AM period used in the simulation analysis presented in section 3 are used. The empirical data may not follow the same assumptions used when generating simulated data and may also be subject to measurement errors. The varational Bayes method is applied assuming the presence of three clusters. The intent is to investigate the estimates of the underlying three matrices against a priori knowledge of the travel patterns on this route to validate whether in fact the three estimated probability OD flow matrices reflect realistic and plausible flow patterns.

When estimating the three clusters, the cells with largest probabilities occurring in the first cluster (which is estimated to be the probability matrix for 42% of trips) are predominantly associated with passengers boarding at the commuter park-and-ride facility and going to the medical campus. The cells with largest probabilities occurring in the second cluster (32% of trips) are predominantly associated with intra-core-academic-campus travel and travel to stops located in the core academic campus. Finally, the cells with largest probabilities occurring in the third cluster (26% of trips) are predominantly associated with passengers traveling from the core academic campus to either the agriculture campus or the commuter park-and-ride facility mentioned previously. Each of the three probability OD flow matrices seems to have a somewhat homogeneous trip purpose.

To further validate the association of these three OD matrices with trip purpose, the estimated cluster assignments for each of the trips, z_l, are analyzed. To do so, the probability of bus trips being assigned to each of the clusters was determined as a function of the departure time of each bus trip from the terminal located at the commuter park-and-ride facility, within the AM period. It is found that the first cluster occurs most often early in the morning, a finding that is consistent with the dominant flows from the park and ride facility. The third cluster peaks at about 8:30 and 9:40. These times correspond to bus departure times from the terminal that would travel through the core academic campus when many classes end at 8:55 and 10:05. Many students in these morning classes who use this route would travel to off-campus jobs or return home and board these buses to access their cars in the commuter park-and-ride facility or in the additional parking lots near the agricultural campus. Finally, the second cluster peaks at around

9:00 and 10:00. Such a temporal pattern could reflect the decreased demand of the other patterns in the other clusters (probability matrices reflect the demand of a flow in one cell relative to demand in all other cells) and increased intra-core-academic-campus travel and demand for travel to the core academic campus.

In summary, the correspondence of the clusters with the timing of different trip purposes supports the reasonableness of the existence of multiple underlying probability OD flow matrices within a time-of-day period. The ability of the developed varitional Bayes method to identify these interpretable matrices supports the validity of the method.

5. FUTURE RESEARCH

Given the promise of the developed methodology, a more comprehensive evaluation would be a valuable effort. That is, different simulations for additional route-periods, from a variety of ground-truths, and with different true numbers of clusters would further validate the use of clustering and the algorithm's effectiveness and set the stage for adopting the developed methodology in future transit passenger OD flow studies for research or practice purposes.

While the preliminary empirical validation results are promising, further in-depth investigation of the results is warranted. Also, this investigation only considered one route and one time-of-day period. Therefore, further research is needed to confirm the presence of multiple underlying probability OD flow matrices and verify the usefulness of recognizing their presence using actual APC data in a more comprehensive manner. Since it is difficult to collect OD flow data for a large number of bus trips, one difficulty of such an assessment is that there is often no known ground-truth probability OD flow matrix to compare the estimates to. The OSU's Campus Transit Lab from which data for the simulation evaluation and empirical validation of this study are used offers the opportunity to amass high-fidelity onboard OD flow survey that could be used for this purpose.

The proposed clustering methodology has limited practical use if there is no way to determine the number of clusters. Each bus route has its own characteristics and it will be difficult to use subject matter expertise to select the proper number of clusters. As was evident in the simulation, using too many clusters can be detrimental in some of the scenarios, but made little difference in others, depending on the number of bus trips with available APC data. The same is true for using too few clusters. Therefore, it is important to derive a model selection criterion that indicates the number of clusters that are best for any APC dataset for a bus route.

6. REFERENCES

1. Ben-Akiva, M., 1987. Methods to combine different data sources and estimate origin-destination matrices, the 10th International Symposium on Transportation and Traffic Theory. Elsevier, New York, Cambridge, MA, 459-481.

2. Ben-Akiva, M., Macke, P., Hsu, P., 1985. Alternative methods to estimate route level trip tables and expand on-board surveys. Transportation Research Record, 1037, 1-11.

3. Campus Transit Lab (CTL), 2016. https://transitlab.web.engadmin.ohio-state.edu/campus-transit-lab.

4. Fraley, C. and Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97, 611 631.

5. Hazelton, M.L., 2010. Statistical inference for transit system origin-destination matrices. Technometrics, 52(2), 221-230.

6. Ji, Y., 2011. Distribution-based approach to take advantage of automatic passenger counter data in estimating period route-level transit passenger origin-destination flows: Methodology development, numerical analyses and empirical investigations. Ph.D. Dissertation, Department of Civil Engineering, The Ohio State University, Columbus, OH.

7. Ji, Y., Mishalani, R.G, McCord, M.R., 2014. Estimating transit route OD flow matrices from APC data on multiple bus trips using the IPF method with an iteratively improved base: Method and empirical evaluation. Journal of Transportation Engineering, 140(5), 040140081-040140088.

8. Ji, Y., Mishalani, R.G., McCord, M.R., 2015. Transit passenger origin-destination flow estimation: Efficiently combining onboard survey and large automatic passenger count datasets. Transportation Research Part C: Emerging Technologies, 58, 178-192.

9. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K., 1999. An introduction to variational methods for graphical models. Machine Learning, 37(2), 183-233.

10. Kikuchi, S., Perincherry, V., 1992. Model to estimate passenger origin-destination pattern on a rail transit line. Transportation Research Record, 1349, 54-61.

11. Li, B.B., 2009. Markov models for Bayesian analysis about transit route origin destination matrices. Transportation Research Part B: Methodological, 43(3), 301-310.

12. Li, Y.W., Cassidy, M.J., 2007. A generalized and efficient algorithm for estimating transit route ODs from passenger counts. Transportation Research Part B: Methodological, 41(1), 114-125.

13. Mishalani, R.G., McCord, M.R., Kumar, P.K., Landgraf, A.J., 2017. Variational Bayes Method for Estimating Transit Route OD Flows Using APC Data. Final Report, UTC Region 5, NEXTRANS, US DOT, Grant No. DTRT07-G-005.

14. Parisi, G., 1988. Statistical field theory. Addison-Wesley, Reading, Massachusetts.

15. Yang, G.L., Le, C., Lucien, M., 2000. Asymptotics in statistics: Some basic concepts 2nd ed. Springer, Berlin.

CONTACTS

For more information:

PI Name: Rabi Mishalani
University: The Ohio State University
Address: 2070 Neil Ave, Rm 483E, Columbus, OH 43210
Phone Number: 614-292-5949
Fax Number: 614-292-8730
Email Address: mishalani@osu.edu
Web Address: https://ceg.osu.edu/people/mishalani.1

NEXTRANS Center
Purdue University - Discovery Park
3000 Kent Ave.
West Lafayette, IN 47906

nextrans@purdue.edu
(765) 496-9724

www.purdue.edu/dp/nextrans