USDOT Region V Regional University Transportation Center Final Report

NEXTRANS Project No. 125OSUY2.1

VARIATIONAL BAYES METHOD FOR ESTIMATING TRANSIT ROUTE OD FLOWS USING APC DATA

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

Prem Goel, Contributing Investigator
Emeritus Professor of Statistics
The Ohio State University
goel.1@osu.edu

Report Submission Date: Jan. 31, 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 efforts of former Graduate Research Assistants Andrew Landgraf are greatly appreciated.

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). 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.

Transit agencies have also been adopting Automatic Fare Collection (AFC) technologies. Transit passenger OD flows could be derived from AFC data. However, many AFC systems, notably those for bus transit, are access-based (i.e., swipe-on or tap on) only and, thus, only record the stops where passengers board and do not record the stops where passenger alight. As a result, assumptions, some of which are difficult to verify, need to be made for inferring individual passenger OD flows (Chan, 2007; Wang el al., 2011) from such AFC data.

The focus of this study is on the use of large quantities of APC data to estimate OD flows for transit bus routes. Since most OD flow estimation methodologies based on boarding and alighting counts were developed before the prevalence of APC technologies, the value of large quantities of APC boarding and alighting data are not effectively utilized in previously developed estimation methodologies (Ji et al., 2014; Ji et al., 2015). More specifically, many OD estimation methodologies – such as the well-known and extensively used Iterative Proportional Fitting (IPF) procedure (Ben-Akiva et al., 1985; Bacharach, 1970) in its most commonly implemented form – predominantly rely on aggregated boarding and alighting counts to determine a time-of-day period OD flow matrix, where the aggregation is performed by stop across bus trips.

In contrast to methodologies that rely on aggregated boarding and alighting counts, Hazelton (2010), Ji (2011), and Ji et al. (2015) proposed similar statistical formulations but different solution methods to estimate transit OD flows based on sets of disaggregate trip boarding and alighting counts. Specifically, Hazelton (2010) used Markov chain Monte Carlo (MCMC) (Robert and Casella, 2005) methods to generate samples from the posterior distribution. This approach can be very time consuming for realistically long transit bus routes and large boarding and alighting datasets. Ji (2011) and Ji et al. (2015) considered the expectation maximization (EM) algorithm (Dempster, 1977) to find the mode of the posterior distribution as an estimate and developed a heuristic approximation of the expectation of the trip volume matrices, given the probability OD flow matrix and boarding and alighting count data, to circumvent the need for the computationally prohibitive enumeration of all OD flow matrices that satisfy the boarding and alighting count data. Ji (2011) and Ji et al. (2015) showed that the resulting algorithm – referred to as the Heuristic Expectation Maximization (HEM) method – is able to efficiently estimate the probability OD flow matrix with a high degree of accuracy, especially with APC data on many bus trips. This method, however, only estimates the posterior mode of the OD flow estimates and not their full distribution. While it is possible to determine a distribution of the flow estimates determined by the HEM method via bootstrapping (Efron and Tibshirani, 1994), doing so is distinctly different from estimating the posterior distribution of the true flows – for example, as estimated by the MCMC method – and introduces an additional computational burden. Variational Bayes methods (Jordan et al., 1999), which approximate the posterior with the closest distribution from a simpler class of distributions leading to efficient estimation algorithms, allows for estimating the distribution of the true flows and offers the opportunity to do so in a computationally efficient manner.

In this study, a computationally efficient variational Bayes method that approximates the posterior distribution of the probability OD flow matrix is developed. In addition, a data-inspired simulation based evaluation is conducted to show that the estimates obtained are as accurate or more accurate than those obtained from competing methods when a single underlying OD flow matrix is assumed.

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, and
q = approximate posterior distribution of the true OD flows.

A model similar to those used by Hazelton (2010) and Ji at al. (2015) is assumed for the data-generating process of boarding and alighting data and is illustrated in Figure 1. 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. 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.

Many strategies can be used to estimate the posterior distribution of probability OD flow matrix α. As argued in the introduction, variational methods (Jordan et al., 1999) are used in this study to approximate the posterior distribution of the true OD flows in a computationally feasible and efficient manner. Variational methods approximate the posterior distribution 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 matrix, it is assumed that the variational distribution of the OD trip volumes is independent of the variational distribution of the probability OD flow matrix, 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 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.

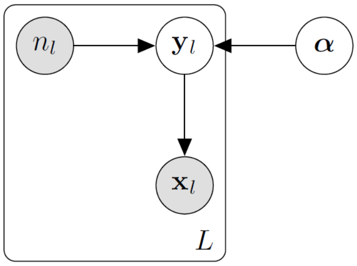

FIGURE 1: Graphical representation of the assumed data-generating process

3. SIMULATION-BASED EVALUATION

The quality of the estimates obtained when using the proposed method are compared to the quality of the estimates obtained by competing methods. While large quantities of empirical boarding and alighting data can be obtained from APC technologies, the true underlying probability OD flows are not available. Reliably determining ground-truth OD flows requires onboard surveys, which are time-consuming, labor-intensive, and expensive to carry out for large studies. It would also be informative to compare the performance of the various methods in a controlled experiment, where different parameters can be adjusted. Therefore, a data-inspired simulation experiment is performed.

In the simulation, bus trip volume OD flow matrices are generated assuming a multinomial distribution based on a probability OD flow matrix and a bus trip volume distribution. To improve the realism of the simulation, the underlying probability OD flow matrix used to generate OD flow matrices is a matrix that is estimated using the method presented here from large quantities of boarding and alighting data obtained from APC technologies on operational bus routes. The total bus trip volumes are also drawn from the same route's ridership data.

To assess performance under a diverse set of scenarios, data from two different time-of-day periods for a route are used to generate underlying probability and volume OD flow values for the simulation. Specifically, boarding and alighting data from The Ohio State University's Campus Transit Lab (CTL, 2016) are used. The CTL is a living lab based on Campus Area Bus Service (CABS) (Campus Transit Lab, 2016), a transit service that serves approximately five million passengers annually on six routes. APC data are taken from the Campus Loop South (CLS) route during the AM (8 AM to 10 AM) and PM (3 PM to 5 PM) peak periods. CLS is over 8 km long and has 141 feasible OD flow pairs. The total number of bus trips with APC data used to estimate the underlying probability OD flow matrices and volume distributions for the CLS-AM and CLS-PM route-periods are 780 and 638, respectively.

For each of the two route-period combinations considered, the generated number of bus trips varies from 25 to 1,000. For each bus trip, an OD volume matrix is generated, and the boarding and alighting data which correspond to the row and column sums of the generated trip volume OD flow matrices, are assumed to be observed without error. For a given set of boarding and alighting data, the probability OD flow matrix is estimated using the variational method and four other methods: the iterative proportional fitting (IPF) procedure, which serves as a representative of the state-of-the-practice, and three other methods that make use of the distribution of the APC data, namely, IPF with iteratively updated base (IPF-IB) (Ji et al., 2014), heuristic expectation maximization (HEM) (Ji, 2011; Ji et al., 2015), and Markov Chain Monte Carlo (MCMC) (Hazelton, 2010). Both the posterior mean and posterior mode of the variational method are considered. Since the mean and mode estimates are found to be very similar to each other except for the cases of very few trips with APC data where the variational mode is more accurate, only the variational mode results are discussed further. To maintain consistency across methods, each method is initialized by the same prior matrix (also known as the base or seed) where the prior probability flows of all feasible OD pairs are assumed to have equal probabilities, and the prior probability flows of all infeasible cells are set to 0.

For each method, the resulting estimated probability OD flow matrix is compared to the underlying probability OD flow matrix used to generate the bus trip volume matrices and APC data in the simulation. The distance between the true and estimated probability OD flow matrices is measured by the squared Hellinger distance (HD^2) metric, which is commonly used to compare two probability distributions. The HD^2 measure 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. (Other measures were computed as well, including the Chi Squared distance, Kullback Leibler divergence, mean squared error, and mean absolute error. The results were qualitatively similar for all metrics.) Finally, to assess the variability of the results, for each set of underlying probability OD flow matrix and number of bus trips considered, 20 sets of volume OD flow realizations and corresponding boarding and alighting counts are generated across the bus trips.

Figure 2 shows the average HD^2 value over the 20 simulation trials for each of the methods, periods, and number of bus trips. Each of the two panels displays the results for the two time-of-day periods. Within each panel, there is a generally decreasing pattern of HD^2 for all methods as the number of trips increases. IPF-IB, HEM, and the variational HD^2 values all seem to decrease towards 0 as the number of bus trips increases. Similar to the results of Ji et al. (2015), the accuracy of the IPF estimates does not improve much with increasing number of bus trips. The IPF method can be more accurate than the other methods when there are very few trips but less accurate than the other methods when collecting boarding and alighting data from more than about 200 bus trips, a fairly reasonable amount for a city-wide transit system equipped with an APC technology.

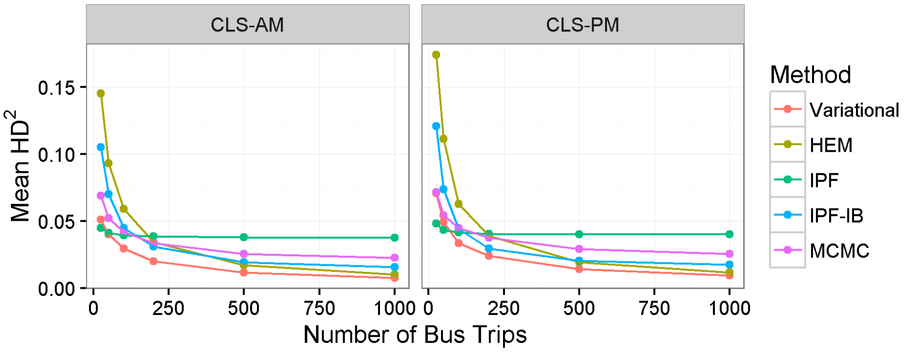

FIGURE 2: Mean accuracy comparisons for two route-periods

It is not easy to observe the differences between the HD^2 values for the estimates obtained by the IPF-IB, HEM, and the variational methods as plotted in Figure 2. Therefore, the differences of the HD^2 values of the variational mode estimates and the estimates of each of the other methods are plotted in Figure 3. In this figure, the distribution of the differences over the 20 simulated APC datasets are also shown in the form of box-plots. A distribution above 0 indicates that the variational mode is more accurate than that method, and vice versa. From Figure 3, it is apparent that the variational method is more accurate than all the other methods in nearly all of the 400 simulations where APC data on at least 100 bus trips are available. In addition to determining better estimates, the variational method is also very feasible. For example, using an Intel(R) Xenon(R) 2.27GHz processor, estimates were determined in 11.1 seconds on average using APC data on 1,000 bus trips.

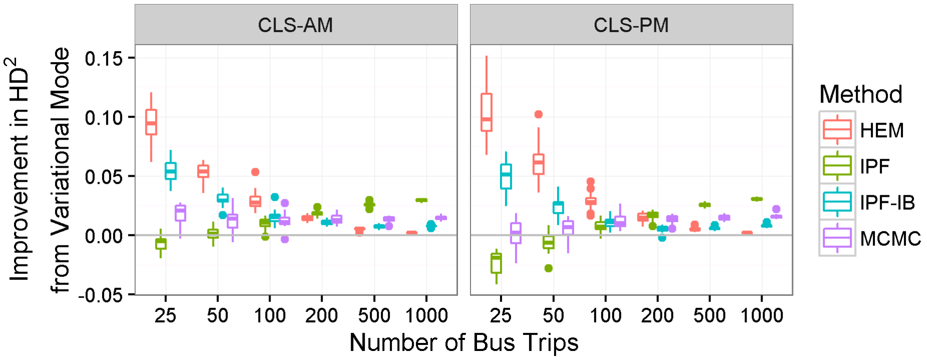

FIGURE 3: Distributional accuracy comparisons considering differences of HD^2 values of the variational mode estimates and the estimates of each of the other methods for two route-periods

4. 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 and from a variety of ground-truths would further validate the algorithm's effectiveness. In addition, empirical validations on a variety of routes would be essential to set the stage for adopting the developed methodology in future transit passenger OD flow studies for research or practice purposes. 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. OSU's CTL from which data for the simulation evaluation are used offers the opportunity to amass high-fidelity onboard OD flow survey that could be used for this purpose.

Given the high quality of the OD flow estimates arrived at by the variational methodology and its computational superiority with respect to other methods reported in the literature, the methodology allows for extending the estimation of OD flow patterns under more complex behavioral assumptions than is typically possible given the computational limitation of other methods. Specifically, due to varying trip purposes, which are associated with different sets of origin-destination pairs and departure times within a time-of-day period, OD flow patterns would be more realistically represented by multiple underlying probability OD flows. That is, it is conceivable that bus trips within a period could carry travelers that collectively exhibit different underlying probability OD flow patterns. The developed variational Bayes method could be extended to capture the presence of multiple underlying OD flow matrices in a computationally efficient manner.

5. REFERENCES

1. Bacharach, M., 1970. Biproportional matrices and input-output change. Cambridge University Press, London.

2. 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.

3. 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.

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

5. Chan, J., 2007. Rail OD estimation and journey time reliability metrics using automated fare data. M.S. Thesis, Massachusetts Institute of Technology, Cambridge, MA.

6. Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B: Methodological. 39(1), 1-38.

7. Efron, B., and Tibshirani, R.J., 1994. An introduction to the bootstrap. CRC press, New York.

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

9. 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.

10. 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.

11. 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.

12. 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.

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

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

15. 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.

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

17. Robert, C. P. and Casella, G., 2005. Monte Carlo Statistical Methods. New York, NY: Springer.

18. Wang, W., Attanucci, J., Wilson, N.H.M., 2011. Bus passenger origin-destination estimation and related analyses using automated data collection systems. Journal of Public Transportation, 14(4), 131-150.

19. 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