TRANSIT ROUTE ORIGIN-DESTINATION MATRIX ESTIMATION USING COMPRESSED SENSING

Pramesh Kumar
Department of Civil, Environmental, and Geo-Engineering
University of Minnesota - Twin Cities
500 Pillsbury Drive S.E.
Minneapolis, MN 55455, USA
Tel: (716) 903-2366
Fax: (612) 626-7750
Email: kumar372@umn.edu

Alireza Khani, Corresponding Author
Department of Civil, Environmental, and Geo-Engineering,
University of Minnesota - Twin Cities,
500 Pillsbury Drive S.E.,
Minneapolis, MN 55455, USA
Tel: (612) 624-4411
Fax: (612) 626-7750;
Email: akhani@umn.edu

Gary A. Davis
Department of Civil, Environmental, and Geo-Engineering,
University of Minnesota - Twin Cities,
500 Pillsbury Drive S.E.,
Minneapolis, MN 55455, USA
Tel: (612) 625-2598
Fax: (612) 626-7750;
Email: drtrips@umn.edu

Word count: 4,985 words text + 1 table x 250 words (each) = 5,235 words

Submitted for publication in Transportation Research Record
Submission Date: November 15, 2018

ABSTRACT

Development of an origin-destination (OD) demand matrix is crucial for transit planning. With the help of automated data, it is possible to estimate a stop-level OD matrix. We propose a novel method for estimating transit route origin-destination (OD) matrix using Automatic Passenger Count (APC) data. The method uses l_0 norm regularizer, which leverages the sparsity in the actual OD matrix. The technique is popularly known as compressed sensing (CS). We also discuss the mathematical properties of the proposed optimization program and the complexity of solving it. We use simulation to assess the accuracy and efficiency of the method and found that the proposed method is able to recover the actual matrix within small errors. With increased sparsity in the actual OD matrix, the solution gets closer to the actual value of the matrix. The method was found to perform more efficiently even for different demand patterns. We also present a real numerical example of OD estimation of A Line BRT route in Twin Cities, MN.

Keywords: rigin-desintion (OD) matrix, transit, compressed sensing, Lasso, sparsity, l_0 norm, l_1 norm, Automatic Passenger Count (APC), automated data

The authors confirm contribution to the paper as follows: study conception and design: Pramesh Kumar, Alireza Khani, Gary A. Davis; data collection: Pramesh Kumar, Alireza Khani, Gary A. Davis; analysis and interpretation of results: Pramesh Kumar, Alireza Khani, Gary A. Davis; draft manuscript preparation: Pramesh Kumar, Alireza Khani, Gary A. Davis, All authors reviewed the results and approved the final version of the manuscript.

INTRODUCTION AND LITERATURE REVIEW

To understand the travel pattern of passengers, transit agencies require an origin-destination (OD) flow matrix of the passengers. This is a fundamental element of interest that helps in designing new routes and schedules, understanding and forecasting demand on transit network, adjusting marketing strategy, etc. The transit OD flow matrix is the quantification of the flow of passengers from one transit stop to another. To evaluate such a matrix, the agencies conduct on-board surveys, which collect data about passenger boarding and alighting stops, the purpose of travel, etc. These surveys are expensive to conduct and cover only a small sample of passengers (*1*). However, with recent advancement in automated data collection systems (ADCS), it is possible to mine the full origin-destination matrix. The automated data such as Automatic Fare Collection (AFC) and Automatic Passenger Count (APC) data are a rich source about information of passenger travel over a continuous period, making it possible to estimate OD matrix more frequently.

The OD estimation problem has attracted the attention of many researchers over several decades. More recently, the use of automated data such as AFC, APC, or cell phone data has become popular for OD estimation. For example, the AFC data can be used to estimate a stop level OD matrix. It usually lacks the passenger alighting stop, which can be inferred using a trip-chaining algorithm based on several assumptions (*2–5*). The inference rate depends on the quality of data, the percentage of passengers using the smart card, and assumptions involved in the trip-chaining algorithm. On the other hand, APC systems collect information about the number of passenger boarding and alighting at each transit stop. OD estimation using the boarding and the alighting counts is a classic problem, which is hard to solve. The problem requires solving an underdetermined system of equations, in which case the number of unknowns to solve is far more than the number of equations available. Usually, multiple solutions are possible for this problem, which satisfy the given equations. Other information is supplemented to produce the accurate OD flows. To deal with this underdetermined problem, various methods have been proposed in the literature, which is summarized below:

1. *Iterative Proportional Fitting (IPF) method*: This is a popular and easy to apply method to evaluate the OD matrix using count data (*6*, *7*). The method starts with a base matrix, which is improved iteratively by multiplying the columns and rows of the matrix by a constant factor. The base matrix can be taken as a null matrix or any other seed matrix. Mishalani et al. found that using onboard survey data as a base matrix gives more accurate results than using null base matrix (*8*). The method has several issues such as the problem of non-structural zeros (*6*), due to which a zero entry remains zero in every iteration. The method also fails to converge if the number of zero entries become large in the matrix.
2. *Bayesian inference methods*: These methods use Bayesian approach to evaluate an OD matrix by formulating the problem as a partially observed Markov chain and utilizing prior information along with current observations of count data (*9–12*).
3. *Optimization methods*: As there are multiple solutions possible for this system of equations, these methods try to find the one, which optimizes an objective. The objective can be maximizing entropy (*13*) or the likelihood (*14–16*) function. With isotropic Gaussian noise, the maximum likelihood estimation turns into a classic least squares problem.

Another class of optimization methods consider above objectives along with a regularizer. The regularizer helps to mitigate the ill-posedness of the system of equations. The regularization can be included as a least square term between the unknown and a prior OD matrix obtained from a survey or from domain knowledge. This technique is quite popular in the literature. For example, Cascetta and Nguyen minimized generalized least square objective with a prior matrix (*7*), Van Zuylen and Willumsen maximized the relative entropy or minimized the Kullback-Leibler (KL) divergence of unobserved and observed flow distributions (*13*). This approach tries to force the solution, as close to the prior matrix as possible which may result in poor estimates if the prior or seed matrix used is not reliable.

In this research, we evaluate the transit route OD matrix using APC data. The problem is the estimation of the flow of passengers between stops for a single trip. The route matrix problem has a special structure that provides an extra piece of information to reduce the ill-posedness of the system of equations. The estimation requires the selection of the correct estimate out of the multiple solutions. We use an estimation method that encourages the sparse OD matrix using l_0 norm regularizer. This helps in mitigating the ill-posedness of the system and offers interpretability (*17*) as there is only a subset of the origin-destination pairs which carries flow in an actual OD matrix. The method is popularly known as compressed sensing (*18*) and can also be viewed as least absolute shrinkage and selection operator (LASSO) regression proposed by (*19*).

The rest of the paper is structured as follows. Section 2 presents the methodology for sparse matrix recovery followed by results of the experiments in section 3, then limitations of this research and directions for future research are discussed in section 4. Finally, conclusions are presented in section 5.

METHODOLOGY

In this section, we present the method to estimate the route level OD matrix using boarding and alighting counts available from APC data. We use the following notations throughout the paper (Table 1):

Let N be the set of stops along a transit route at which passenger board or alight. We consider the boarding and alighting in a single direction. Let b_i and a_i be the observed number of passengers who board and alight at stop i = (1, 2, …, |N|) respectively. The values of b_i and a_i are obtained from APC data. Let $X = \{x_{ij}\} \in R^{|N|\times|N|}$ be the origin-destination flow matrix, where x_{ij} denotes the number of passengers boarding at stop i and alighting at stop j. The overall setup is shown in Figure 1. Let $x \in R^{|N|^2}$ be the vectorized form of matrix X i.e., $x = Vect(X)$. The estimation procedure is subject to the following constraints, which are taken from (*11*, *20*).

Constraints

1. If we sum the values of x_{ij} along all the columns, then we get the total number of passengers boarding at stop i i.e. b_i,

$$\sum_{j=1}^{|N|} x_{ij} = b_i \; \forall \; i \in N \tag{1}$$

2. Similarly, if we sum the values of x_{ij} along all the rows, then we get the total number of passengers alighting at stop j, i.e., a_j

$$\sum_{i=1}^{|N|} x_{ij} = a_j \; \forall \; j \in N \tag{2}$$

3. The total number of boarding at all the stops should be equal to the total number of alighting.

$$\sum_{j=1}^{|N|} b_j = \sum_{i=1}^{|N|} a_i \tag{3}$$

4. The number of boarding and alighting at the same stop is zero, which means the diagonal elements of the matrix X should be equal to zero.

$$x_{ii} = 0 \; \forall \; i \in N \tag{4}$$

5. As the transit vehicle runs in a single direction, a passenger boarding at one stop cannot alight at the previous stops that vehicle has already visited. This means,

$$x_{ij} = 0 \;\; \forall \; i > j, \quad \forall \; i,j \in N \tag{5}$$

7. The total load on a link between two stops is equal to the passengers boarding between those stops.

$$\sum_{i=1}^{k} (b_i - a_i) = \sum_{i=1}^{k} \sum_{j=k+1}^{n} x_{ij} \tag{6}$$

By imposing these constraints, the structure of the matrix will look as in Figure 2.

We can express the linear constraints (1)-(6), in form of a matrix as

$$\mathcal{A}(x) = b \qquad (7)$$

Where, $\mathcal{A} \in R^{p \times |N|^2}$ is the linear map (which is a matrix in this case) for p number of constraints and $b \in R^p$ represents the constant vector for these constraints. In the following subsections, we describe the proposed solution to the given problem.

Transit route OD estimation using compressed sensing technique

As discussed, (7) is usually an ill posed problem for which one can expect multiple solutions. A generic regularizer can help in mitigating the ill posedness of the problem. One such regularizer is the generalized least square of the difference between the unkown and a prior matrix obtained from the survey data. The quality of the solution depends upon the availability of a good prior matrix as the optimal solution is forced to be as close as possible to the prior matrix. We can use other regularizers based on the domain knowledge on the space of the plausible OD flows in the network (*17*). To use one such regularizer, we make the following assumption:

Assumption *The planted OD matrix in the set of linear equations is sparse which means that the flow between many of the OD pairs should be equal to zero. The observed flow is only due to a small subset of* $\frac{N(N-1)}{2}$ *pairs.*

The intuition behind the above assumption is that there is a large number of OD pairs for a transit route, but the travel happens only along few pairs. For example, during the morning peak hours, there are only a few popular origin stops such as residential locations and few destination stops such as central business areas, park and rides, etc. Moreover, it is unlikely that passengers boarding at initial stops of the route will alight at all the following stops. This makes the flow between most of the OD pairs equal to zero. This is opposite to the solution evaluated using entropy maximization, which tries to achieve the solution, as uniform as possible to minimize the errors. The sparsity as a regularizer has been used before for highway network OD estimation and has found promising results (*17*, *21–24*). For example, (*17*) leverages sparsity in highway OD matrix to estimate a set of suitable traffic analysis zones (TAZs) and use those zones to evaluate an OD matrix. The method proposed in (*17*) has a bi-level structure with sparse OD estimation on upper level and traffic assignment using user equilibrium at lower level. The use of non-negativity constraints for improving the solution is also emphasized. This paper uses similar optimization for the transit route OD estimation problem, which has a special structure as we get an extra set of constraints because of transit movement in one direction. We also describe the conditions under which sparse recovery is possible.

Using sparsity as the regularizer for OD estimation

To achieve the sparsity in the solution, we minimize the number of non-zero entries in the solution, which can be done by minimizing l_0 norm of the vector x. We can state the problem as the minimization of l_0 norm of x subject to linear constraints. The optimization formulation is given below:

$$\begin{aligned} &minimize \ \|x\|_0 \\ s.t. \quad &\mathcal{A}(x) = b \\ &x \geq 0 \end{aligned} \tag{8}$$

Where, $\|x\|_0$ is the l_0 norm of vector x, which is defined as $lim_{p\to 0} \sum_j |x_j|^p$. The non-negativity should not be dropped from (8) as it helps to mitigate the ill-posedness of the problem (*17*). Using Lagrangian relaxation, the linear constraints can be included in the objective function as a least square term and formulated as following:

$$minimize_{x \geq 0} \ \ \|\mathcal{A}(x) - b\|_2 + \mu\|x\|_0 \tag{9}$$

The problem (9) tries to find the sparse vector x planted in the given ill-posed system of linear equations. The regularization parameter μ controls the sparsity of the vector and requires tuning to get the best results. A higher value of the μ will impose more sparsity in the solution. When $\mu = 0$, (9) reduces to an ordinary least squares problem. The optimization program (9) is useful for the APC data when the total number of boarding and alighting do not match as the least square term will try to find a solution which best explains the observed flows. This happens quite often in the APC systems due to the errors in recording data. The given problem (9) is an NP-hard as the minimization of l_0 norm cannot be done in polynomial time. Recent work in compressed sensing has proposed a tightest convex relaxation of the l_0 norm which is l_1 norm (*25*). The problem (9) can be restated as follows.

$$minimize_{x \geq 0} \ \ \|\mathcal{A}(x) - b\|_2 + \mu\|x\|_1 \tag{10}$$

where, $\mu\|x\|_1 = \sum_i |x_i|$. (10) is a convex optimization program as the absolute value of x_i can be written as a set of linear inequality constraints. The use of l_1 norm is better than the l_2 norm (also called ridge regression) to achieve sparsity. This is because the l_1 norm ball has corner points that can intersect the given plane at the sparsest solutions, unlike l_2 norm ball. The problem can also be viewed as least absolute shrinkage and selection operator (or Lasso regression) proposed by (*19*) as given a set of observations, we try to estimate the coefficients which satisfies the given equations. However, there is a key difference between compressed sensing and LASSO. The former provides conditions under which the linear map $\mathcal{A}$ is nicely behaved and the uniqueness of the solution can be proved (these conditions are discussed in the next subsection). In other words, we can design $\mathcal{A}$ in such a way that it can guarantee to recover the actual solution. On the other hand, LASSO is a regression method in which we have no control over the data and we try to find the best coefficients which are sparse and satisfy the equations obtained from data. We can also interpret these estimates as a Bayesian posterior mode estimate when the regression parameters have independent Laplace (i.e., double exponential) priors (*26*). Now the natural question which arises is that when does solving (10) gives a good solution to (9). In other words, what natural conditions can be applied on linear map $\mathcal{A}$ so that we can say that the solution is unique. Candès and Tao, 2005 proposed the idea of restricted isometry property (RIP) of the matrices, which states that if $\mathcal{A}$ satisfies the isometry property, then there exists a unique solution to the problem (10) which is equal to the solution of (9).

Restricted Isometry Property (RIP)

The linear map $\mathcal{A}$ has RIP with constants k and δ_k, if $\forall\ \|x\|_0 \le k$, $\mathcal{A}$ behaves almost as an isometry in following sense i.e. l_2 norm of $\mathcal{A}(x)$ is close to the l_2 norm of vector x:

$$(1-\delta_k)\|x\|_2^2 \le \|\mathcal{A}(x)\|_2^2 \le (1+\delta_k)\|x\|_2^2 \tag{11}$$

RIP matrices are extremely common in practice and most of the random matrices satisfy this property. Based on the above definition, a theorem is proposed by Candès and Tao, 2005 (*25*).

Theorem (Candès and Tao, 2005 (*25*))

If $\mathcal{A}(x) = b$ and b is constructed using a sparse solution with $\|x\|_0 \le k$, and the RIP condition is satisfied with constants δ_{2k} and δ_{3k} , satisfying $\delta_{2k} + \delta_{3k} < 1$, then (10) can obtain a unique solution to the problem (9) with as few as $\mathcal{O}\left(klog\left(\frac{|N|^2}{k}\right)\right)$ number of equations

As the passenger flow cannot be negative, we can replace the l_1 norm with sum of the components of vector x, which allow us to use the gradient-based approaches to solve the optimization program (10) efficiently. If we have some idea about the number of non-zero entries (say less than k), we can constraint the solution as follows:

$$\begin{aligned} &minimize\ \|\mathcal{A}(x) - b\|_2 \\ &s.t. \quad \|x\|_0 \le k \\ &\qquad\quad x \ge 0 \end{aligned} \tag{12}$$

We use the optimization program (8) with l_1 norm for solving the transit route OD estimation problem. The problem is convex and can be solved easily using a standard convex optimization solver such as CVX (*27*). We could also employ an iterative algorithm proposed in *(19)* to evaluate a sparse solution but the algorithm does not gurantee convergence to a unique solution. In the next section, we present numerical examples to show the application of the porposed method.

RESULTS

In this section, we present two numerical examples of OD estimation using the proposed methodology. First, simulation is used to assess the consistency and accuracy of the estimation method. Second, the OD estimation of a bus route in Twin Cities, MN is presented.

OD estimation using simulation

We use the APC data provided by Metro Transit, which is a primary transit service provider in the Twin Cities, MN area offering an integrated network of buses, light rail, bus rapid transit, and a commuter train. To prepare a synthetic OD matrix, we make some assumptions on the probability distribution of arrival of the passengers on different stops. We consider 10 stops along a transit route to facilitate the presentation of results. The passenger arrival at the stop is assumed to follow a Poisson distribution.

$$b_i \sim Poisson(k)\ \forall i \tag{13}$$

where k is the mean arrival rate at the stop and b_i is the number of boarding at stop i. We recommend fitting a Poisson distribution to the real data to calculate the value of k. We calculated the mean value of arrival rate of the passengers on the A line, an arterial BRT route in Twin Cities, MN. The mean arrival rate was found to be equal to 0.86 during peak hours, which is quite low. To assess the significant errors produced by the estimation, we assumed the mean value equal to 15 passengers. Then we set the sparsity level for the O-D matrix. The sparsity level will make the value of probability of flow from one stop to another stop zero if this probability value is less than the threshold sparsity level. This is done to create sparsity in the matrix and to test whether the method works more efficiently when the sparsity is high. We then assign the flow from one stop to others by assuming a multinomial distribution i.e.

$$x_{ij} \sim MNL(b_i, p_{i1}, p_{i2}, \dots, p_{i|N|}) \quad (14)$$

Where p_{ij} is the probability of movement from stop i to stop j. We then make the diagonal and lower triangle of the matrix zero because of the constraints (4-5). To calculate the boarding and alighting flows for O-D estimation, we sum the rows and columns of the simulated matrix. After that, we set up an optimization model using the python API of CVX (*27*). To avoid choosing the value of μ in optimization program (10), we solved the program (8) with l_1 norm. However, we recommend using the optimization program (10) when the sum of boarding and alighting count do not match in the APC data, which happens because of the errors in data collection. We generated 200 Monte-Carlo samples of OD matrices and calculated the l_2 error between the actual OD x and estimated OD vector x_{est} as:

$$\|x - x_{est}\|_2 = \sqrt{\sum_i (x_i - x_{est,i})^2} \quad (15)$$

Figure 3 shows the histogram of l_2 error in the estimation for each sample. We can observe that the mean value of the error is 4.99 and with a standard deviation of 1.32. The 95% confidence interval of the l_2 error was found to be equal to (4.81, 5.19). This shows that the results obtained from this estimation method are consistent and small. To see how the method performed in predicting the individual origin-destination pair flow value, we created a box plot for the estimation error (Figure 4). The proposed method predicted the actual value of the non-zero entries 41.5 % of the time. In case of errors, the method seems to overpredict the values except some of the O-D pairs such as 0-4, 1-2 and 5-6.

Figure 5(a) shows the average load profile of the passengers on the transit route. The width of the 95% confidence interval is small which shows that the method is reliable in estimating demand and therefore in deciding the adequate frequency to handle the load of the passengers. We can also observe that the errors in estimating the load of the passengers is also quite small (Figure 5(b)).

To understand the effect of sparsity, we solved the problem for several levels of sparsity and calculated the root mean square error (RMSE) between the estimated and actual OD matrix. Figure 6 shows the RMSE value with respect to the sparsity in the matrix. We can observe that the RMSE value is reduced with increased sparsity. For example, when the OD matrix has only 10% non-zero values, the corresponding RMSE value was found to be less than 0.35, which is quite impressive. This shows that the accuracy of the method is improved when there is more sparsity. Comparing the results to common least squares solution (Figure 6), the proposed method is able to recover solutions with lower RMSE value. It can also be observed that when the sparsity is low, the proposed method is more efficient than least squares as the gap between two lines is high but when there are a greater number of non-zero entries, the RMSE gap between these two methods reduces.

To see how different demand patterns affect the OD estimation, we performed a similar simulation for several mean arrival rates(k) of passengers at stops. Figure 7 shows normalized RMSE values with respect to sparsity in the random matrix for different mean arrival rate. The normalization is done by simply dividing RMSE by mean arrival rate. The results are presented in separate panels. We can see that the normalized RMSE decreases with an increase in demand. At k = 2, the normalized RMSE value was found to be almost equal to 1, which is still quite low. At lower demand, the matrix is already sparse, so we see less effect of sparsity parameter.

OD estimation of A Line BRT route in Twin Cities

We use the APC data from Twin Cities, MN to calculate the origin-destination flow of a route. The Automatic Passenger Count (APC) data used for this research contains transit trip information, such as date and time of the operation, routeID, stopID, departure and arrival time, number of boarding and alighting on each stop, and the geographical coordinates of the stops. We select A Line, which is a bus rapid transit (BRT) route in Twin Cities for this analysis. It serves 20 stations along Snelling Av and 46th St. We select a trip from the data during peak hour. The number of boarding and alighting at different stops in the northbound direction is shown in Figure 8. We can observe the popular boarding locations such as 46th street station, 46th & Minnehaha station, and Snelling & Highland station and alighting stops such as Rosedale transit center, Snelling & Highland station and Snelling & Clair st. station. We use the optimization program (10) to solve the given problem with a value of $\mu = 0.2$. Some recommendations for choosing the value of μ is given in (*19*).

The total ridership of the trip is 16. Because of low ridership, flow along most of the O-D pairs should be equal to zero. We apply the proposed method to the given data and calculate the origin-destination flows. Figure 9 shows the origin-destination flows between different O-D pairs. We can see that the flow occurred only between 11 O-D pairs out of 400 pairs (2.75%). The highest flow was observed between Snelling & Highland Av and Rosedale Transit Center, which is the last station along this route. Other popular OD pairs are 46th St and Snelling & St. Clair, Snelling & Minnehaha and Snelling & Highland Av. Because of the low ridership, the sparse matrix recovery seems to perform well.

LIMITATIONS AND DISCUSSION

In this section, we discuss the limitations of the proposed method and provide some recommendations for future research to address these limitations. Various studies use a prior matrix as a regularizer which can also be included in the proposed framework as follows:

$$minimize_{x \geq 0} \ \|\mathcal{A}(x) - b\|_2 + \mu_1\|x\|_0 + \mu_2\left\|x - x^{prior}\right\|_2 \qquad (16)$$

The choice of parameters μ_1 and μ_2 will control the weight of different objectives which can be obtained by observing the error rate for different values of these parameters. Due to lack of the suitable prior matrix, we have not included any results using this program in this study. Also, no unique choice of μ_1 and μ_2 can make this method unattractive to practitioners.

The problem of OD estimation has been well studied in the literature both in the context of road and transit network. This is an interesting problem with solution methods using both optimization and statistical techniques. Depending on the available data, the problem can be formulated as an underdetermined or overdetermined system of equations. The classic techniques such as entropy maximization, least squares, etc. produce some good results but may not evaluate the correct solution as there can be infinitely many or no solutions, which again depends on the quality of data and the set of equations obtained from the setup. Recent work in the field of compressed sensing has established that under suitable conditions, we can evaluate a unique sparse solution out of given ill-posed system of linear equations. We tried to solve the problem using this approach and found impressive results but not an exact solution. We also did not prove that the linear mapping produced by a set of linear equations in this case satisfies the restricted isometry property, which may be the source of error in our results. The condition is hard to prove and can be a future research topic.

The compressed sensing technique can also be used to design a linear map $\mathcal{A}$ so that we can guarantee an exact solution to this problem. The future work in this regard should be focused on how to engineer a system in order to create a linear map $\mathcal{A}$ so that the recovery of an exact solution can be guaranteed. This can be done by finding optimal sensor locations on highway and transit network or integrating different data sources to produce an appropriate $\mathcal{A}$.

This research can also be expanded in multiple directions. The method can be used to estimate a full transit network OD matrix. The problem can be formulated as a bi-level program with sparse recovery optimization at the upper level and transit assignment (*28*, *29*) at the lower level to capture route choice behavior in the model. We believe that the network level OD will also be sparse because it is unlikely that passengers boarding at one stop can alight at all other stops in the network. The concept can also be extended to matrix sensing which will be helpful in estimating a time-dependent transit OD matrix. As the boardings and alightings follow a regular pattern during various hours of the day, data from several days can be used to learn this pattern. This means the high dimensional data for several days can be used to minimize the rank of the matrix to extract a regular pattern. This can be done by minimizing the nuclear norm of the matrix, which is a convex surrogate for the rank of the matrix. The problem is computationally challenging and needs further attention.

CONCLUSIONS

In this research, we proposed a method for estimating an origin-destination OD matrix for a transit route along one direction. The problem can be formulated as an undetermined system of linear equations. The adopted strategy was to estimate a sparse O-D matrix, using l_0 norm. Using its convex surrogate l_1 regularizer, the problem can be solved efficiently. The sparsity in the matrix is generated because there are only a few popular O-D pairs along a transit route where the flow occurs. We also discussed the complexity of solving the proposed optimization program. The constraints and sparsity try to force the solution to an actual value. We tested the efficiency of the estimator using simulation. The errors were found to be bound within a small range. With an increased level of sparsity in the matrix, the method was able to recover more accurate results. We also found small errors even for higher demand. For example, the normalized RMSE between estimated and actual matrix value was found to be at most 0.1. We also presented a numerical example for A-line BRT route in Twin Cities, MN. Finally, we discussed various limitations and directions for future research in section 4. Further studies are required to show under which constraints, the OD linear map satisfy the RIP property. Other statistical methods are also required to assess the accuracy of the estimation.

ACKNOWLEDGEMENT

This research was conducted at the University of Minnesota Transit Lab (http://umntransit.weebly.com/), currently supported by the following, but not limited to, projects:

- National Science Foundation, award CMMI-1637548

- Minnesota Department of Transportation, Contract No. 1003325 Work Order No. 15
- Minnesota Department of Transportation, Contract No. 1003325 Work Order No. 44
- Transitways Research Impact Program (TIRP), Contract No. A100460 Work Order No. UM2917

The authors are grateful to Metro Transit for sharing the data. We are also grateful to the anonymous referees for their constructive input to improve the quality of this article. Any limitation of this study remains the responsibility of the authors.

REFERENCES

1. Attanucci, J. & Wilson, N. H. M. Bus Transit Monitoring Manual: Volume 1: Data Collection Program Design. *US Department of Transportation*, Vol. 1, No. August, 1981.
2. Kumar, P., A. Khani, and Q. He. A Robust Method for Estimating Transit Passenger Trajectories Using Automated Data. *Transportation Research Part C: Emerging Technologies*, Vol. 95, 2018. https://doi.org/10.1016/j.trc.2018.08.006.
3. Nassir, N., A. Khani, S. Lee, H. Noh, and M. Hickman. Transit Stop-Level Origin-Destination Estimation Through Use of Transit Schedule and Automated Data Collection System. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2263, 2011, pp. 140–150. https://doi.org/10.3141/2263-16.
4. Trépanier, M., N. Tranchant, and R. Chapleau. Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems*, Vol. 11, No. 1, 2007, pp. 1–14. https://doi.org/10.1080/15472450601122256.
5. Zhao, J., a Rahbee, and N. H. . Wilson. Estimating a Rail Passenger Trip Origin-Destination Using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering*, Vol. 22, No. 5, 2007, pp. 376–387.
6. Ben-Akiva, E. al. Alternative Methods to Estimate Route-Level Trip Tables and Expand On-Board Surveys. *Transportation Research Record 1037, TRB, National Research Council, Washington, D.C.,* p. 1–11.
7. Cascetta, E., and S. Nguyen. A Unified Framework for Estimating or Updating Origin/Destination Matrices from Traffic Counts. *Transportation Research Part B*, Vol. 22, No. 6, 1988, pp. 437–455. https://doi.org/10.1016/0191-2615(88)90024-0.
8. Mishalani, R., Y. Ji, and M. McCord. Effect of Onboard Survey Sample Size on Estimation of Transit Bus Route Passenger Origin-Destination Flow Matrix Using Automatic Passenger Counter Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2246, 2011, pp. 64–73. https://doi.org/10.3141/2246-09.
9. Maher, M. J. Inferences on Trip Matrices from Observations on Link Volumes: A Bayesian Statistical Approach. *Transportation Research Part B: Methodological*, Vol. 17, No. 6, 1983, pp. 435–447.
10. Gary A . Davis. Estimating Freeway Demand Patterns and Impact of Uncertainity on Ramps Controls. *ASCE Journal of Transportation Engineering*, Vol. 119, No. 4, 1993, pp. 489–503.
11. Li, B. Markov Models for Bayesian Analysis about Transit Route Origin-Destination Matrices. *Transportation Research Part B: Methodological*, Vol. 43, No. 3, 2009, pp. 301–310. https://doi.org/10.1016/j.trb.2008.07.001.
12. Hazelton, M. L. Statistical Inference for Transit System Origin-Destination Matrices. *Technometrics*, Vol. 52, No. 2, 2010, pp. 221–230. https://doi.org/10.1198/TECH.2010.09021.
13. Van Zuylen, H. J., and L. G. Willumsen. The Most Likely Trip Matrix Estimated from Traffic Counts. *Transportation Research Part B: Methodological*, Vol. 14, No. 3, 1980, pp. 281–293. https://doi.org/10.1016/0191-2615(80)90008-9.
14. Nihan, N. L., and G. A. Davis. Application of Prediction-Error Minimization and Maximum Likelihood to Estimate Intersection O-D Matrices from Traffic Counts. *Transportation Science*. 2. Volume 23, 77.

http://ezproxy.unicartagena.edu.co:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=6798710&lang=es&site=ehost-live.
15. Spiess, H. A Maximum Likelihood Model For Estimating Origin-Destination Matrices. *Transportation Research Board*, Vol. 21B, No. 5, 1987, pp. 395–412. https://doi.org/10.1016/0191-2615(87)90037-3.
16. Vardi, Y. Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data. *Journal of the American Statistical Association*, Vol. 91, No. 433, 1996, pp. 365–377. https://doi.org/10.1080/01621459.1996.10476697.
17. Menon, A. K., C. Cai, W. Wang, T. Wen, and F. Chen. Fine-Grained OD Estimation with Automated Zoning and Sparsity Regularisation. *Transportation Research Part B: Methodological*, Vol. 80, 2015, pp. 150–172. https://doi.org/10.1016/j.trb.2015.07.003.
18. Candes, E. J., and M. B. Wakin. An Introduction To Compressive Sampling. *IEEE Signal Processing Magazine*, Vol. 25, No. 2, 2008, pp. 21–30. https://doi.org/10.1109/MSP.2007.914731.
19. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1, 1996, pp. 267–288.
20. Kikuchi, S., and N. Kronprasert. Constructing Transit Origin-Destination Tables from Fragmented Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2196, 2010, pp. 34–44. https://doi.org/10.3141/2196-04.
21. Sanandaji, B. M., and P. Varaiya. Compressive Origin-Destination Estimation. *Transportation Letters*, Vol. 8, No. 3, 2016, pp. 148–157. https://doi.org/10.1179/1942787515Y.0000000018.
22. Mardani, M., and G. B. Giannakis. F EF. No. 00, pp. 1–5.
23. Chawla, S., Y. Zheng, and J. Hu. Inferring the Root Cause in Road Traffic Anomalies. 2012. https://doi.org/10.1109/ICDM.2012.104.
24. Zhang, Y., Z. Ge, A. Greenberg, M. Roughan, and F. Park. Network Anomography.
25. Candès, E., and T. Tao. Decoding by Linear ProgrammingEmmanuel Candes†. *IEEE Trans. Inf. Theory*, Vol. 51, No. 12, 2005, pp. 4203–4215.
26. Park, T., and G. Casella. The Bayesian Lasso. Vol. 103, No. 482, 2008, pp. 681–686. https://doi.org/10.1198/016214508000000337.
27. Michael Grant and Stephen Boyd. CVX: Matlab Software for Disciplined Convex Programming, Version 2.1. http://cvxr.com/cvx.
28. Alireza Khani, Elizabeth Sall, Lisa Zorn, M. H. Integration of the FAST-TrIPs Person-Based Dynamic Transit Assignment Model, the SF-CHAMP Regional, Activity-Based Travel Demand Model, and San Francisco's Citywide Dynamic Traffic Assignment Model. 2013.
29. Khani, A., M. Hickman, and H. Noh. Trip-Based Path Algorithms Using the Transit Network Hierarchy. *Networks and Spatial Economics*, Vol. 15, No. 3, 2014, pp. 635–653. https://doi.org/10.1007/s11067-014-9249-3.

TABLE 1. Notations

Variable	*Definition*
N	Set of stops/stations along a transit route
i	Index for transit stop
b_i	Number of passengers boarding at stop i
a_i	Number of passengers alighting at stop i
X	Origin destination flow matrix
x	Vector form of OD matrix X
$\|x\|_0$	l_0 norm of vector x, $\|x\|_0 = \lim_{p\to 0} \sum_i \lvert x_i \rvert^p$
$\|x\|_1$	l_1 norm of vector x, $\|x\|_1 = \sum_i \lvert x_i \rvert$
A	Linear map on a vector
Z	Set of integers

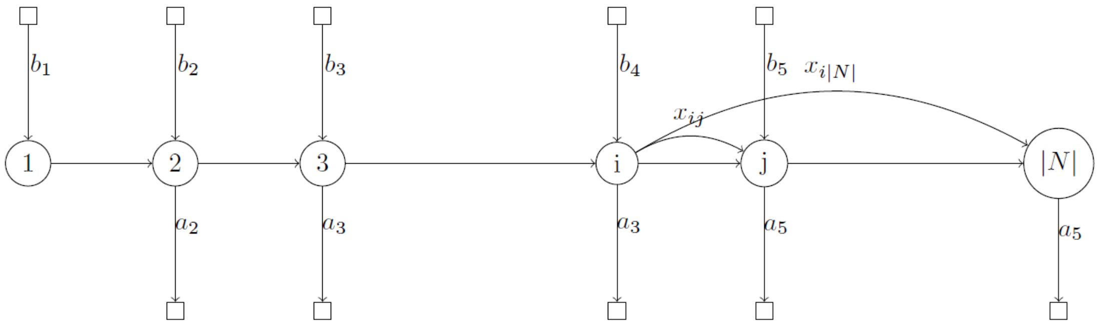

FIGURE 1. Transit route origin-destination (OD) flow

Alighting

		1	2	.	.	.	n	
	1	0						b_1
	2	0	0					.
Boarding	.	0	0	0				.
	.	0	0	0	0			.
	.	0	0	0	0	0		.
	n	0	0	0	0	0	0	b_n
		a_1	.	.			a_n	T

FIGURE 2. OD matrix for a route in a single direction

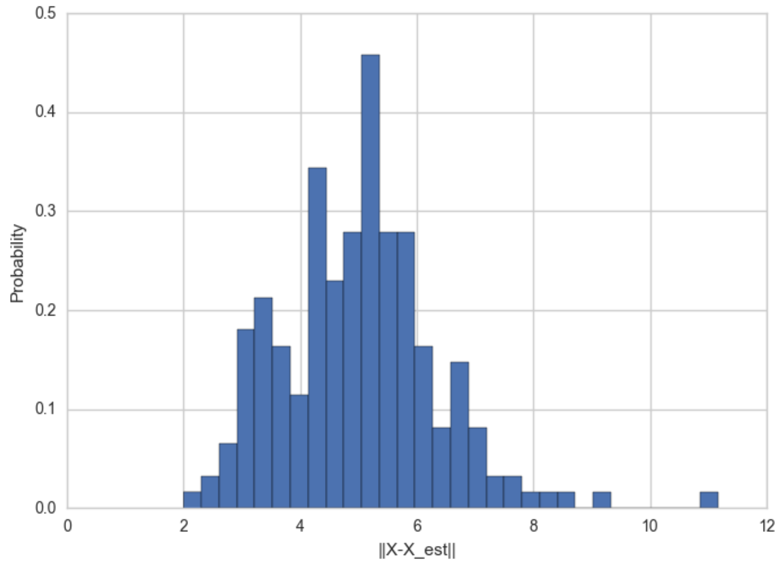

FIGURE 3. l_2 error between the actual and estimated OD matrix

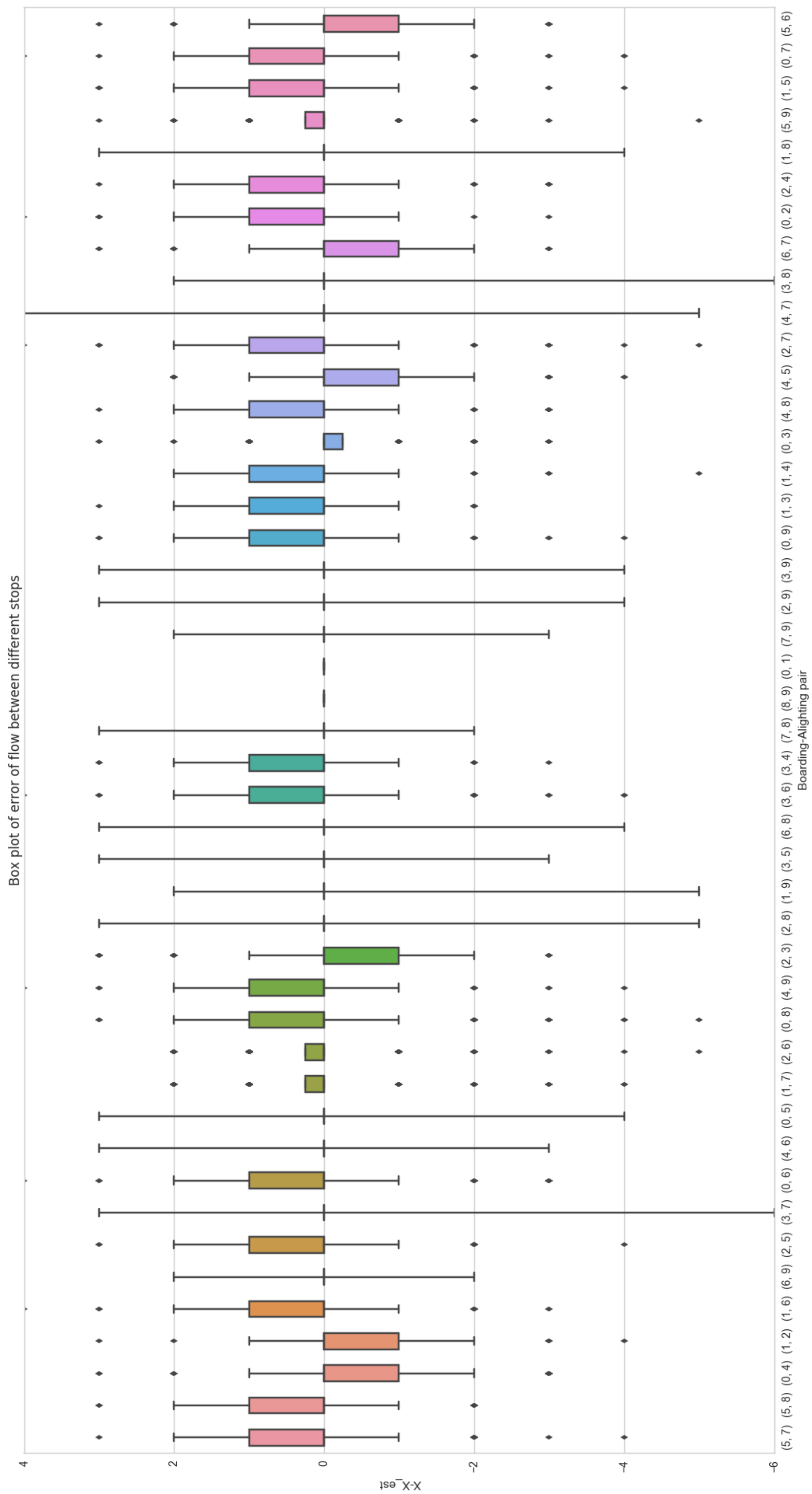

FIGURE 4. Box plot for the errors in estimation of O-D flows

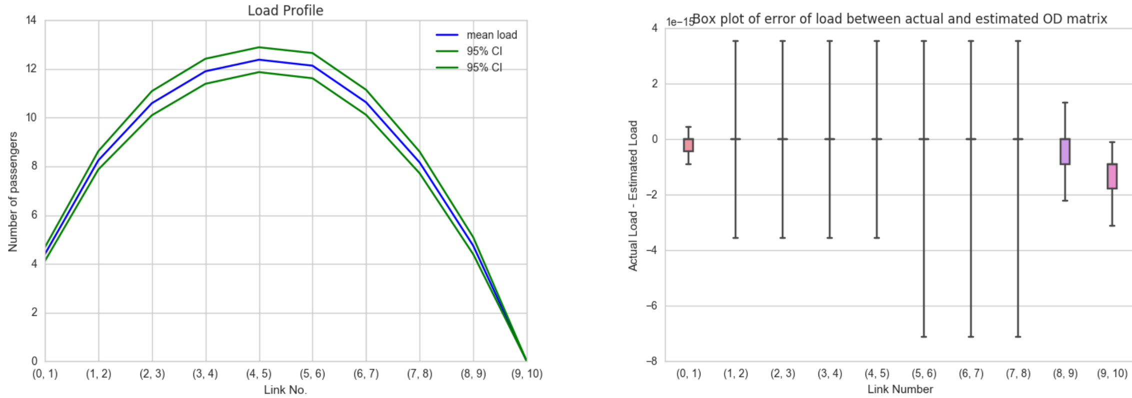

FIGURE 5. (a) Average load profile of the transit route. (b) Box plot of error between actual load and estimated load

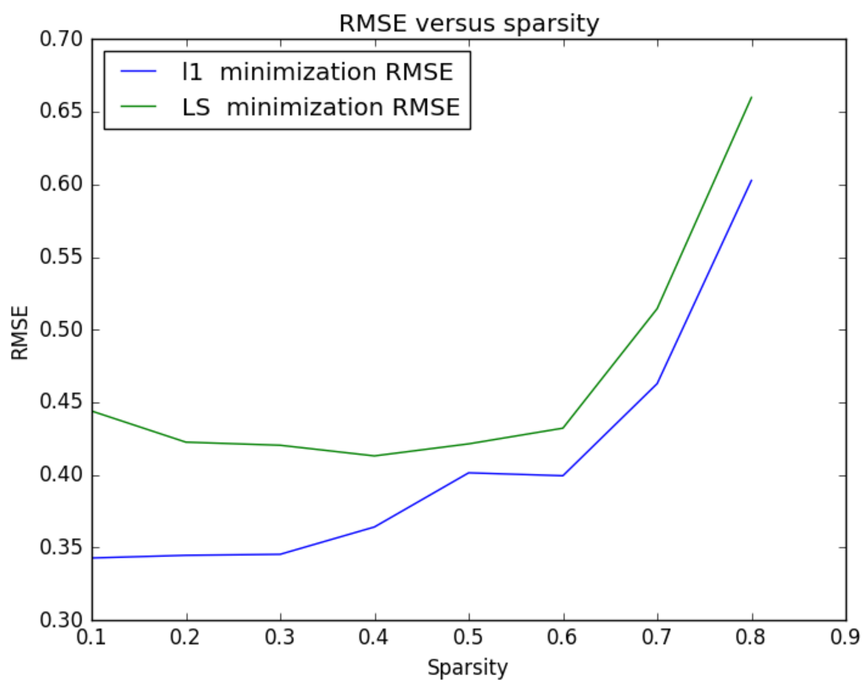

FIGURE 6. Root mean square error (RMSE) versus sparsity in OD estimation (Sparsity is in terms of proportion of non-zero values)

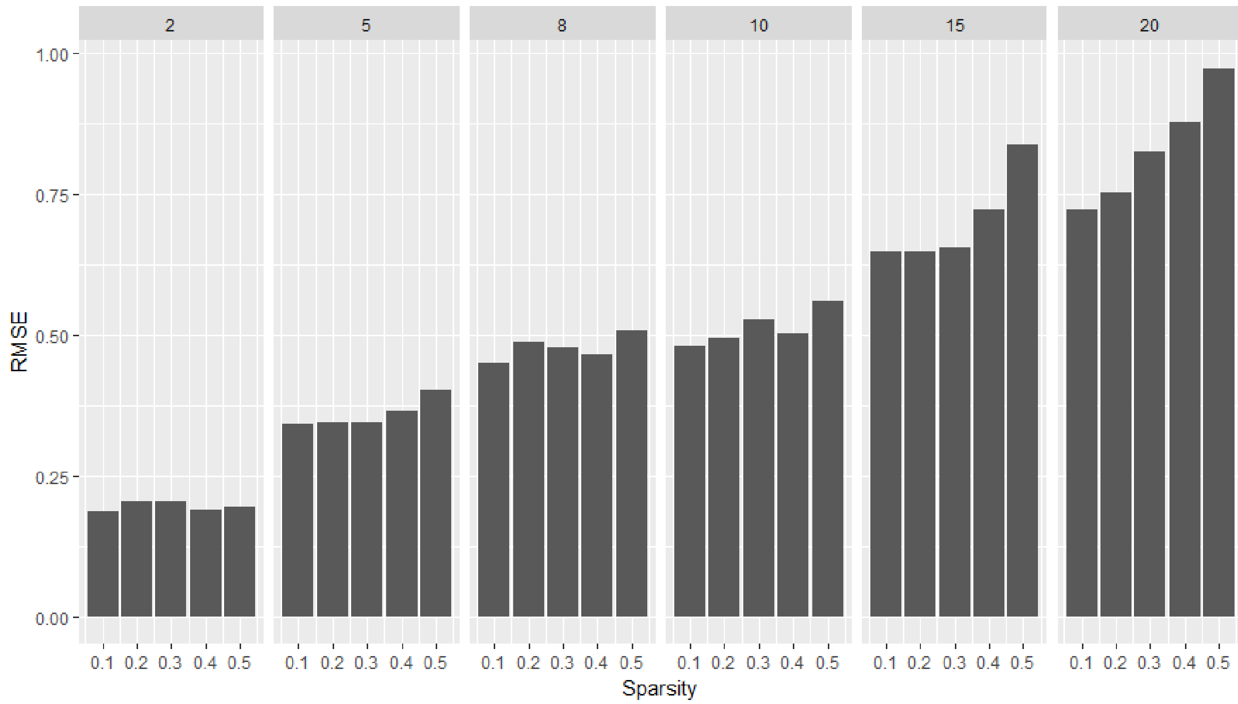

FIGURE 7. Comparing RMSE with sparsity for different mean arrival rates (each panel represents different mean arrival rate of passengers)

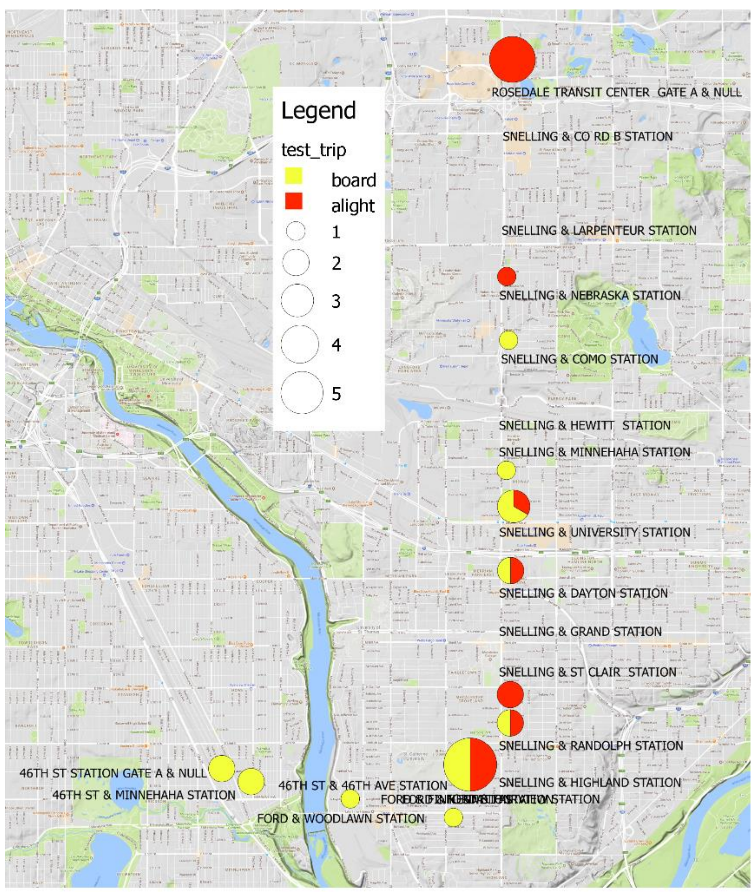

FIGURE 8. Boarding and alighting count for A Line

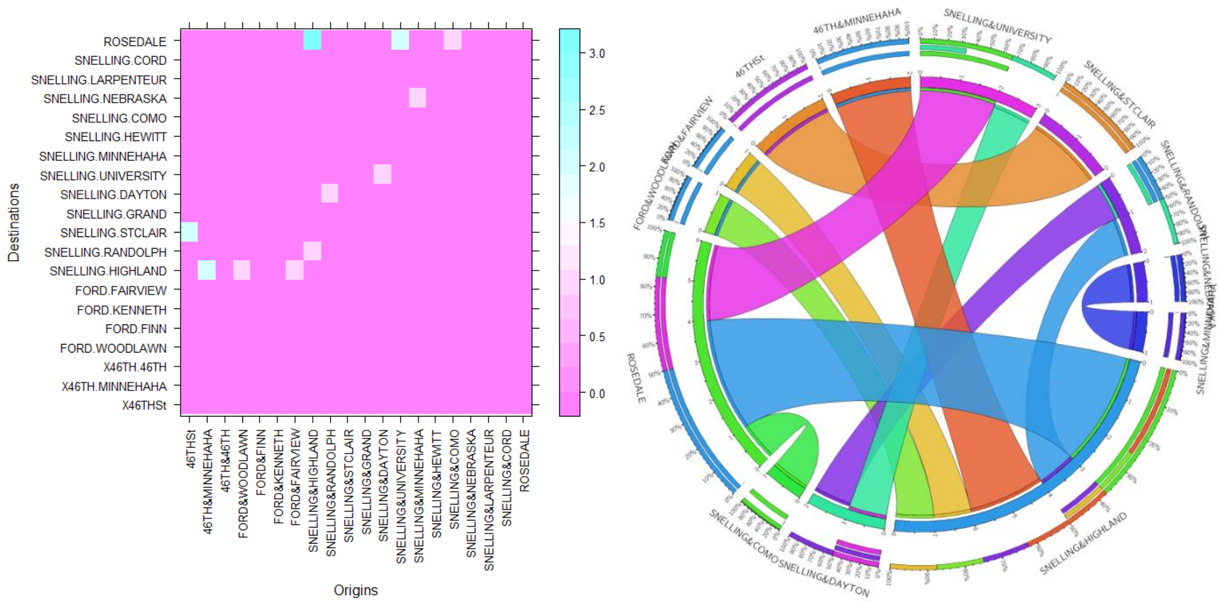

FIGURE 9. Origin-Destination flow for A Line, Twin Cities, MN