From Physical Properties of Transportation Flows to Demand Estimation: An Optimization Approach

Dimitris Bertsimas,a,b Julia Yanb

a Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; b Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Abstract. Efficient management of transportation systems requires accurate trip demand data. It is sometimes possible to track anonymized users through their commutes, accomplished through previous studies on smart cards, license plates, and mobile phones. However, frequently the main public transit data sources are in aggregated forms such as entry and exit counts, and one must recover the original demand from these aggregated counts. Such problems are generally underspecified. To address this, we present an optimization framework to recover origin–destination matrices under minimal assumptions, incorporating reasonable physical constraints such as flow conservation, smoothness, and symmetry. The proposed method is evaluated and shows strong (∼5%–10%) improvement in $R^2$ over the maximum entropy method on a variety of real-world datasets from Boston, New York City, and San Francisco, comprising tens to hundreds of stations.

Funding: This work was partially supported by a grant to MIT by the Mitsubishi Electric Corporation.

Keywords: demand estimation • transportation • optimization

1. Introduction

Time-dependent origin–destination (OD) demand data are a necessary input to public transportation planning models. However, even when transit data sets are available, they are frequently missing the origin, destination, or time component. Data sets in their ideal, most detailed form can be found, for example, in subway systems with distance-based or zone-based fares; both swipe-in and swipe-out of the system are required for fare collection, preserving the time, origin, and destination components of the commute. However, in the case of flat-fare subway systems, the lack of individual swipe-out mechanisms means the OD component of the commute is lost, and data are usually collected only as aggregated turnstile entrances and exits. Transportation planning surveys provide another source of OD demand data, but they too are missing a dimension of the data: in particular, they lose any time resolution above the “peak” versus “nonpeak” level. A further issue is that such surveys are costly to conduct and are collected only once or twice a decade, meaning they are rarely up-to-date.

Because of the aforementioned limitations in the available data, the problem of estimating OD demand from any given data set is generally underspecified and can give rise to many solutions. The challenge of finding a high-quality unique solution is well studied in the literature.

For traffic problems, the most common data sources are in link counts and survey data. Prior works have largely aimed to constrain the estimation by unifying these data sets. Such approaches account for the unreliability of survey data sets relative to link count data sets by using methods such as Bayesian statistics in Maher (1983); generalized least squares in Cascetta (1984), Cascetta et al. (2013) and Bell (1991); maximum likelihood in Spiess (1987); and multiobjective optimization in Brenninger-Göthe, Jörnsten, and Lundgren (1989).

In response to the influence that the target matrix can have on OD estimates, other works aim to reduce the dependence on costly and out-of-date survey data. Zhou and Mahmassani (2007) and Ashok and Ben-Akiva (2000, 2002) model the deviations from survey data explicitly and use link counts to update a historical trends model of the OD demand. Other works use the link count data alone. To further specify the problem, assumptions are made about structural properties of the OD matrices. For example, Van Zuylen and Willumsen (1980) and Fisk (1988) both search for OD matrices of maximum entropy (ME), while Yang et al. (1992) and Maher, Zhang, and Van Vliet (2001)
ensure that their estimations satisfy network equilibrium properties.

The increasing recent focus on big data has also made newer, richer data sources available for use. As more public transit systems turn to smart-card-based systems, it becomes possible to track individual cardholders through the transit system, even when considering flat-fare swipe-in systems where only the origins are known. Trip chaining, which assumes a trip’s destination to be the (known) subsequent trip’s origin, has been leveraged to great effect in generating OD demand matrices for the New York City subway system by Barry et al. (2002) and for the Chicago rail system by Zhao, Rahbee, and Wilson (2007). Although there are no smart-card data sets for traffic OD problems, additional data sources have been successfully used to supplement link count data: Castillo, Menéndez, and Jiménez (2008) and Zhou and Mahmassani (2006) use automatic vehicle identification data, and Calabrese et al. (2011) and Iqbal et al. (2014) use mobile phone data.

However, the data sets that are publicly available are most often in the form of turnstile entry and exit counts, and more detailed data sets can be difficult to come by. As such, we propose a method of demand estimation that makes the following contributions:

1. It utilizes publicly available data sets.
2. It utilizes noninvasive, aggregated data sets to respect passenger privacy concerns.
3. We impose rather light assumptions on the physical attributes of the transportation flows: temporal and spatial smoothness and symmetry.

Practically, our method offers advantages over existing approaches in that it is significantly less invasive than tracking card-level swipe information, as well as less resource intensive than conducting a survey. Many cities are already equipped to collect aggregate turnstile entry and exit data in their subway systems, and this method can be generally applied across many data sets without extra effort. Furthermore, although our focus is on transportation demand estimation, the assumptions imposed by the model are minimal, and we expect this approach to be applicable to other problems that take a similar matrix deaggregation-like form.

Our work is similar in philosophy to works focusing on maximum entropy or network equilibrium, which both specify desired properties of the estimation. However, we offer an interpretable list of desired physical properties rather than specifying maximum entropy alone. Additionally, these desired properties are intuitive and lighter than imposing a model on passenger behavior, as in the network equilibrium literature.

The rest of this paper is structured as follows. Section 2 presents the method for OD demand estimation: it requires only periodic entry and exit counts at each station in the network, and leverages intuitive physical properties of the solution matrices to find high-quality solutions. Section 3 presents a baseline method for comparison with that in Section 2. In Section 4, we cover real-world case studies, using bike and subway data in Boston, taxi data in New York City, and subway data in San Francisco, and show that our method performs well in general on multiple data sets. We offer concluding remarks in Section 5.

2. An Optimization Approach

We propose a formulation of the OD estimation problem using quadratic optimization (QO). A traditional approach in statistics is to minimize the deviations of the estimation from ground truth. However, in typical applications of OD demand estimation, we do not have a reliable target matrix, necessitating a different approach. As such, we use QO to impose constraints that our estimation properly satisfies the given data as well as physical properties of real-world transit matrices: namely, smoothness and symmetry. In Section 4, we will confirm the validity of such a method on data sets where ground truth is known and extend to a data set in a typical application where ground truth is not known.

2.1. Deaggregating the Data

For a system with \( N \) stations over \( T \) periods, we consider the case where our input data are of the form of periodic turnstile entries \( a \in \mathbb{R}^{N \times T} \) and exits \( b \in \mathbb{R}^{N \times T} \), where each element \( a_{u}^{t} \) \((b_{v}^{t})\) represents the number of entrances (exits) to station \( u \) during period \( t \). The OD estimation problem is then to construct from the input data an OD demand matrix \( d \in \mathbb{R}^{N \times N \times T} \), where each element \( d_{u \rightarrow v}^{t} \) represents the demand from station \( u \) to station \( v \) during period \( t \).

Assuming that the length of each time period (typically an hour) given in the data is long relative to the typical trip taken, the given input data are an aggregation of the true matrix elements of \( d \). The problem is then essentially one of recovering matrices where only the row sums and column sums are known. One fundamental constraint for such a problem is that the sum of the matrix elements across the rows and columns should agree with the provided data. For the OD estimation problem in particular, the number of entrances into a station should equal the total demand from that station over all possible destinations, that is,

\[
\sum_{u=1}^{N} d_{u \rightarrow v}^{t} = a_{u}^{t}, \quad \forall u = 1, \ldots, N, \forall t = 1, \ldots, T. \quad (1a)
\]

Similarly, the number of exits from a station should equal the total demand to that station over all possible origins, that is,

\[
\sum_{u=1}^{N} d_{u \rightarrow v}^{t} = b_{v}^{t}, \quad \forall v = 1, \ldots, N, \forall t = 1, \ldots, T. \quad (1b)
\]
In practice, it may be impossible to satisfy all of these constraints at once because of noise in the data. For example, subway data sets often show exit counts significantly lower than the entrance counts: this data incongruity can occur due to undercounting at the exit gates, as the gates can let many passengers out each time they open. Furthermore, some longer trips or trips that begin toward the end of a period may end in a later period from the starting period, so that the entry and exit data are not an exact aggregation of the OD matrix. To ensure feasibility, we introduce slack variables \( \epsilon^t_{u,v} \) and \( \xi^t_{u,v} \) to the constraints in Equation (1) as follows:

\[
\sum_{t=1}^N d^u_t - a^u_t \leq \epsilon^u_t, \quad \forall u = 1, \ldots, N, \forall t = 1, \ldots, T, \quad (2a)
\]

\[
\sum_{u=1}^N d^u_t - b^v_t \leq \epsilon^v_t, \quad \forall v = 1, \ldots, N, \forall t = 1, \ldots, T, \quad (2b)
\]

and add the slack variables to the objective function to be minimized, that is,

\[
\min \sum_{u,t} \left( \left( \frac{\epsilon^u_{u,v}}{a^u_t} \right)^2 + \left( \frac{\xi^u_{u,v}}{b^v_t} \right)^2 \right), \quad (3)
\]

so that the inferred OD demand matrix is one that best explains the provided data.

The system with only these data satisfaction constraints is, in general, underspecified. To further limit the degrees of freedom in the problem and find a unique, high-quality solution, we impose further constraints to ensure that the OD matrix will satisfy physical properties of real-world data.

### 2.2. Smoothness

A natural property of the estimate is it should be smooth over time while also explaining the data. In particular, the proportion of commuters entering station \( u \) who aim to make some commute \( u \to v \), relative to the entrance counts at station \( u \), should not change dramatically from period to period. Trip demand should be similarly smooth relative to the exit counts. More formally,

\[
\frac{d^u_{t} - d^u_{t-1}}{a^u_{t-1}} \leq \delta^u_{t-1}, \quad \forall u, v = 1, \ldots, N, \forall t = 2, \ldots, T, \quad (4a)
\]

\[
\frac{d^u_{t} - d^u_{t-1}}{b^v_{t-1}} \leq \lambda^u_{t-1}, \quad \forall u, v = 1, \ldots, N, \forall t = 2, \ldots, T. \quad (4b)
\]

Note that Equations (4a) and (4b) work with proportions of entrances and exits rather than the absolute volume of demand, as spiking traffic during peak hours is in clear violation of such a smoothness assumption.

In addition to the temporal smoothing constraints, a physical OD matrix should be reasonably smooth over space as well. To express this, we constrain neighboring stations to see similar demand patterns, that is,

\[
\frac{d^u_{t} - d^u_{t}}{a^u_t} - \frac{d^u_{t} - d^u_{t}}{a^u_t} \leq \delta^u_{t}, \quad \forall u, v = 1, \ldots, N, \forall t = 1, \ldots, T, \quad (5a)
\]

\[
\frac{d^u_{t} - d^u_{t}}{b^v_t} - \frac{d^u_{t} - d^u_{t}}{b^v_t} \leq \rho^u_{t}, \quad \forall u, v = 1, \ldots, N, \forall t = 1, \ldots, T, \quad (5b)
\]

where \( N(u) \) denotes the nearest neighbor of \( u \).

The temporal and spatial smoothing of Equations (4) and (5) take slightly different forms, but they both capture the natural intuition that across any features of a data array, nearest neighbors should behave similarly. This is made explicit in the spatial smoothing of Equation (5), but is also implicit in the temporal smoothing of Equation (4), where the “nearest neighbor” of period \( t \) is taken as the previous period \( t-1 \). In the case where some geometrically nearby stations are expected to show different demand patterns, \( \lambda(u) \) can be generalized to denote the nearest neighbor of \( u \), where distance is measured across some feature space, as in the \( k \)-nearest-neighbors algorithm.

We can find the smoothest possible function that also explains the data by adding the smoothing residual terms \( \delta^u_{t}, \lambda^u_{t}, \theta^u_{t}, \rho^u_{t} \) to the objective function as follows:

\[
\min \left\{ N \sum_{u,v} \left( \left( \frac{\epsilon^u_{u,v}}{a^u_t} \right)^2 + \left( \frac{\xi^u_{u,v}}{b^v_t} \right)^2 \right) + \sum_{u,v,t} \left( (\delta^u_{t})^2 + (\lambda^u_{t})^2 + (\theta^u_{t})^2 + (\rho^u_{t})^2 \right) \right\}. \quad (6)
\]

As observed by Brenninger-Götze, Jörnsten, and Lundgren (1989), who also take a multiobjective optimization approach, it is possible to change the coefficients on each objective (data satisfaction, temporal smoothing, and spatial smoothing) to reflect each objective’s relative importance to the multiobjective formulation. A natural weighting that we choose for each of the spatial smoothing terms uses the squared distance between the station and its nearest neighbor, which we denote \( D_u \) for each station \( u = 1, \ldots, N \). For the data satisfaction terms, we weight by the number of stations \( N \), to prevent the time smoothing and distance smoothing terms from dominating the data satisfaction terms when the number of stations is large.

### 2.3. Symmetry

Another reasonable property to impose on the estimation is that flow across the network should be symmetric; in particular, for any pair of stations \( u \) and \( v \),
the volume of $u \rightarrow v$ commutes should be roughly balanced by the volume of $v \rightarrow u$ commutes, capturing the intuition that passengers often make round trips. Although there are certainly cases where passengers do not make simple round trips, the demand flows should be roughly balanced in aggregate. To this end, we introduce the constraints

$$\sum_{t=1}^{T} d_{i}^{u\rightarrow v} - \sum_{t=1}^{T} d_{i}^{v\rightarrow u} \leq \phi_{u\rightarrow v}, \quad \forall u, v = 1, \ldots, N \ (7)$$

and add the slack variables $\phi_{u\rightarrow v}$ to the objective function to produce

$$\min \left\{ N \sum_{u, t} \left( \frac{\varepsilon_{u}}{a_{u}^{t}} \right)^{2} + \left( \frac{\varepsilon_{u}}{b_{u}^{t}} \right)^{2} \right\} + N^{2} T \sum_{u, v, t} \left( \frac{\phi_{u\rightarrow v}}{\sum_{u', v'} a_{u'}^{t}} \right)^{2} + \sum_{u, v, t} \left( (\delta_{u\rightarrow v}^{t})^{2} + (\lambda_{u\rightarrow v}^{t})^{2} \right)
+ \sum_{u, v, t} \left( (\theta_{u\rightarrow v}^{t})^{2} + (\rho_{u\rightarrow v}^{t})^{2} \right) \right) \right\} \ (8)$$

where the flow symmetry terms $\phi_{u\rightarrow v}$ are weighted to prevent them from dominating the objective function for high demand volumes.

Incorporating the data deaggregation, flow symmetry, and smoothness constraints, the full formulation is

$$\min \left\{ N \sum_{u, t} \left( \frac{\varepsilon_{u}}{a_{u}^{t}} \right)^{2} + \left( \frac{\varepsilon_{u}}{b_{u}^{t}} \right)^{2} \right\} + N^{2} T \sum_{u, v, t} \left( \frac{\phi_{u\rightarrow v}}{\sum_{u', v'} a_{u'}^{t}} \right)^{2} + \sum_{u, v, t} \left( (\delta_{u\rightarrow v}^{t})^{2} + (\lambda_{u\rightarrow v}^{t})^{2} \right)
+ \sum_{u, v, t} \left( (\theta_{u\rightarrow v}^{t})^{2} + (\rho_{u\rightarrow v}^{t})^{2} \right) \right) \right\} \ (9a)$$

s.t.

$$\sum_{t=1}^{N} d_{i}^{u\rightarrow v} - a_{i}^{u} \leq \varepsilon_{i}^{u}, \quad \forall u = 1, \ldots, N, \forall t = 1, \ldots, T, \ (9b)$$

$$\sum_{t=1}^{N} d_{i}^{u\rightarrow v} - b_{i}^{v} \leq \varepsilon_{i}^{v}, \quad \forall v = 1, \ldots, N, \forall t = 1, \ldots, T, \ (9c)$$

$$\left| \frac{d_{i}^{u\rightarrow v}}{a_{i}^{u}} - \frac{d_{i}^{u\rightarrow (v-1)}}{a_{i}^{u-1}} \right| \leq \delta_{u\rightarrow v}^{t}, \quad \forall u, v = 1, \ldots, N, \forall t = 1, \ldots, T, \ (9d)$$

$$\left| \frac{d_{i}^{u\rightarrow v}}{b_{i}^{v}} - \frac{d_{i}^{u\rightarrow (v-1)}}{b_{i}^{v-1}} \right| \leq \lambda_{u\rightarrow v}^{t}, \quad \forall u, v = 1, \ldots, N, \forall t = 2, \ldots, T, \ (9e)$$

$$\left| \frac{d_{i}^{u\rightarrow v}}{a_{i}^{u}} - \frac{d_{i}^{(v-1)\rightarrow u}}{a_{i}^{v-1}} \right| \leq \theta_{u\rightarrow v}^{t}, \quad \forall u, v = 1, \ldots, N, \forall t = 2, \ldots, T, \ (9f)$$

$$\left| \frac{d_{i}^{u\rightarrow v}}{b_{i}^{v}} - \frac{d_{i}^{(v-1)\rightarrow u}}{b_{i}^{v-1}} \right| \leq \phi_{u\rightarrow v}, \quad \forall u, v = 1, \ldots, N, \forall t = 1, \ldots, T, \ (9g)$$

$$\sum_{t=1}^{T} d_{i}^{u\rightarrow v} - \sum_{t=1}^{T} d_{i}^{v\rightarrow u} \leq \phi_{u\rightarrow v}, \quad \forall u, v = 1, \ldots, N, \forall t = 1, \ldots, T, \ (9h)$$

$$d \geq 0. \quad \ (9i)$$

3. The Maximum Entropy Approach

The classic ME approach taken by Van Zuylen and Willumsen (1980) was motivated by defining the most likely trip matrix as the one associated with the greatest number of microstates. Their method was applied to link count data, but can be easily adapted to the entry and exit data set through the following constrained nonlinear optimization problem:

$$\max_{d} \sum_{u, v, t} [-d_{i}^{u\rightarrow v} \log(d_{i}^{u\rightarrow v})] \ (10a)$$

s.t.

$$\sum_{u, v} d_{i}^{u\rightarrow v} = a_{i}^{u}, \quad \forall u = 1, \ldots, N, \forall t = 1, \ldots, T, \ (10b)$$

$$\sum_{u, v} d_{i}^{u\rightarrow v} = b_{i}^{v}, \quad \forall v = 1, \ldots, N, \forall t = 1, \ldots, T. \ (10c)$$

For simplicity, in Equations (10b) and (10c), we have assumed that the input data can be satisfied exactly.

Writing the Lagrangian relaxation for Equation (10), we obtain

$$L(d, \lambda, \mu) = \sum_{u, v, t} [-d_{i}^{u\rightarrow v} \log(d_{i}^{u\rightarrow v})] + \sum_{u, t} \lambda_{i}^{t} \left( \sum_{v=1}^{N} d_{i}^{u\rightarrow v} - a_{i}^{u} \right)
+ \sum_{v, t} \mu_{i}^{t} \left( \sum_{u=1}^{N} d_{i}^{u\rightarrow v} - b_{i}^{v} \right), \ (11)$$

and applying the Lagrange conditions, we find that the optimal solution satisfies the property

$$d_{i}^{u\rightarrow v} = \frac{a_{i}^{u} b_{i}^{v}}{\sum_{v'} b_{i}^{v'}}, \quad \forall u, v = 1, \ldots, N, \forall t = 1, \ldots, T. \ (12)$$

From Equation (12), we observe that the maximum entropy solution forces preferences for a particular destination $v$ at time $t$ (given by the ratio $b_{i}^{v}/\sum_{v'} b_{i}^{v'}$) to be identical across all origins, a strong condition that is not entirely credible. Our approach in Section 2 alleviates the proportional condition, constraining the proportions to be similar only between nearest neighbors instead of across all stations.

Another shortcoming of the maximum entropy approach is that if we consider each individual snapshot of the OD matrix in time, Equation (10) is entirely separable along the time dimension, and therefore
does not leverage temporal information in the estimation. Our approach is able to take advantage of temporal information, further constraining the OD matrix.

Having formally compared the smoothing assumptions behind our approach and the maximum entropy approach, we now compare the performance of the two approaches on real-world data in Section 4.

4. Computational Results

In this section, we present the results of our approach on three real-world data sets where the ground-truth origin–destination matrices are known, and one real-world data set where ground truth is not known. Entry and exit data are generated from the pickup and drop-off times in the data sets; these are then used as input to the formulation described in Section 2 to infer the original OD matrix.

Performance is evaluated using the metrics $R^2$ and conditional value at risk (CVaR), the former because of its prevalence in the wider statistics community, and the latter because of its usefulness in measuring extreme losses. Because $R^2$ will mostly reflect differences between the large OD matrix elements, we report CVaR for the percentage error to illustrate the distribution of percentage error. The metric CVaR is defined as

$$\text{CVaR}_\alpha = \mathbb{E}[X \mid X \geq \text{VaR}_\alpha],$$

where $X$ is the error metric of interest (percentage error), and $\text{VaR}_\alpha$ is defined as the top $\alpha$ quantile error, that is,

$$\text{VaR}_\alpha = \inf\{s \in \mathbb{R} \mid \Pr(X \geq s) \leq \alpha\}.$$  

For the results that follow, we report the average percentage error for the top 10% of errors, that is, $\alpha = 0.1$.

We also include measures of the various physical properties described in Section 2

Temporal irregularity:

$$\frac{1}{N^2 T} \sum_{u,v,t} (\delta_{t}^{u \rightarrow v} + \lambda_{t}^{u \rightarrow v});$$  

Spatial irregularity:

$$\frac{1}{N^2 T} \sum_{u,v,t} (\theta_{t}^{u \rightarrow v} + \rho_{t}^{u \rightarrow v});$$  

Flow imbalance:

$$\frac{\sum_{u,v,t} \phi_{t}^{u \rightarrow v}}{\sum_{u,v,t} \delta_{t}^{u \rightarrow v}}.$$  

For comparison, we include the maximum entropy solution obtained in closed form from the proportional formula in Equation (12) as a baseline. Since the closed-form solution assumes a perfect data regime where the entries and exit counts are an exact aggregation of the OD matrix, we apply the maximum entropy equation to the perfect data obtained through simple aggregation of the OD matrix.

The ME baseline and our method (QO) are tested on all data sets, with the results shown in Table 1. Additionally, the physical properties of the OD matrix estimates on the various data sets, including one where ground truth is not known, are summarized in Table 2. The physical properties of the actual OD matrices are included when possible.

All methods were implemented using the Julia language (Bezanson et al. 2017) and the optimization package JuMP (Lubin and Dunning 2015). Computational experiments were run on a laptop with an Intel i7-6500U processor and 16 GB RAM, with the exception of the larger-scale Hubway stations and New York City taxi experiments, which were run on a server with an Intel E5-2660 v4 processor and 256 GB RAM.

4.1. An Application with Bike-Sharing Data in Boston

The first data set was made publicly available by Hubway, a bike-sharing company located in the greater
Bertsimas and Yan: From Transportation Flows to Demand Estimations
Transportation Science, 2018, vol. 52, no. 4, pp. 1002–1011, © 2018 INFORMS

Figure 1. Goodness-of-Fit Comparison of Actual Demand (Black), Maximum Entropy Baseline (Blue), and QO (Red) on the Hubway Data Set

Notes. Origins are on the vertical grid axis, and destinations are on the horizontal grid axis. The x-axis and y-axis of each chart in the grid represent the hour of the day and actual demand, respectively. The y-axis varies between the panels. As a result, the OD matrices appear to violate the symmetry assumptions in Section 2.3. However, these OD matrices are actually quite symmetric; for example, the Cambridge–Somerville trip volume is 14,478 trips, which is balanced by a Somerville–Cambridge trip volume of 14,424 trips.

Boston area, through the Hubway Data Visualization Challenge. The data set contains records for 1,029,739 trips from October 1, 2012, through November 30, 2013, and includes bike pickup and drop-off times and stations for each trip. From this data set, we picked weekday trips between currently active stations in the Hubway network, which comprised 768,948 trips in total. To construct a smaller-scale example that could be easily visualized, the data set was grouped by hour and municipality (Boston, Cambridge, Somerville, and Brookline). Each municipality was taken as a “station” to produce 4 x 4 hourly OD matrices, for a total of 384 elements to be recovered. In this example, the QO was solved in six seconds.

The estimations using the maximum entropy baseline and the QO are plotted in Figure 1 for a visual comparison. While Figure 1 qualitatively shows that the maximum entropy method performs well on this data set, it tends to overstate demand to Boston because of the volume of trips ending in Boston. In comparison, the QO more accurately recovers the OD matrix elements: in particular, the overprediction of Cambridge–Boston and Somerville–Boston demand and underprediction of Cambridge–Cambridge and Somerville–Cambridge demand are corrected. This is reflected in the $R^2$ and CVaR values in Table 1 as well: the maximum entropy method attains $R^2 = 0.916$ and CVaR = 5.29, but the QO attains $R^2 = 0.991$ and CVaR = 3.14.

Examining the physical properties of the estimated OD matrices in Table 2, the QO shows an increase in temporal smoothness (4.5% irregularity in the baseline to 3.7% irregularity in our method) and a decrease in spatial smoothness (0.0% irregularity in the baseline to 0.5% irregularity in our method). The trade-off of spatial smoothness for temporal smoothness reflects the relaxation of the proportionality condition across all stations in Equation (12), holding only for neighboring stations in the QO (see Equation (5)). The flow imbalance is also reduced from 0.2% in the maximum entropy method to 0.1% with the QO; both of these remain close to the true imbalance of 0.3%.

We also tested both methods on the full Hubway station network, comprising 131 stations; this produced an OD matrix with 411,864 elements to be recovered. In this example, the QO was solved in 2.5 hours.
The maximum entropy method performed relatively poorly on the full Hubway network, achieving only $R^2 = 0.360$. This suggests that the main assumption behind the maximum entropy method, which is that preferences for a particular destination should be the same regardless of origin station, is a poor one for the large-scale network. Relating this condition in the QO and leveraging the other physical properties, we improve to $R^2 = 0.413$. However, this improvement in $R^2$ comes at a slight cost to CVaR, which rises from 0.97 to 1.12, which indicates that in this case, the QO does not estimate the smaller matrix elements as well.

Table 2 shows that the QO trades off increased spatial irregularity for temporal smoothing. The spatial irregularity increases from 0.0% to 0.6%, while the temporal irregularity decreases from 0.5% to 0.4%, a computational validation of the formal result in Section 3.

The lower performance of the QO on the stations-level data set relative to the municipalities-level data set can be attributed to the fact that the more granular data set is noisy. To show this, we test the validity of the physical assumptions by evaluating the following prediction methods:

- **Last hour**:
  \[
  \frac{d_{t}^{u,v}}{a_{t}^{v}} = \frac{d_{t-1}^{u,v}}{a_{t-1}^{v}}, \quad \forall u, v = 1, \ldots, N, \forall t = 2, \ldots, T; \tag{16a}
  \]

- **Nearest neighbor**:
  \[
  \frac{d_{t}^{u,v}}{a_{t}^{v}} = \frac{d_{1}^{s(u),v}}{a_{1}^{v}}, \quad \forall u, v = 1, \ldots, N, \forall t = 1, \ldots, T; \tag{16b}
  \]

- **Return flow**:
  \[
  \sum_{t} d_{t}^{u,v} = \sum_{t} d_{t}^{v,u}, \quad \forall u, v = 1, \ldots, N, \tag{16c}
  \]

corresponding to Equations (4a), (5a), and (7), respectively; that is, we evaluate how well elements of our OD matrix can be estimated using information from neighboring or similar elements, which were constrained to be similar in the QO.

Table 3 reports the $R^2$ for each prediction method in Equation (16). Unsurprisingly, the smoothing constraints do not hold as well at the granular level, with both the last-hour and nearest-neighbor predictions showing lower $R^2$ values for the station-level data set. The nearest-neighbor prediction also performs relatively poorly on both data sets. The low quality of the nearest-neighbor predictions is likely due to the large distance between each municipality and its nearest neighbor in the municipality-level data set, and the low volume and relative noisiness of the station-level data set.

However, despite the low quality of the nearest-neighbor prediction, removing the spatial smoothing constraints actually produced worse results. For the municipality-level OD matrix, $R^2$ stayed comparable but CVaR increased from 3.14 to 6.19, while for the station-level data set, CVaR was comparable but $R^2$ dropped from 0.413 to 0.382. These results indicate that such constraints are important to resolve the underdetermined OD matrix, even when the actual OD matrix is noisy.

### 4.2. An Application with Taxi Data in New York City

Our largest example is obtained from a public record of taxi trips made in New York City (NYC Taxi and Limousine Commission 2013). It includes the pickup and drop-off times and locations (by latitude and longitude) for each trip. We used the data for trips made on May 14, 2013, which comprised 474,708 trips in total. To construct “stations,” we grouped by pickup and drop-off districts (for example, Midtown or the Upper East Side). This produced $266 \times 266$ hourly OD matrices, for a total of 1,698,144 elements to be recovered: a truly large-scale application of our method. In this example, the QO was solved in eight hours.

The baseline method shows competitive performance with $R^2 = 0.805$ and CVaR = 1.24, and the QO improved on that with $R^2 = 0.870$ and CVaR = 1.03.

Table 2 shows a decrease in temporal irregularity, with spatial irregularity being kept at a low level and flow balance lowered. The OD matrices for such large-scale problems are generally sparse, and, in this case, only 2.2% of the elements of the original OD matrix are nonzero. Despite not having accounted for sparsity specifically in our formulation, our solution was also naturally sparse, with only 2.0% of the estimated OD matrix being nonzero.

In contrast to the Hubway station-level data set, the prediction methods of Equation (16) perform competitively, with $R^2$ values of 0.884, 0.803, and 0.992, respectively. This shows that the physical properties can hold well for large-scale examples, and, in this scenario, our QO is able to leverage the physical properties to produce an accurate estimation on hundreds of stations.

### 4.3. An Application with Subway Data in San Francisco

In our final example with ground-truth data, we use a publicly available data set from the subway system in San Francisco (Bay Area Rapid Transit 2016). The data set contains hourly origin–destination demand through 2016. We used the data from a high-volume

<table>
<thead>
<tr>
<th>Data set</th>
<th>Last hour</th>
<th>Nearest neighbor</th>
<th>Return flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hubway–Municipalities</td>
<td>0.977</td>
<td>0.173</td>
<td>1.000</td>
</tr>
<tr>
<td>Hubway–Stations</td>
<td>0.385</td>
<td>0.119</td>
<td>0.918</td>
</tr>
</tbody>
</table>

*Note.* The table shows the accuracy of estimating elements of the OD matrix from neighboring elements, detailed in Equation (16).
day on February 5, 2016, which comprised 523,802 trips between 46 stations. In this data set, which required recovery of a total of 50,784 matrix elements, the QO was solved in 80 seconds. Both the QO and maximum entropy methods were altered slightly to disallow self-looping at the stations.

The baseline method achieves $R^2 = 0.676$ and CVaR $= 1.02$, which the QO improves on with $R^2 = 0.764$ and CVaR $= 0.87$. Table 2 shows that temporal irregularity decreases, spatial irregularity increases, and flow imbalance decreases from the ME to the QO, reflecting the constraints of Equation (9). These results validate our method on real-world data from an actual subway network, which is exactly the type of system we envision will benefit from demand estimation using entry and exit count data.

### 4.4. An Application with Subway Data in Boston

In this section, we use our method to infer OD matrices for the Boston subway network, operated by the Massachusetts Bay Transportation Authority (MBTA). The Boston subway system comprises 113 stations on four main subway lines (Red, Green, Blue, and Orange). Since Boston uses a flat-fare policy, it does not collect swipe-out information, and data are collected only as aggregated turnstile entrances and exits. This makes the Boston subway case a prime example of the usefulness of our method.

A minute-level turnstile data set for the month of February 2014 was made available by Barry and Card (2014). It contains station entry and exit data for the period from February 1, 2014, to March 3, 2014, for 61 of the stations on the main subway lines. This excludes the Green Line surface stations that do not have turnstiles. The data set was filtered down to weekdays, resulting in 9,621,170 entries across the 61 stations (corresponding to an average daily ridership of approximately 458,151) and then grouped by hour. In this form, the turnstile data served as the input to infer a $61 \times 61$ hourly OD matrix, for a total of 89,304 matrix elements to be recovered. In this data set, the QO was solved in just under an hour.

In addition to the constraints described in Section 2 and applied to the Boston Hubway and New York City taxi data sets, we included grouped versions of relevant constraints to further constrain the QO. In particular, we expect the proportions of transfers between lines to remain relatively smooth over time. We also introduced constraints in the QO and slightly modified the baseline maximum entropy method to disallow self-looping demand at any station in the network. Liu and Zhou (2016) use capacity and passenger rationality constraints to model transit dynamics in a similar public transit setting, which is particularly useful when passengers have many viable route choices available to them. Although the Boston network is relatively simple and has few OD pairs where there are multiple reasonable route choices, our formulation could naturally be extended to incorporate such constraints for a more complex network.

Since ground truth is unavailable for this data set, we cannot evaluate our method’s performance as we did for the previous data sets; however, given the strong performance of the QO for earlier large-scale examples from a variety of transit modes and with up to hundreds of stations, we expect to see a similar improvement for this 61-station data set. Comparing the physical properties in the maximum entropy method to the QO in Table 2, we observe that the decreased temporal irregularity and the realistic increase in spatial irregularity follow the same pattern as before.

We can also compare the inferred OD matrices to survey data. Since the most recent survey data set is several years older than the turnstile data, it is inappropriate to incorporate it into the estimation models and allow it to influence the estimation. However, the general trends should remain relevant, even if the values themselves are out-of-date. Approximate estimations of transferring behavior are computed from survey data (Central Transportation Planning Staff 2008, Massachusetts Bay Transportation Authority 2010). The transfer matrix is plotted in Figure 2, with the origin line on the horizontal axis, the destination line on the vertical axis, and darker shades corresponding to higher demand. From Figure 2, it is clear that passengers tend to stay on a single line for the duration of their commute, since most of the demand is concentrated on the diagonal.

For comparison, the transfer matrices for the baseline-estimated and QO-estimated OD matrices are plotted in Figure 3 for the average weekday in the

![Figure 2. Daily Transfer Patterns on the Boston MBTA System for a Typical Weekday in July 2010](image)

Note. The demand was calculated from the 2010 Blue Book of Ridership and Service Statistics (containing total daily ridership by line) and the 2008–09 Central Transportation Planning Staff Survey (containing proportions of transfers by line).
February 2014 turnstile data set. The baseline method tends to allocate most demand to the Red Line, regardless of origin, since the Red Line shows the highest volume of exits. This is particularly inaccurate in the case of Blue–Red demand; two transfers are required to commute between these lines, and it is unlikely that many of the passengers on the Blue Line would be making two transfers. Meanwhile, the QO more accurately captures transferring behavior, concentrating demand largely on the diagonal. The main discrepancy with the QO is that it is unable to capture the high volume of Green–Green commutes. This is explained by the fact that most of the Green Line station data were not captured in the turnstile data set. If given these data, we would expect our method to more accurately capture the Green Line commutes.

5. Conclusion

To conclude, the collection of aggregate entrance and exit data provides a noninvasive and economical means for transportation planners to estimate OD demand. We expect that our estimations could take the place of outdated and less-detailed survey data used in transit planning models and in real-time demand estimation. Although aggregate data alone are not enough to fully specify the problem, we can recover unique, high-quality OD matrices by requiring that our estimations be physically realistic. We demonstrate that our method is tractable for large transit networks with tens and hundreds of stations, and show improved performance over a competitive baseline for real-world case studies in multiple cities.

Acknowledgments

The authors thank Sebastien Martin for his help with the New York City taxi data set. The authors thank the editor, associate editor, and reviewers for their helpful comments.

References


