A Stochastic Programming Method for OD Estimation Using LBSN Check-In Data

Qing-Long Lu[1], Moeid Qurashi[2], and Constantinos Antoniou[3]

[1]Research Associate, Technical University of Munich, Germany
[2]Research Associate, Technical University of Dresden, Germany
[3]Full Professor, Technical University of Munich, Germany

SHORT SUMMARY

Dynamic OD estimators based on traffic measurements inevitably encounter the indeterminateness problem on the posterior OD flows as such systems structurally have more unknowns than constraints. To resolve this problem and take advantage of the emerging urban mobility data, the paper proposes a dynamic OD estimator based on location-based social networking (LBSN) data, leveraging the two-stage stochastic programming framework, under the assumption that similar check-in patterns are generated by the same OD pattern. The search space of the OD flows will be limited by integrating a batch of realizations/scenarios of the second-stage problem state (i.e., check-in pattern) in the model. The two-stage stochastic programming model decomposes in a master problem and a set of subproblems (one per scenario) via the Benders decomposition algorithm, which will be tackled alternately. The preliminary results from experiments conducted with the Foursquare data of Tokyo, Japan, show that the proposed OD estimator can effectively recurrent the check-in patterns and result in a good posterior OD estimate.

Keywords: LBSN data, OD estimation, stochastic programming

1 Introduction

OD estimation/updating based on traffic measurements is realized with the aid of a prior OD matrix (for restricting the stochasticity of demand patterns) and a traffic assignment method (for mapping the OD matrix into traffic measurements). Link-level traffic counts data have been extensively used within such modeling frameworks benefiting from the convenience in collection and the capability of representing the aggregating contribution of OD flows to the link-based traffic underneath proper route selection procedures. However, as the ratio between the number of equations and the number of unknowns therein is always far smaller than one worsening with network size (OD disaggregation), the resulting OD flow estimates are generally confronted with the high indeterminateness issue (Cascetta et al., 2013; Antoniou et al., 2016; Qurashi et al., 2022). To this end, some have tried reducing the OD dimensions by adopting the quasi-dynamic framework established upon the assumption of constant OD shares across a reference period., e.g., (Cascetta et al., 2013; Bauer et al., 2017), or by employing dimension reduction techniques like principal component analysis, e.g., (Qurashi et al., 2019; Krishnakumari et al., 2020). Nevertheless, the availability of big urban data provides an unparalleled opportunity to address this problem from another direction, that is restricting the search space of the posterior OD flows by using or accommodating additional data sources relevant to urban mobility patterns. LBSN data is among the many alternatives that have shown the potential on this theme, attributed to its broad urban spatial and temporal coverage and confirmed trip purposes (Hu & Jin, 2017). Accordingly, some works, e.g., (Jin et al., 2014; Yang et al., 2015), exploiting along this research direction have been presented in recent years. However, most of them were dedicated to modeling static systems/networks, and all were built upon conventional gravity models, which cannot satisfy the accuracy requirement of many time-sensitive transport engineering applications. For instance, dynamic demand estimation is one of the critical inputs to the simulation models used for evaluating traffic management and policy measures (Antoniou et al., 2016).

To fill this gap, we aim to develop a within-day dynamic OD flow estimator driven by LBSN data, by leveraging the two-stage stochastic programming framework. Specifically, in the lower (second) stage, mobility flows among activity categories within each zone will be optimized. The inter-zone OD flows will then be estimated in the upper (first) stage, fulfilling the specific constraints on OD flows and the constraints imposed by lower-stage scenarios.

2 Methodology

We assume that similar check-in patterns at a specific time interval in different days during the reference period are generated by the same OD pattern. This assumption is plausible given the OD patterns will not change dramatically within a short period without any disruptive events. As such, we can apply the two-stage stochastic programming approach to alleviate the intractable high indeterminateness issue that challenges the OD estimators based on traffic measurements by integrating a group of check-in patterns extracted from the LBSN data.

Normally, LBSN data are generated when users post with geo-location information or "check-in" to a point-of-interest (POI) via mobile devices. Generally, The released LBSN data only tells the number of check-ins at POIs within specific time intervals. However, the changes in check-ins between successive intervals are capable of capturing some information on urban mobility patterns. Considering the stochasticity of human mobility, we tend to aggregate the POIs based on the category hierarchy adopted by social networking sites. Our approach essentially reconstructs the activity-based mobility flows while the OD matrix is just a derivative.

Specifically, as shown in Figure 1, we defined an activity node for one type of POIs. For each traffic analysis zone (TAZ), we define a virtual source and a virtual sink to: (i) "memorize" the sum of in- and out-flows; (ii) bridge the lower-stage and the upper-stage models.

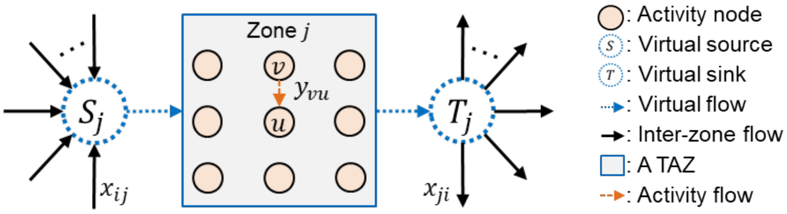


Figure 1: Graphical illustration of the model.

2.1 Lower-stage model

The lower-stage model focuses on zonal activity flows, i.e., optimizing the activity flows within TAZs. The following generic objective function is applied.

$$\min_{\{\mathbf{y}_z, \forall z \in \mathbb{Z}\}} \sum_{z \in \mathbb{Z}} f_2(\mathbf{y}_z, \mathbf{q}_z^{\tau-1}, \mathbf{q}_z^{\tau}) \tag{1}$$

where $\mathbf{y}_z$ is the vector of activity flows in TAZ z, $\mathbf{q}_z^{\tau-1}$ and $\mathbf{q}_z^{\tau}$ are the vector of check-ins at time interval $\tau - 1$ and τ, respectively, $f_2(\cdot)$ is the goodness-of-fit function, $\mathbb{Z}$ is the set of TAZs within the study area.

Since the number of check-ins cannot be negative, it is straightforward to have the first set of constraints (named inventory constraints), saying that the sum of leaving flows cannot be greater than the addition of the sum of coming flows and the number of check-ins recorded at the previous interval, which can be expressed as

$$\sum_u y_{vu,z} - \left(\sum_u y_{uv,z} + q_{v,z}^{\tau-1} \right) \leqslant 0 \qquad \forall v \in \mathbb{V}_z, \forall z \in \mathbb{Z} \tag{2}$$

where y_{vu} is the activity flow from activity node v to u.

The second set of constraints is on the virtual sources and sinks, named in- and out-flow balance constraints, for restricting the difference of the results from two stages with a plausible deviation bound considering the randomness and incompleteness of the activity information.

$$(1-\underline{\epsilon}_s)\sum_m x_{mz} \leqslant \sum_v y_{sv,z} \leqslant (1+\overline{\epsilon}_s)\sum_m x_{mz} \qquad \forall z \in \mathbb{Z} \tag{3}$$

$$(1-\underline{\epsilon}_t)\sum_m x_{zm} \leqslant \sum_v y_{vt,z} \leqslant (1+\overline{\epsilon}_t)\sum_m x_{zm} \qquad \forall z \in \mathbb{Z} \tag{4}$$

where $\{\underline{\epsilon}_s, \overline{\epsilon}_s, \underline{\epsilon}_t, \overline{\epsilon}_t\}$ are deviation parameters in the range (0,1), y_{sv} is the activity flow from the source to the activity node v, y_{vt} is the activity flow from the activity node v to the sink, x_{mz} is the inter-zone flow (the upper-stage decision variable) from TAZ m to z.

To avoid over-fitting, we also impose constraints on the flow distribution, named activity share constraints, for each activity node. These constraints integrate the prior knowledge about the activity chain into the model to restrict the search space of activity flows so as to make the inference more realistic.

$$(1-\underline{\epsilon}_a)\rho_{vu}q_v^{\tau-1} \leqslant y_{vu} \leqslant (1+\overline{\epsilon}_a)\rho_{vu}q_v^{\tau-1} \qquad \forall v,u \in \mathbb{V}_z, \forall z \in \mathbb{Z} \tag{5}$$

where $\underline{\epsilon}_a$ and $\overline{\epsilon}_a$ are predefined threshold parameters in the range (0,1).

The final constraints are non-negativity constraints on activity flows given as below.

$$y_{vu} \geqslant 0 \qquad \forall v,u \in \mathbb{V}_z, \forall z \in \mathbb{Z} \tag{6}$$

2.2 Upper-stage model

From substantial empirical data, we found that the sum of out-flows of a TAZ is linearly related to the number of check-ins, as shown in Figure 2. It follows that the objective function of the upper-stage model can be expressed as

$$\min_{\mathbf{x}} \quad f_1\left(\mathbf{x}, \mathbf{x}^{(p)}\right) + \kappa f_c\left(\mathbf{x}, \Phi\right) \tag{7}$$

where $\mathbf{x}$ is the vector of OD flows to be estimated, $\mathbf{x}^{(p)}$ is the given prior OD flows, Φ is the vector of the number of check-ins.

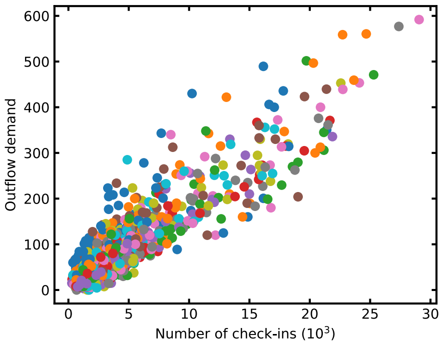


Figure 2: The linear relationship between the number of check-ins and outflows.

The constraints for the upper-stage model will be simply the bound constraints.

$$\underline{\epsilon}_b x_{ij}^{(p)} \leqslant x_{ij} \leqslant \overline{\epsilon}_b x_{ij}^{(p)} \qquad \forall i,j \in \mathbb{Z} \tag{8}$$

where $\underline{\epsilon}_b$ and $\overline{\epsilon}_b$ are threshold parameters.

2.3 Two-stage stochastic within-day dynamic OD estimator

As aforementioned, the two-stage stochastic programming framework provides a way to address the indeterminateness of the posterior OD flows by taking into account numerous lower-stage scenarios.

The proposed OD estimator can then be written as

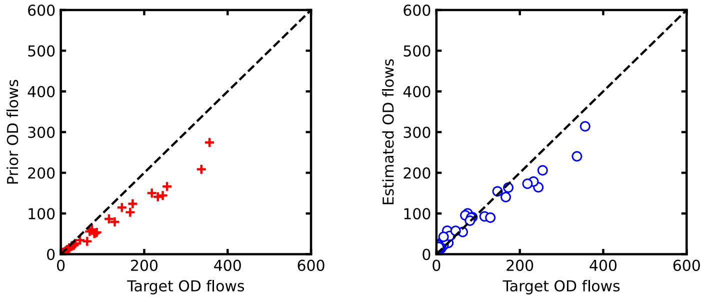
$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{y}} \quad & f_1\left(\mathbf{x}, \mathbf{x}^{(p)}\right) + \kappa f_c\left(\mathbf{x}, \Phi\right) + \omega \mathbb{E}_\xi \left[\sum_{z \in \mathbb{Z}} f_2\left(\mathbf{x}, \mathbf{y}_z, \mathbf{q}_z^{\tau-1}(\xi), \mathbf{q}_z^{\tau}(\xi)\right) \right] && && (9)\\
\text{s.t.} \quad & (1-\underline{\epsilon}_s) \sum_m x_{mz} \leqslant \sum_v y_{sv,z} \leqslant (1+\bar{\epsilon}_s) \sum_m x_{mz} && \forall z \in \mathbb{Z} && (10)\\
& (1-\underline{\epsilon}_t) \sum_m x_{zm} \leqslant \sum_v y_{vt,z} \leqslant (1+\bar{\epsilon}_t) \sum_m x_{zm} && \forall z \in \mathbb{Z} && (11)\\
& \text{Bound constraints on } \mathbf{x} && && (12)\\
& \text{Other lower-stage constraints} && && (13)
\end{aligned}
$$

where ξ is a random vector describing the problem state.

Apparently, it is a complicated non-convex model as the expectation $\mathbb{E}_\xi$ is usually an integral of a complex function. $\mathbf{x}$ is the complicating variable of the model which prevents the distributed solution. Accordingly, we apply the sample average approximation to approximate the expectation, and employ the Benders decomposition (BD) technique to decompose the problem into numerous subproblems (one per scenario) to accelerate the computation by parallel computing. If we apply the generalized least squares (GLS) estimator to $f_1(\cdot)$, $f_2(\cdot)$ and $f_c(\cdot)$, all subproblems will be convex and easy to solve. Due to space limitations, the solution algorithm is not given here.

3 RESULTS AND DISCUSSION

This section exhibits some preliminary results of implementing the GLS objective function. The Foursquare check-in data of Tokyo city, Japan, collected in May 2012 (about 90,000 check-in records), are used for validating the approach. The study area is divided into seven TAZs. The experiment was conducted for the interval from 9 am to 10 am. The result shows that the proposed model achieves a 30% improvement in Normalized Root Mean Square Errors (NRMSE) (from 51% to 36%) as shown in Figure 3. Furthermore, Figure 4 depicts that the model also leads to a good fit between the true (i.e., $\mathbf{q}^\tau - \mathbf{q}^{\tau-1}$) and the estimated check-in changes (i.e., $\sum_u \mathbf{y}_{vu} - \sum_u \mathbf{y}_{uv}, \forall v$). The experimental results and details are kept limited due to space limitations.



(a) Target vs. initial

(b) Target vs. estimated

Figure 3: Comparison of initial, target and estimated OD flows.

4 CONCLUSIONS

The indeterminateness problem of the posterior OD flows significantly impairs the reliability of traffic-measurements-based dynamic OD estimators. In contrast to reducing the dimension of OD flows as in most existing literature, this study tries to restrict the search space of the OD flows by adopting two-stage stochastic programming. This can be attained since the framework allows considering a batch of realizations of the second-stage problem state.

The model is specially devised based on the LBSN data characteristics. Yet, it is also applicable to the normalized crowdsourced datasets like the Google popular times data. In the present framework, check-in patterns play similar to traffic counts in previous OD estimators. The results

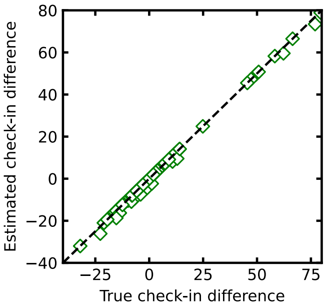


Figure 4: Comparison of true and estimated check-in changes.

show that the model can effectively replicate the check-in patterns. Moreover, it also results in a good posterior OD estimate with an improvement of NRMSE from 51% to 36%.

Further, considering the involvement of activity nodes/trip purposes, the model also has the potential to infer activity-based mobility patterns in urban areas. The exploration of this will be presented at MFTS 2022 conference. We will also include the sensitivity analysis on the threshold parameters, and examine and validate the model by implementing other estimators in the objective function other than GLS, such as the negative entropy. Integrating the model with appropriate sampling techniques for scenario generation, imperative in stochastic programming methods, will also be evaluated at the conference.

ACKNOWLEDGEMENTS

This work was supported by the European Interest Group CONCERT-Japan DARUMA project (Grant Number: 01DR21010) funded by the German Federal Ministry of Education and Research (BMBF).

References

Antoniou, C., Barceló, J., Breen, M., Bullejos, M., Casas, J., Cipriani, E., … van Lint, H. (2016). Towards a generic benchmarking platform for origin–destination flows estimation/updating algorithms: Design, demonstration and validation. *Transportation Research Part C: Emerging Technologies*, *66*, 79-98.

Bauer, D., Richter, G., Asamer, J., Heilmann, B., Lenz, G., & Kölbl, R. (2017). Quasi-dynamic estimation of od flows from traffic counts without prior od matrix. *IEEE Transactions on Intelligent Transportation Systems*, *19*(6), 2025–2034.

Cascetta, E., Papola, A., Marzano, V., Simonelli, F., & Vitiello, I. (2013). Quasi-dynamic estimation of o–d flows from traffic counts: Formulation, statistical validation and performance analysis on real data. *Transportation Research Part B: Methodological*, *55*, 171–187.

Hu, W., & Jin, P. J. (2017). An adaptive hawkes process formulation for estimating time-of-day zonal trip arrivals with location-based social networking check-in data. *Transportation Research Part C: Emerging Technologies*, *79*, 136–155.

Jin, P. J., Cebelak, M., Yang, F., Zhang, J., Walton, C. M., & Ran, B. (2014). Location-based social networking data: exploration into use of doubly constrained gravity model for origin–destination estimation. *Transportation Research Record*, *2430*(1), 72–82.

Krishnakumari, P., Van Lint, H., Djukic, T., & Cats, O. (2020). A data driven method for od matrix estimation. *Transportation Research Part C: Emerging Technologies*, *113*, 38–56.

Qurashi, M., Lu, Q.-L., Cantelmo, G., & Antoniou, C. (2022). Dynamic demand estimation on large scale networks using principal component analysis: The case of non-existent or irrelevant historical estimates. *Transportation Research Part C: Emerging Technologies*, *136*, 103504.

Qurashi, M., Ma, T., Chaniotakis, E., & Antoniou, C. (2019). Pc–spsa: employing dimensionality reduction to limit spsa search noise in dta model calibration. *IEEE Transactions on Intelligent Transportation Systems*, *21*(4), 1635–1645.

Yang, F., Jin, P. J., Cheng, Y., Zhang, J., & Ran, B. (2015). Origin-destination estimation for non-commuting trips using location-based social networking data. *International Journal of Sustainable Transportation*, *9*(8), 551–564.