Using Big Data Sources for Origin-Destination Matrix Adjustment

Mahboobeh Moghaddam[1], Mahmoud Mesbah[2], Mark Hickman[2]

[1]School of Economics, The University of Queensland

[2]School of Civil Engineering, The University of Queensland

Email for correspondence: m.moghaddam@uq.edu.au

Abstract

Origin-Destination (OD) matrix adjustment is a well-known approach to generate up-to-date matrices. It is an efficient method to avoid conducting unnecessary travel surveys. The main practice in OD matrix adjustment is to use a base year matrix as the seed matrix and adjust it for subsequent years using up-to-date traffic information such as links' traffic counts. The advent of big data has presented new opportunities for less costly and more effective origin-destination adjustment methods. However, in the extant literature the use of real traffic counts in the research studies on OD matrix adjustment has been very limited. In this paper, we present our effort to exploit two of the transport big data sources in Queensland, known as STREAMS and SCATS, to adjust the origin-destination matrices developed for the South-East Queensland Strategic Model (SEQSTM). Our experiments show that, just as improving the quality of the data can improve the quality of the adjusted matrix, increasing the quantity of the data can significantly enhance the quality of the matrix as well.

Keywords. Origin Destination estimation, Big data, SEQSTM, STREAMS, SCATS

1 Introduction

1.1 Motivation

The Origin-Destination (OD) matrix is the source of traffic demand information for transport planning as well as control and management of transport systems (Doblas and Benitez 2005). This matrix is usually created based on data collected from large scale travel surveys such as home interviews, roadside interviews and licence plate methods (Bera and Rao 2011). However, the generated OD matrix can become obsolete in a relatively short time due to various events and the frequent changes in travel patterns, which then requires the matrix to be re-estimated (Bera and Rao 2011). This makes the use of travel surveys relatively expensive and therefore not practical to be repeated frequently even in most developed cities (Toole et al. 2015). Therefore, in practice, after an original OD matrix estimation, this matrix is usually *adjusted* for subsequent time intervals.

One main approach to OD matrix adjustment is based on updating an existing matrix using current traffic information. In the era of big data, the availability of huge amount of traffic data collected from sensors such as loop detectors presents new opportunities for less costly and more effective OD matrix adjustment approaches. However, we are led to believe that the real traffic counts have been used in research studies only at a small scale. To the best of our knowledge, the effect of the quantity of the data (more precisely, the number of traffic counts) on the outcome of the OD matrix adjustment methods (i.e. the quality of the adjusted matrix) has not been investigated before.

To study the impact of employing big data for OD matrix adjustment, we performed OD matrix adjustment with two big data sources in South-East Queensland (SEQ). Known as STREAMS and SCATS, these data sources provide an almost-live feed of traffic information. These data sources provided us with counts on more than 6,800 links from across SEQ.

We performed the OD matrix adjustment using a well-known transportation planning software, called EMME, with an iterative approach through several steps. Each step provided us the information for further improving the results by improving the quantity or the quality of the traffic counts required for the matrix adjustment. Our experiments show that improving the quality of the data can improve the quality of the adjusted matrix, which is not surprising. However, our results also demonstrate that improving the quantity of the traffic count data can significantly enhance the quality of the adjusted matrix too, a proposition that, to the best of our knowledge, has never been investigated before.

1.2 Related Work

There is a great deal of research on OD matrix adjustment methods, considering the fact that one of the early studies in this domain goes back to 40 years ago (Bell 1983). As the focus of our research is on employing big data sources for OD matrix adjustment, in this section we only discuss the literature in terms of the data that is used in evaluation of matrix adjustment proposals. We refer the interested reader to (Bera and Rao 2011) for a more detail review of OD matrix adjustment methods.

Many researchers have acknowledged the opportunity presented by the existence of new big data sources for OD matrix adjustment, including El Faouzi (2011). However, despite the reliance of the OD matrix adjustment methods on traffic counts, not many of the published studies have used real traffic count data, collected from real-world road networks. Generally, the evaluations are based on simulated traffic counts and artificial networks, such as in (Walpen et al. 2015; García-Ródenas and Verastegui-Rayo 2008). Even in cases considering a real-world network, the traffic data is often generated through simulation, as in (Gómez et al. 2015; Cantelmo et al. 2015; Barceló et al. 2013; Toledo and Kolechkina 2013; Cipriani et al. 2011).

We are led to believe that the real traffic counts have been used in research studies only at a small scale, e.g. 65 traffic counters in (Doblas and Benitez 2005), 18 link flows in (Russo and Vitetta 2011), and 30 sensors to provide link counts in (Lu et al. 2013). One of the largest set of traffic counts has been used in (Pinjari et al. 2011) with 665 counting stations.

In our study, despite the availability of big data sources of traffic counts, we initially performed OD matrix adjustment with just 200 links. In section 3, we have elaborated on some of the challenges that prevented us from more extensively utilizing the available data sources. By gradually addressing some of these challenges, we could increase the number of links with traffic counts to more than 6,800 in the final round of experimentation. The results of our experiments clearly demonstrate the impact of the quality and quantity of the data used in OD matrix adjustment on the quality of the adjusted matrix.

2 Data

In this study, we have relied on data from three main sources as follows and explained next:

1. SEQ Strategic Transport Model (SEQSTM)
2. STREAMS: SEQ Traffic Flow, excluding Brisbane's Central Business District (CBD)
3. SCATS: Brisbane's CBD traffic counts

2.1 SEQSTM

We used the demand matrix (morning peak) from the South East Queensland Strategic Transport Model (SEQSTM) as the initial seed matrix to be adjusted. This model is a traditional long-term, four-step transport planning model developed and maintained by Queensland's Department of Transport and Main Roads (TMR) based on a household travel survey conducted in 2006 for 3 days and recalibrated in 2011. In the initial survey, all trips made by all household members were recorded for a randomly selected weekday, along with household socio-economic information. The model is used to forecast road network conditions until 2031 (DTMR 2015).

For our purposes, the SEQSTM provides not only the matrix for the intended analysis, but also delivers other parameters required for the OD matrix adjustment method, such as penalty functions for turns and the volume delay functions (VDF) for links. It also provides a fully connected road network of the SEQ at the level of motorways, arterials, and major collectors, which is required for traffic assignment iterations of the OD matrix adjustment process.

2.2 STREAMS

The second source of data is a system called STREAMS[1]. This system collects traffic data, including traffic volumes, occupancy, and speed, from more than 21,000 links and 10,000 intersections across SEQ, recorded and updated in 3-minute intervals. The data have been collected since 2009 at the University of Queensland under an agreement with the Queensland Department of Transport and Main Roads. However, STREAMS only collects data on the state-controlled highways, meaning that locally-controlled roads are not included in this data set. Our STREAMS data are taken from the state-controlled traffic network as observed on 20 March 2013.

2.3 SCATS

The third source of data is SCATS, the Sydney Coordinated Adaptive Traffic System, which is used by Brisbane City Council to manage traffic on highways and streets. Currently, Brisbane City Council manages more than 6800 km of roadways and over 50,000 intersections, of which more than 800 intersections are signalised. The SCATS system records and updates traffic volumes at 5-minute intervals, for individual detectors operating at the stop line of each intersection approach and at a significant number of upstream (mid-block) locations. These data have been provided to the University of Queensland under an agreement with the Brisbane City Council. Ultimately, these data complement those covered by STREAMS. In our experiments, we used the SCATS traffic data on 20 March 2013 to be consistent with the STREAMS data.

3 Challenges

We faced a number of challenges in exploiting the big data sources, STREAMS and SCATS, as discussed below.

3.1 Volume of Data

The volume of the data makes traditional methods for accessing and analysing the data inefficient. To cut the complexity, we limited our analysis to one day (20 March 2013, a typical weekday, neither school holiday nor a public holiday). However, considering the fact that there are more than 21,000 links in SEQSTM with traffic information that is collected and stored every 3 minutes, the size of the data for one day is still quite challenging. Therefore, for data preparation processes such as reading the data, cleaning it, and combining multiple sources, efficient programming practices are required.

[1] < https://www.transmax.com.au/cms/streams-intelligent-transport-system >

3.2 Incomplete Data

Although STREAMS covers more than 21,000 links across SEQ, many of these links do not report the traffic information. An important area missing from STREAMS is the Brisbane CBD. As one of the major sources of trip generation and distribution in Brisbane, lack of traffic information from the CBD can seriously damage the quality and trustworthiness of the adjusted OD matrix. As a result, we needed to combine traffic information from two major sources, STREAMS and SCATS, to cover the important links in SEQ as much as possible.

3.3 Lack of Interoperability

In OD matrix adjustment, if the traffic counts are collected from the same road network representation as the one from the seed matrix, assignment of counts to the seed matrix is trivial. In our case, however, traffic counts and seed matrix were based on two different road representations of SEQ. As a result, the assignment of traffic counts from STREAMS to the links in the SEQSTM network was not a trivial job. The reasons are that:

(1) the representation of links in STREAMS is not an exact match to the representation of links in SEQSTM network,
(2) the two networks have different levels of detail,
(3) the representations of intersections, roundabouts, and on- or off-ramp links differ in the two networks. SEQSTM generally offers a more detailed and closer-to-reality representation of the area compared to STREAMS.

Therefore, we needed to match the network of STREAMS with the underlying network of SEQSTM. We had a similar interoperability problem between the road network representation of SCATS and that of the SEQSTM.

As the existing approaches to matching road networks are usually very complex and not easy to implement, we decided to develop our own matching algorithm. Accordingly, we developed a heuristic algorithm and implemented it using ArcGIS and Python.

We evaluated the results of the matching for STREAMS and SEQSTM by visually inspecting the output matches for 13 suburbs in the study area, covering more than 5% of the total number of links in the SEQSTM network. Our evaluation results show that the overall precision of the algorithm is around 89%. The overall precision is defined in terms of: (1) finding an accurate match where a match exists, and (2) not assigning a match when there is no match (Moghaddam et al. 2017).

4 Experiment

To evaluate whether the quantity of the data has any impact on the quality of the adjusted matrix, we performed five rounds of experiments in an iterative way. In each round, we included more traffic counts and repeated the matrix adjustment procedure. The results are summarized in Table 1.

The traffic counts are extracted from two data sources: STREAMS and SCATS. Based on the level of the matching available in that round, we assigned the traffic counts to the SEQSTM network. SEQSTM also provides us with the seed matrix.

We performed OD matrix adjustment using the “traffic demand adjustment” function in EMME 4.2.2. This function is an implementation of the gradient method (Heinz Spiess 1990). In the EMME procedure, it is possible to define an input weight parameter, $0 \leq \alpha \leq 1$, that determines how much to deviate from the seed matrix and fit to the counts. If the procedure aims to minimize the difference between the observed traffic counts and the flows resulting from the traffic assignment, $\alpha = 1$. On the contrary, if the objective is to minimize the difference

between the seed matrix and the adjusted matrix, $\alpha = 0$. Consequently, this procedure reports two different R^2: flow (or link) R^2, and demand (or OD) R^2. The flow R^2 represents the correlation between observed and model flows; in other words, how close the assigned flows are to the actual, observed counts. The demand R^2 represents the correlation between the adjusted matrix and the seed matrix.

We experimented with different values of α . At the end, we decided to keep it equal to 1 to control for this parameter across different rounds of the experiment. As the value of α is set to be 1, we relied on the flow R^2 as the evaluation metric for the quality of the adjusted matrix. Nevertheless, we have reported the demand R^2 in Table 1 as well.

For data preparation, we developed a java program that can efficiently handle the data of one day. Data preparation includes reading traffic counts from STREAMS and SCATS, reading the SEQSTM network layout, assigning the traffic counts to the SEQSTM links based on the matching of the network representations, and writing this information in an acceptable format for EMME. Another program is responsible for reading the network layout of STREAMS and SCATS (in csv and json format) and for generating an acceptable input format for ArcGIS (kml) to execute the matching algorithm.

4.1 Round 1

The first round was performed with limited traffic counts, only from STREAMS. The matching of STREAMS network and SEQSTM network was performed using two simple rules, and therefore, only 200 links could be assigned with a traffic count value. We let the demand adjustment function go through 30 iterations; however, the best R^2 achieved was 0.077.

4.2 Round 2

To increase the number of traffic counts in the adjustment process, we tried to improve the matching of the SEQSTM network and STREAMS network in this round. STREAMS provides us with the lat/long coordinates of some detectors. However, these detectors are not directly associated to the links, even within the STREAMS data set. Therefore, we matched the detectors to both the STREAMS and SEQSTM links using their coordinates. Based on the results of these matchings, we identified the STREAMS and SEQSTM links that were matched to the same detector.

We tried different thresholds for the matching and performed visual inspection to evaluate the accuracy of the matching outcome. At the end, we decided to apply a strict 1m threshold for the matching of SEQSTM links and the detectors, and a more liberate threshold of 10m for the matching of STREAMS links and the detectors.

There are 13,639 detectors in the STREAMS network. However, only 3,133 of them have lat/long coordinates. After performing these matching processes and extracting the links with traffic count information from STREAMS, we managed to increase the number of traffic counts in the adjustment process to 1960. As depicted in Table 1, this increase in the size of traffic counts significantly improved the R^2, from 0.077 to 0.483.

4.3 Round 3

At this stage, we developed a heuristic algorithm for matching the SEQSTM and STREAMS networks. The proposed algorithm can be broken down into three major stages: pre-processing, intersection, and post-processing. The pre-processing stage splits the road features in each network such that road features represented with similar shapes in the two networks will also have similar lengths. After the road segments are split, a function calculates the direction of each road feature. Next, an intersection function, which evaluates the spatial “intersection” of two sets of polyline features, is used to match the split road features from the two networks. Splitting the road features improves the accuracy of the matches found by the intersection function. Finally, the post-processing procedure removes incorrectly matched

road features such as the links that are identified as matches but are oriented in opposite directions.

The more comprehensive matching algorithm increased the number of links with traffic counts to 5132. Compared to the previous round, this increase led to 27% improvement in R^2, reaching 0.661 (Moghaddam et al. 2017).

4.4 Round 4

In this round, we improved the quality of the input data. Our investigation of the STREAMS data revealed that there are links with very high flow that do not seem reasonable. Therefore, we developed a filtering mechanism to exclude the links with unreasonable flows. The threshold for exclusion is based on the capacity and the number of lanes of each link, adopted from SEQSTM strategic model. After discussing with experts from TMR, we decided to remove the links which had a flow greater than twice the total roadway capacity, as defined within the SEQSTM network. The improvement in the quality of the data increased the R^2 from 0.661 to 0.682.

4.5 Round 5

In this final round, we increased the number of links with traffic counts by using the SCATS data set. We used our heuristic algorithm (as in Rounds 2 and 3) to match the SCATS network to the SEQSTM network. This increased the number of counts to 6,840, which subsequently led to $R^2 = 0.723$.

5 Results

The results are summarized in Table 1 below. The R^2 of the estimation process clearly shows that not just the quality of the input data can affect the result of the OD matrix adjustment, but also the quantity of the data matters too; more traffic counts have improved the R^2 from 0.077 to 0.723.

Table 1. The result of performing OD matrix adjustment

Round	Data Description	# Iterations (where best R^2 achieved)	# Links (with traffic counts)	Best R^2 (flow)	Best R^2 (demand)
1	+STREAMS 2013: flow data from the initial limited matching +SEQSTM 2014	22	200	0.077	NA*
2	+STREAMS 2013: flow data from matching through detector sites +SEQSTM 2014	30	1960	0.483	0.690
3	+STREAMS 2013: flow data from full matching +SEQSTM 2014	20	5132	0.661	0.776
4	+STREAMS 2013: threshold applied to filter flows (full matching) +SEQSTM 2014	20	5208	0.682	0.887
5	+STREAMS 2013: threshold applied to filter flows (full matching) +SCATS 2013 traffic counts +SEQSTM 2014	20	6840	0.723	0.848

*Not available as it was not reported in the previous version of EMME.

6 Conclusion and Future Work

In this paper, we studied how an increase in the number of traffic counts can affect the result of OD matrix adjustment. We designed and performed a set of experiments, improving the quality and quantity of the traffic counts gradually. We used the flow R^2 (how close the observed flows are to the ones resulting from the traffic assignment based on the adjusted matrix). The results show that not just improving the quality, but also improving the quantity of the data can lead to a significant improvement in the adjusted matrix. This is an important observation, which we hope draws the attention of the research community to the importance of the volume of the traffic data on the quality of the adjusted matrix.

For future work, we intent to focus on the problem of determining the optimal (minimum) set of traffic counts and their locations for OD matrix estimation. The main question would be if we have covered this optimal set, what would be the impact of growing beyond this set on the quality of the adjusted matrix? Would we still observe significant improvement? Another potential direction for future research is to extend the data set to more than one day and estimate an OD matrix.

7 Acknowledgment

The authors would like to acknowledge Queensland's Department of Transport and Main Roads, and Brisbane City Council for providing the data used in this study. The authors are grateful to Frans J Dekker, David C Gyles, Veit-Rudolf Roth, Scott Cormack, Mehdi Z Taghavi, and Jason N Kruger for providing the data and clarifying data issues.

8 References

Barceló, J., Montero, L., Bullejos, M., Linares, M. & Serch, O., 2013. Robustness and Computational Efficiency of Kalman Filter Estimator of Time-Dependent Origin-Destination Matrices. *Transportation Research Record: Journal of the Transportation Research Board*, 2344, pp.31–39.

Bell, M.G.H., 1983. The Estimation of an Origin-Destination Matrix from Traffic Counts. *Transportation Science*, 17(2), pp.198–217.

Bera, S. & Rao, K. V, 2011. Estimation of origin-destination matrix from traffic counts: the state of the art. *European Transport / Trasporti Europei*, 49, pp.2–23.

Cantelmo, G., Viti, F., Cipriani, E. & Marialisa, N., 2015. A Two-Steps Dynamic Demand Estimation Approach Sequentially Adjusting Generations and Distributions. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, pp. 1477–1482.

Cipriani, E., Florian, M., Mahut, M. & Nigro, M., 2011. A gradient approximation approach for adjusting temporal origin–destination matrices. *Transportation Research Part C: Emerging Technologies*, 19(2), pp.270–282.

Doblas, J. & Benitez, F.G., 2005. An approach to estimating and updating origin–destination matrices based upon traffic counts preserving the prior structure of a survey matrix. *Transportation Research Part B: Methodological*, 39(7), pp.565–591.

DTMR, 2015. Southeast Queensland Strategic Transport Model (SEQSTM), Multimodal, Version 1.

El Faouzi, N.-E., 2011. Data fusion in intelligent transportation systems: Progress and

challenges – A survey. *Information Fusion*, 12(1), pp.4–10.

García-Ródenas, R. & Verastegui-Rayo, D., 2008. A column generation algorithm for the estimation of origin–destination matrices in congested traffic networks. *European Journal of Operational Research*, 184(3), pp.860–878.

Gómez, P., Menéndez, M. & Mérida-Casermeiro, E., 2015. Evaluation of trade-offs between two data sources for the accurate estimation of origin–destination matrices. *Transportmetrica B: Transport Dynamics*, 3(3), pp.222–245.

Heinz Spiess, 1990. A Gradient Approach for the O-D Matrix Adjustment Problem.

Lu, C.-C., Zhou, X. & Zhang, K., 2013. Dynamic origin–destination demand flow estimation under congested traffic conditions. *Transportation Research Part C: Emerging Technologies*, 34, pp.16–37.

Moghaddam, M., Bertolaccini, K., Hickman, M. & Mesbah, M., 2017. A Heuristic Network Matching Algorithm to Address the Interoperability of Transport Network Representations. *[Submitted to] Computer-Aided Civil and Infrastructure Engineering*.

Pinjari, A.R., Pendyala, R.M., Bhat, C.R. & Waddell, P.A., 2011. Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation*, 38(6), pp.933–958.

Russo, F. & Vitetta, A., 2011. Reverse assignment: calibrating link cost functions and updating demand from traffic counts and time measurements. *Inverse Problems in Science and Engineering*, 19(7), pp.921–950.

Toledo, T. & Kolechkina, T., 2013. Estimation of Dynamic Origin–Destination Matrices Using Linear Assignment Matrix Approximations. *IEEE Transactions on Intelligent Transportation Systems*, 14(2), pp.618–626.

Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A. & González, M.C., 2015. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58, pp.162–177.

Walpen, J., Mancinelli, E.M. & Lotito, P.A., 2015. A heuristic for the OD matrix adjustment problem in a congested transport network. *European Journal of Operational Research*, 242(3), pp.807–819.