A scalable and flexible single-level framework for O-D matrix inference using IoT data
Wei Sun¹, Akshay Vij¹, Nicholas Kaliszewski²
¹University of South Australia
²Department for Infrastructure and Transport
Email for correspondence: simon.sun@unisa.edu.au

Abstract
This study proposes a scalable and flexible single-level framework for Origin-Destination matrix (ODM) inference using data from IoT (Internet of Things) and other sources. The framework allows the analyst to integrate information from multiple data sources, while controlling for differences in data quality across sources. We assess the effectiveness of the framework through a real-world experiment in Greater Adelaide (GA), Australia. We infer car OD flows within the region using three separate data sources: site-level traffic counts from loop detectors, vehicle trajectories recorded by roadside Bluetooth sensors, and journey-to-work data collected by the Australian Census. We compare our OD inferences with those from the current version of the Metropolitan Adelaide Strategic Transport Model (MASTEM), calibrated using data from traditional household travel surveys. We find remarkable consistency between our inferences and those from MASTEM, despite differences in input data and methodologies. For example, for the morning peak period, we predict the total number of trips made within GA to be around 556 thousands, while the corresponding prediction from MASTEM is about 484 thousands. The two predictions are within 15 per cent of each other. When we compare the spatial distribution of trips, in terms of origins and destinations, we find that our inferred OD matrix has an 86 per cent cosine similarity to the corresponding MASTEM matrix. In summary, our results show that the proposed framework can produce highly comparable ODM, trip production and trip attraction patterns to those inferred from traditional household travel survey-based transportation demand modelling methods.

1. Introduction
In transport engineering and planning, the Origin-Destination matrix (ODM) refers to the distribution of trips in terms of their origins and destinations between zones within a planning area. The ODM provides insights on travel patterns within the planning area, and is therefore one of the key outputs of the strategic transport modelling and forecasting process. Traditionally, ODMs produced from comprehensive models such as MASTEM have been relied on data collected through household travel surveys (HTS) that ask participating household members about their travel patterns over a 1 or 2-day observation period. However, HTSs are expensive. Hartgen and San Jose (2009) report average costs of $487,000 per HTS, and roughly $150 per response, though they note that "many surveys cost considerably more than the average, and the spread of the data is substantial". Stopher et al., (2011) find that a computer-assisted telephone interview survey in Australia would cost $150-200 per household, face-to-face surveys are likely in the order of $350 plus per household, and a 15-day GPS survey would cost around $300 per household. Furthermore, HTSs are typically cross-sectional, only providing a static snapshot, and most urban areas around the world conduct their HTS only once every 5-10 years. Older surveys can lose their representativeness, especially if
there are structural shifts in travel patterns. For example, the recent COVID-19 pandemic and accompanying public health concerns are likely to have long-term impacts on trip-making and mode-use patterns that would not be captured by surveys conducted before the pandemic. Similarly, the emergence of new modes of transport, such as electrical scooters recently, and ridesharing in the recent past, have had similar if smaller impacts on travel patterns that may not be fully captured by older surveys.

The rapid diffusion of smartphones, Wi-Fi and Bluetooth networks, and the digitization of transport planning, booking and payment systems, in conjunction with broader advances on the Internet of Things (IoT) across all sectors of the economy, imply that we have more data than ever before on how people use transport infrastructure, and how these patterns are likely to change in the future. These new ICT technologies offer a more cost-effective alternative for the collection of transport data, but at far greater volumes. For example, the Sydney Household Travel Survey currently samples roughly 5,000 households each year from a population of roughly 5 million, and observes their travel patterns over a 24-hour period. In contrast, data from IoT sources could offer a continuous stream of travel information for a large majority of that population over rolling time periods.

As a result, in recent decades, numerous studies have developed methods for the inference of ODMs based on data collected from roadside loop detectors, Bluetooth detectors, license plate scanners, GPS-connected devices and other IoT sources. Spiess (1990) proposed an approach to estimate the ODM using observed link flows, and his study has paved the way for subsequent work in this area. Apart from traffic counts data, with increased data availability, subsequent studies have utilized other IoT sources, such as mobile phone data, vehicle turning data and Bluetooth data, to estimate the ODM. These data sources have either been used independently to infer the ODM (Carpenter et al., 2012, Barceló et al., 2013, Alexander et al., 2015), or used jointly with traffic counts to improve inference accuracy (Iqbal et al., 2014, Behara et al., 2021a, Behara et al., 2021b, Michau et al., 2019, Cipriani et al., 2021).

A review of the literature suggests there are two key limitations to these previous studies that the present study aims to overcome. First, existing studies estimate the ODM based on data from only one or two IoT sources. It is not clear how information from new sources can be integrated within their frameworks, and if and how the analyst can control for differences in quality across datasets. We develop a flexible framework that allows the analyst to integrate multiple data sources for ODM inference, and to ascribe weights to each dataset as a measure of their quality to be used accordingly within the ODM inference algorithm. For example, in our real-world experiment, we use information from roadside loop detectors, Bluetooth sensors, and commute mode choice data collected by the national Census to infer ODMs in Adelaide, Australia. During inference, we give the greatest weight to roadside loop detector data and the lowest weight to Bluetooth sensor data, to capture known differences in observation quality.

Second, many studies have followed a bi-level framework that uses an equilibrium or optimal model to perform path assignment, i.e., to translate ODMs into equivalent link counts. More recently, studies such as Dey et al. (2020) and Behara et al. (2021b) propose to replace the traditionally equilibrium-based path choice assignment with inferred or pre-determined path choice functions based on historic data. By using the inferred path choice function from the observed data, Behara et al. (2021b) simplify the traditional bi-level framework to a single-level, which is much more computationally efficient, especially when applying to a large road network. Therefore, instead of testing ODM inference quality on a relatively small part of the city/region with only a few hundred OD pairs (Cipriani et al., 2021, Behara et al., 2021a, Ma and Qian, 2018, Cantelmo et al., 2014), the framework can be easily deployed to a city-wide
scale (Behara et al., 2021b). For example, our real-world experiment covers the Adelaide metropolitan region in Australia, comprising roughly 1.3 million residents living and working across 3,260 square-km, and contains ten thousand OD pairs. By adopting a single-level framework, we can scale our algorithms to infer ODMs for the entire region.

In summary, we propose a flexible and scalable algorithm for OD inference, which can leverage travel information from multiple sources with varying degrees of quality, and which can be applied to large planning areas at a network-wide and area-wide level. The framework offers a viable alternative to traditional survey-based methods of ODM inference.

The reminder of this paper is structured as follows. Section 2 presents the proposed framework for the ODM inference. Section 3 presents the settings for the case study using the proposed inference framework based in the Adelaide region, South Australia. Section 4 describes the results of the case study and brings discussions. Section 5 summarises the conclusions.

2. The proposed method and its instance in the real-world case study

The proposed method is developed from several existing studies (Michau et al., 2019, Behara et al., 2021b, Spiess, 1990, Cipriani et al., 2021), and the following presents the formulation in its generalised form.

\[ \text{Find } \hat{T} \text{ that minimize: } f(w_q D_q(\bar{q}, q), w_1 D_1(\bar{T}_1, T_1), w_2 D_2(\bar{T}_2, T_2), \ldots, w_k D_k(\bar{T}_k, T_k)) \]

Subject to: \( q = F(T) \)

In the generalised target function above:

- As suggested by existing studies (Michau et al., 2019, Cipriani et al., 2021), the format of the above target formulation is usually structured as a weighted sum of different components.
- \( \bar{q} \) is an input into the formula. It denotes the observed SCATS traffic counts. The traffic count data provide crucial information for estimating OD matrices as it contains the information from this traffic counts data could be the highly granular link-level data if it is available in the study zone (e.g., in the Melbourne CBD area as introduced in Dey et al. (2020)). However, these granular link-level traffic counts may not be available in some cities or regions. Therefore, the site-level traffic count which has a lower granular level could be used instead and achieves reasonably good performance, as will be demonstrated in the real-world experiment section.
- \( q \) is an output from the optimization and it denotes the estimated traffic counts.
- \( \bar{T}_1 \) to \( \bar{T}_k \) are the prior knowledge of the OD matrix from multiple data sources. The prior knowledge could be an OD matrix learnt from a specific data source, such as the estimated OD matrix estimated from the observed Bluetooth data (Carpenter et al., 2012, Michau et al., 2014, Michau et al., 2019, Cipriani et al., 2021), GPS-based travel data (Ge and Fukuda, 2016), mobile data (Montero et al., 2019, Alexander et al., 2015, Caceres et al., 2007), or from a license plate recognition system (Rostami Nasab and Shafahi, 2019). It also can be an estimated complete OD matrix for a specific purpose from some reliable data source, such as the home-work commute by car ODM obtained from the ABS census database (as will be used in the following real-world experiment).
- It is worth noting that the prior knowledge of the OD matrix does not need to be complete, instead, as discussed in some studies (Sherali et al., 2003, Liou and Hu, 2010, Behara and Bhaskar, 2021), even a prior knowledge of the OD matrix could help to improve the quality of estimated OD matrix.
• $T_1$ to $T_k$ denote the estimated OD matrix that has the same scope that corresponds $\tilde{T}_1$ to $\tilde{T}_k$. They are outputs from the optimization process.

• $D_q$ denotes the distance between the observed SCATS traffic counts $\tilde{q}$ and the estimated counts $\tilde{q}$. $D_1$ to $D_k$ denote the distance between each of the prior knowledge of the OD matrix and their estimated equivalent, respectively. It is worth noting that each distance can be measured with its own distance measurement. As introduced in Michau et al. (2019), some typical distance measurements include Least Squares, Entropy maximization, Information minimization and Maximum Likelihood. The modeller should choose appropriate distance measurement to facilitate the optimization process integrating information from all inputs.

• $w_q$ and $w_1$ to $w_k$ denote the assigned weightings to each component of the distance measurements. These weightings are assigned by the modeller, and they can reflect the relative belief or importance of the data (Michau et al., 2019).

Lastly, the $q = F(T)$ is an input and it indicates that the estimated SCATS site counts follow a path choice function which is determined by $T$. In the bi-level estimation method, the traffic assignment usually follows an equilibrium process (Spiess, 1990, Cipriani et al., 2021), not only it is computationally slow because it is iterative in nature, but the literature also suggests the observed link counts may be inconsistent with the link counts assumed by the equilibrium state (Behara et al., 2021b, Fisk, 1989). Behara et al. (2021b) propose to adopt a set of observed path choices from a data source. Therefore, the path choice function could be simplified from a dynamic process into a static process. As a result, the optimization process can be much faster to reach the result without compromising on the quality.

The optimization formula stated above aims to minimise the total weighted distance between each piece of prior information, including the observed traffic counts, and the prior knowledge from different data sources about the target OD matrix.

Ideally, as suggested by Michau et al. (2019), each component in the target function is expected to be convex, which makes the target function itself a convex function that has a single (global) optimum. Then, the inference algorithm will end in the same result regardless of the starting point. However, it is aware that the features of prior knowledges may impact choices of distance measurements $D_1$ in the above generalised form of target function, and the choices of distance measurements could impact the quality of the inference result as they could introduce nonconvexity into the function. As result, the weighted sum of distance components could have multiple local optimums instead of having a single global optimum in a convex function. We may encounter this issue when consider utilising a piece of prior knowledge of the ODM which its scale is different from the expected underlying ODM, and we want to capture its structural information to infer the underlying ODM.

The real-world case study performed in this research falls into the nonconvex problem category. In the Adelaide case study, we infer the underlying ODM in the AM and PM period from three sources of information: the site-level SCATS traffic counts, an observed partial ODM of Bluetooth-connected vehicles and an estimated ODM obtained from the Census database (referred as the seed ODM in this case study). Therefore, there are three distance components in the optimization formula:

(1) The distance between the observed and estimated SCATS counts.

(2) The distance between the observed Bluetooth OD matrix and the estimated partial OD matrix which has an equivalent scope; and

(3) The distance between the seed matrix and the estimated OD matrix.
The optimization algorithm minimizes the sum of three distance components to reach an estimated OD matrix which considers available information from different data sources. Therefore, it is important to choose appropriate distance metrics. It is worth noting that in this real-world case study, three distance components have different scales. Firstly, the distance between the observed and inferred SCATS counts focuses on the distance between the absolute counts. Secondly, the distance between the observed and inferred Bluetooth-based partial OD matrix focuses on the structural comparison since the Bluetooth site network only captures a proportion of the underlying traffic movement. Lastly, the distance between the seed OD matrix and the inferred OD matrix mainly provides structural information to the inference algorithm but the absolute figures are also relevant since the home-work trips are expected to account for a considerable proportion of trips, especially during the AM and PM peak periods.

Another consideration is that, since the three components are in a summation form when inferring the underlying OD matrix, and the target of the inference algorithm is to minimize the total distance, it is important to have the three distance components on a similar scale. Given the above considerations, in the current modelling exercise, we choose to use cosine distance to reflect:

1. The distance between the observed and inferred partial Bluetooth OD matrix; and
2. The distance between the seed OD matrix and the inferred OD matrix.

The cosine distance is calculated as 1 minus cosine similarity (here we flatten the matrices and treat them as vectors), in a formula it is:

\[
\text{Cosine Distance} = 1 - \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}
\]

Where:

- \(n\) stands for the length of the vector.
- \(A_i\) and \(B_i\) stand for the corresponding figures in the observed and inferred OD matrices (in the form of a vector), respectively.

Since the value of cosine distance is between 0 and 1, we also want to make the distance between the observed and inferred SCATS counts between 0 and 1. Therefore, for the distance between the observed and inferred SCATS counts, we calculate the Euclidean distance and normalize it into the 0 to 1 range.

Therefore, given the above considerations, the target function to minimize is structured as:

\[
f = w_q \times \frac{\sqrt{\sum_{i=1}^{n} (q_i - \hat{q}_i)^2}}{\sqrt{\sum_{i=1}^{n} (0 - \hat{q}_i)^2}} + w_{bt} \times \left(1 - \frac{\sum_{i=1}^{m} \hat{T}_{bt_i} \times T_{bt_i}}{\sqrt{\sum_{i=1}^{m} \hat{T}_{bt_i}^2} \times \sqrt{\sum_{i=1}^{m} T_{bt_i}^2}}\right) + w_s \times \left(1 - \frac{\sum_{i=1}^{k} \hat{T}_{s_i} \times T_{s_i}}{\sqrt{\sum_{i=1}^{k} \hat{T}_{s_i}^2} \times \sqrt{\sum_{i=1}^{k} T_{s_i}^2}}\right)
\]

Subject to: \(q = F(T)\)

Where:
\( \tilde{q} \) denotes the observed site-level SCATS counts,
\( q \) denotes the estimated site-level SCATS counts based on the inferred ODM,
\( n \) denotes the number of SCATS sites to be used in the calculation,
\( \tilde{T}_{bt} \) denotes the observed Bluetooth-based (partial) ODM,
\( T_{bt} \) denotes the part of the inferred ODM that has the same scope to \( \tilde{T}_{bt} \),
\( m \) denotes the number of OD pairs in the observed Bluetooth-based ODM,
\( \hat{T}_{bt} \) denotes the observed Bluetooth-based (partial) ODM,
\( T_{bt} \) denotes the part of the inferred ODM that has the same scope to \( \hat{T}_{bt} \),
\( m \) denotes the number of OD pairs in the observed Bluetooth-based ODM,
\( F(T) \) denotes the calculated empirical average path choice (represented by fractions of SCATS site counts) for each OD pair in the scope of study. This empirical path choice is calculated from Bluetooth-based trajectories and then translated into the corresponding SCATS sites. Since the empirical path choice is predetermined and not to vary with the number of trips between each OD pair, the inference process remains at a single level.

Last, as explained previously, it is aware that because of the cosine similarity as a distance measurement, the inferred ODM may not reach the global optimum and the iteration could end up in a local optimum which is sub-optimal. To examine the stability of the inference result, we repeated the inference using random initial points.

3. A real-world case study in the Adelaide region

To demonstrate the effectiveness of the proposed inference framework, we performed a real-world case study in the Adelaide region and compare the inferred ODM using the proposed inference framework to those from the MASTEM system used by the Department for Infrastructure and Transportation (DIT) in South Australia. The results show the effectiveness of the proposed ODM inference framework.

3.1. The scope of study

The study area in the real-world experiment covers the Adelaide region and it contains 102 Statistical Area Level 2 (SA2) zones which will be used as origins and destinations in the experiment. The geographical information for SA2 zones used in this experiment is collected from the Australian Bureau of Statistics website\(^1\). The following figure presents the geographic scope of the case study.

---

3.2. The study periods

This real-world experiment focuses on the workday in May 2019. This is mainly for the convenience to compare with the ODM from the MASTEM system. Also, in this case study, we focus on ODM in the morning and evening peak periods in which the transport network tends to be under pressure:

- 07:00:00 – 09:59:59 AM peak period
- 16:00:00 – 18:59:59 PM peak period

3.3. Data sources

The information utilised to infer the ODM includes a set of trip data obtained by the Bluetooth scanning site network, the traffic counts obtained by the SCATS detectors, and the prior knowledge of Work-Home travel ODM from the 2016 Census performed by the ABS. Therefore, raw data collected in this experiment include:

- The location (coordinates) for Bluetooth sites and SCATS sites (obtained from the Addinsights system from the DIT).
- The Bluetooth-based trip records in May 2019 (obtained from the Addinsights system from the DIT).
- The site-level SCATS traffic counts (obtained from the Addinsights system from the DIT).
- The estimated travel-to-work-by-car number in each SA2 zone (2016 census data, obtained via ABS table builder).
3.4. Data processing workflow

The following presents the pre-processing workflow to convert raw data into inputs into the inference algorithm.

Figure 2 - Data sources and data pre-processing workflow

3.5. Obtaining the empirical path choice

Based on the observed Bluetooth trip data, the average observed trajectories (as presented in a list of Bluetooth sites passed) for each pair of origin and destination could represent the empirical path choice for trips between each pair of origin and destination. With this assumption of the trip path choice, we could estimate the trip path choice between each pair of origin and destination using the raw Bluetooth record following several steps:
(1) Inferring Bluetooth trips from the raw Bluetooth record.
(2) For each Bluetooth trip, presenting its trajectory using a list of Bluetooth sites passed.
(3) For each pair of origin and destination, calculating the average Bluetooth trajectory.
(4) For each average Bluetooth trajectory, translate the Bluetooth sites to their corresponding SCATS sites.

It is worth reminding that there are still 8 out of 102 SA2 zones not covered by Bluetooth sites. This means out of all possible 10404 (102*102) OD pairs, we only have observations on 8836 (94*94) of them. Therefore, to obtain complete path choice data as an input of the inference algorithm, we used the following procedure to approximate the Bluetooth-based trajectories for the missing 1568 OD pairs. For this purpose, we constructed a simplified network for the scope of study which uses its centroid to represent each SA2 zone. In this network, assuming each SA2 zone is only directly to its neighbours (the neighbours for each SA2 zone are calculated based on the GIS information). Therefore, nodes stand for zonal centroids and edges stand for the direct connection between neighbouring zones. Then, the following procedures are adopted:

(1) For the trips originated from a non-covered SA2 zone X. Find zone X’s neighbouring zones.
   a. If the destination zone is covered by Bluetooth sites and the destination is not one of zone X’s neighbours, find the one (zone Y) that is nearest to the destination, and assume the trips from zone X to the destination will pass zone Y. Therefore, we could use the observed average path choice from zone Y to the destination to approximate the path choice from zone X to the destination.
   b. If the destination zone is covered by Bluetooth sites and is one of X’s neighbours, we could use the observed average path choice from the destination to the destination itself to approximate the path choice from zone X to the destination.
   c. If the destination is not covered by Bluetooth sites, and if the destination is one of X’s neighbours or is X itself, the path choice is an empty list of Bluetooth sites.
   d. If the destination is not covered by Bluetooth sites, but it is not one of X’s neighbours nor X itself, find X’s neighbour (zone Y) which is nearest to the destination, find the destination’s neighbour (zone Z) which is nearest to zone X. Then, we could use the observed path choice from zone Y to zone Z to represent the path choice from X to the destination.

(2) For the trips destined to a non-covered SA2 zone X. Find zone X’s neighbouring zones.
   a. If the origin zone is covered by Bluetooth sites and it is not one of X’s neighbours, find X’s neighbour (zone Y) which is nearest to the origin. Then, we could use the observed path choice from the origin to zone Y to approximate the path choice from the origin to zone X.
   b. If the origin zone is covered by Bluetooth sites and it is one of X’s neighbours, then, we could use the observed path choice from the origin zone to the origin zone to approximate the path choice from the origin to zone X.
   c. If the origin is not covered by Bluetooth sites, and if the origin zone is X’s neighbour or X itself, the path choice is an empty list of Bluetooth sites.

If the origin is not covered by Bluetooth sites, but it is not X’s neighbour nor X itself, find X’s neighbour (zone Y) which is nearest to the origin, find the origin’s neighbour (zone Z) which is nearest to zone X. Then, we could use the observed path choice from zone Z to zone Y to approximate the path choice from the origin to X.
4. Results

In this section, we present the inferred OD matrices, examine the results, and compare the inferred OD matrices to those from the MASTEM systems. In Section 4.1, we present inference results for the AM peak period. Similar ODMs were estimated for other time periods as well, with similar results, but for the sake of brevity we do not include them here. To demonstrate the value of allowing incorporating multiple pieces of prior knowledge of the ODM to facilitate inferring the ODM, Section 4.2 presents the inferred ODM without incorporating the seed ODM obtained from the ABS census data. The following table presents the weightings for the two scenarios:

Table 1 - Weightings for scenarios

<table>
<thead>
<tr>
<th>Period</th>
<th>Use Seed ODM</th>
<th>Weighting for traffic counts</th>
<th>Weighting for seed ODM</th>
<th>Weighting for Bluetooth ODM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM peak</td>
<td>Yes</td>
<td>1</td>
<td>0.75</td>
<td>0.5</td>
</tr>
<tr>
<td>AM peak</td>
<td>No</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
</tr>
</tbody>
</table>

4.1. AM peak period

For the AM peak period, with the incorporation of the seed OD matrix, the total number of trips in the inferred ODM is around 556 thousand, and the total number of trips in the MASTEM ODM is about 484 thousand. From a structural perspective, the cosine similarity between two matrices is 0.8643 (correlation coefficient 0.8487) and this suggests the inferred ODM in the AM peak period is similar to that from the MASTEM system.

In terms of the trip production, the cosine similarity between that from the inferred ODM and that from the MASTEM ODM is about 0.9695 (correlation coefficient 0.8548). Figure 3 compares choropleth maps of the inferred trip production and that from MASTEM, and it shows inferred trip production is close to that from MASTEM in most SA2 zones.

Figure 4 presents the hot zones for trip production in the AM peak period. We can see that in most hot trip production zones, the inferred number of trips produced is close to that from MASTEM.

For trip attractions, the cosine similarity is also about 0.9765 (correlation coefficient 0.9676). Figure 5 and figure 6 suggest that the inferred trip attraction is close to that from MASTEM in most of SA2 zones. However, while the Adelaide CBD area is the hottest trip attraction zone in both choropleth maps, the inferred number of trips attracted is about 18,500 higher than that from MASTEM, this is mainly because the seed ODM from the Census data suggests there are about 11% (8.9% from MASTEM ODM) of trips attracted to the Adelaide CBD for commute purpose by car, and this structural information is captured by the inference algorithm.
Figure 3 - Trip production in the weekday AM peak period (Inferred with the seed ODM)

Figure 4 - Top trip production zones in the AM peak period
4.2. Comparison: Inference without the seed OD matrix

As another set of comparison, we inferred the ODM without incorporating of the seed OD matrix from the census. For the AM peak period, the total number of trips in the inferred ODM is around 783 thousand, which is about 41% higher than that from the inferred ODM with incorporating the seed ODM and 62% higher than the from MASTEM estimation. From a structural perspective, the cosine similarity between the inferred ODM and the MASTEM one drops to 0.3731 (correlation coefficient 0.3672) and this suggests the inferred ODM is much less similar to that from the MASTEM estimation.

For the trip production, the cosine similarity between that from the inferred ODM and that from the MASTEM ODM is about 0.7409 (correlation coefficient 0.4448). And for trip attractions, the cosine similarity is around 0.7883 (correlation coefficient 0.6161). The decrease in the similarity level is visually obvious in figure 7 and figure 8.
In general, the comparison between the inferred OD matrices with and without incorporating the seed OD matrix from the census data and the estimated OD matrix from the MASTEM system shows that, the proposed inference framework could noticeably improve the inference quality by integrating multiple pieces of prior information into the inference process.

4.3. Discussion

The real-world experiment in the Adelaide region demonstrates the effectiveness of the proposed ODM inference framework. By integrating information from multiple data sources (three in the experiment), even each piece of input data only contains a fraction of information and could be biased to some extent, the proposed framework can produce reasonable and highly comparable ODM, trip production and trip attraction to those from the traditional traffic modelling process. In terms of the benefits, from a flexible perspective, the proposed inference framework allows the modeller incorporating multiple pieces of prior knowledge of the underlying ODM (e.g., the partial ODM collected from GPS-connected cars, the partial ODM obtained from the mobile data). By incorporating multiple data sources, the inferred ODM is less likely to be distorted by the data sources with relatively lower quality and is expected to deliver more reasonable and realistic results. Besides, since the proposed inference framework does not necessarily require the study area to have a single piece of prior knowledge of the underlying ODM that provides complete and high-quality information, it potentially allows cities and regions that has multiple fractions of information regarding the underlying ODM to start their modelling practice. From a scalable perspective, since the proposed inference framework adopts a single-level approach by assuming a static path choice set, the computational speed is expected to be fast, and it allows the practitioner to easily infer the ODM in a city or region thousands of OD pairs and test the impact of different model configurations. As an example, the Adelaide case study demonstrated in this section is finished on a normal personal laptop in a Python environment (using the Pyomo library and the ipopt solver), with authors’ non-professional coding skills, it takes around 10 minutes to infer the ODM among 10404 OD pairs, and this computational time could be significantly improved with better programming skills and/or more computational resources.

Figure 7 - Trip production in the weekday AM peak period (Inferred without the seed ODM)
Figure 8 - Trip attraction in the weekday AM peak period (Inferred without the seed ODM)

5. Conclusion

In this study, we proposed a flexible and scalable framework for the ODM inference. The flexibility of the proposed framework allows integrating traffic data from multiple data sources and allows each data source to be less granular or only provide a fraction of information or biased to some extent. The scalability brought by the single-level design makes the proposed framework to be computationally fast and allows the modellers to easily estimate the ODM among thousands of OD pairs. The proposed framework is expected to allow more cities and regions to perform ODM inference using their available data sources. To prove the effectiveness of the framework, we performed a real-world experiment in the Adelaide region that uses site-level traffic counts, Bluetooth data, and a piece of prior knowledge obtained from the Census for the ODM inference. The result shows that for the AM and PM periods, the proposed framework can produce highly comparable ODM, trip production and trip attraction to those from the traditional transportation modelling processes (MASTEM system in this case).

Acknowledgements

This research is funded by iMOVE CRC and supported by the Cooperative Research Centres program, an Australian Government initiative; and the South Australian Department for Infrastructure and Transport.

Reference


SPIESS, H. 1990. Gradient Approach for the O-D Matrix Adjustment. University of Montreal, Canada: Centre for research on transportation.