Rethinking traffic count methodologies to derive OD matrices from aggregated mobile phone data

Clémence de ROLLAND*[1, 2] and Caroline BAYART[3]

[1]Laboratoire Aménagement Economie Transport, Université Lumière Lyon 2, Lyon, France
[2]Ecole nationale des ponts et chaussées, Marne-la-Vallée, France
[3]Laboratoire Chrome, Université de Nîmes, Nîmes, France

Short summary

Among the recently identified Big Data sources identified for mobility analysis, Mobile Phone Data (MPD) has been is considered as a promising passive source of information to complement traditional travel surveys thanks to large samples that are not limited to one mode of transport. Unlike individual MPD, aggregated MPD avoids privacy concerns and can be collected over long periods of time, but it only provides the number of people in a given area during a given time interval and requires further processing for mobility management. This paper therefore proposes a theoretical framework for generating OD matrices from AMPD, using methods derived from traffic counts. To ensure proper data transformation, several algorithms are tested, first in simulation and then using real data. The results demonstrate the potential of AMPD to generate high quality OD matrices on a continuous basis.
Keywords: Big Data Analytics, Transport demand analysis, Mobile Phone Data, Origin-Destination matrix

1 Introduction

Household Travel Surveys have long been the primary data source in transport studies, providing tools such as origin-destination (OD) matrices for studying mobility behaviors and planning transport services. However, they face limitations such as low frequency, small sample sizes, and the inability to capture less used modes or weekend/holiday travel patterns. With increasing uncertainty and rapid changes in mobility behavior, new passive data sources have been explored to complement these surveys and overcome their limitations. Among these data sources, Mobile Phone Data (MPD) are particularly promising, as they provide large-scale; continuous data, collected from larger samples and from users of all transport modes (unlike, for example, smartcard data, which only collects data from public transport users).
Among MPD, two types of datasets coexist : individual datasets, that follow users throughout a day or a week, allowing OD matrices to be built directly from the traces, but that are limited to smaller samples and time periods due to privacy concerns; and aggregated datasets (AMPD), which do not suffer from these privacy limitations but provide very raw information (a number of persons per zone or per OD pair, per time interval). AMPD provided by OD pair have shown a good correlation with other data sources, but may underestimate short trips because they mostly use a time threshold value to define a trip from individual traces (Dypvik Landmark et al., 2021; Casassa et al., 2024) and are also limited by anonymization thresholds.
AMPD provided by zone have been less explored, while they are less limited by these thresholds and can provide interesting information on mobility behaviors with further processing. However, to our knowledge, their potential to generate Origin-Destination (OD) matrices has been underexplored. Our motivation is to exploit these data to their full potential. To this end, we draw an analogy with traffic counts data, as both data sources provide information on the number of individuals passing through a given zone. Our research objectives are to: (1) develop a theoretical framework to convert AMPD zonal data into variables similar to traffic counts data, which we will call *cellcounts*, (2) test existing methods for building OD matrices from traffic counts data with *cellcounts* data and (3) evaluate the transferability of this approach through simulations and real data experiments.

2 Methodology

The AMPD studied in this paper consist of three indicators: a **Presence** indicator n, an **Entrance** indicator e and an **Occupancy** indicator o. Each indicator is provided for each zone i, for a 1h interval Δt_k. The indicators are defined as follows:

- $n_i(\Delta t_k)$ is the number of persons who were detected at least once in i during Δt_k;
- $e_i(\Delta t_k)$ is the number of persons who were detected at least once in i during Δt_k and who were not detected in i during Δt_{k-1};
- $o_i(\Delta t_k)$ is the number of persons who spent the relative majority of Δt_k in i.

Our framework is based on an analogy between the AMPD and traffic counts: in both cases, an individual going from i to j triggers events (mobile phone events or traffic counts) along the way. If the routes from i to j are known for all OD pairs, the sum of a given counter ("mobile phone network" counter or classic traffic counter) can be expressed as the sum of the flows of all OD pairs passing through it. For the AMPD data, the problem can be formulated as follows:

For a given time interval Δt, for each ij pair where $i \neq j$,

$$\text{find } t_{ij}(\Delta t) \text{ so that } \forall k \in \mathcal{Z} : \begin{cases} x_k(\Delta t) = \sum_{ij \in \mathcal{Z}^2} t_{ij}(\Delta t) \pi_{ij}^k \\ v_j(\Delta t) = \sum_{i \in \mathcal{Z}, i \neq j} f_{ij}(\Delta t) \\ w_i(\Delta t) = \sum_{j \in \mathcal{Z}, j \neq i} t_{ij}(\Delta t) \end{cases} \tag{1}$$

with:

- $v_j(\Delta t)$ is the number of persons who are present in j when Δt ends, i.e. the entering persons;
- $w_i(\Delta t)$ is the number of persons who are present in i when Δt starts, i.e. the persons leaving;
- $x_k(\Delta t)$ is the number of persons who cross zone k during Δt without staying in zone k, which is the variable analog to traffic counts. We will call this variable *cellcounts*;
- $\mathcal{Z}$ is the set of zones in the studied area ;
- $t_{ij}(\Delta t)$ is the number of persons going from i to j during Δt (flow);
- π_{ij}^k is the proportion of the flow from i to j going through k, which depends on the affectation model used to calculate the itineraries.

The pipeline to obtain an OD matrix is divided into two stages: (S1) obtaining the *cellcounts* values $x_i(\Delta t)$, the *cellcounts* sums $v_i(\Delta t)$ and $w_i(\Delta t)$, as well as the number of intrazonal trips $u_i(\Delta t)$ from the AMPD datasets, and (S2) solving the problem (1), which is identical to the fundamental problem of obtaining the OD matrix from traffic counts.
To obtain the variables in (S1), we relied on the link between the *cellcounts* and the information in the AMPD. We have added the relationship between the variables at Δt_k and a Δt_{k+1} due to the conservation of the number of persons between the two time intervals. We have added three parameters $p_{vi}(\Delta t)$, $p_{wi}(\Delta t)$ and $p_{xi}(\Delta t)$ corresponding to the proportion of users $v_i(\Delta t)$, $w_i(\Delta t)$ and $x_i(\Delta t)$ who spent the majority of Δt in i and are counted as occupants $o_i(\Delta t)$ in the AMPD. We obtain the following system of equations:

$$\begin{pmatrix} n_i(\Delta t) \\ o_i(\Delta t) \\ e_i(\Delta t) \\ n_i(\Delta t+1) - e_i(\Delta t+1) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & p_{vi}(\Delta t) & p_{wi}(\Delta t) & p_{xi}(\Delta t) \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_i(\Delta t) \\ v_i(\Delta t) \\ w_i(\Delta t) \\ x_i(\Delta t) \end{pmatrix} \tag{2}$$

which is invertible when $p_{vi} + p_{wi} - p_{xi} \neq 1$, allowing us to obtain expressions for the *cellcounts* x_i as well as the *cellcounts* sums v_i, w_i and the intrazonal trips u_i.
The pipeline for (S1) also includes a module to obtain the parameters p_{vi}, p_{wi} and p_{xi} from a random draw guarantying that the *cellcounts* are positive.
Once the *cellcounts* have been obtained, the fundamental traffic counts equation can be applied and solved in (S2).
To solve this fundamental traffic counts equation, we test the classic methods described in Bera & Rao (2011): a Gravity Matrix model (GM), simple Least Squares (LS), Entropy maximization (EM), Bayesian Inferences (BI), Generalized Least Squares (GLS) and Maximum Likelihood (ML). We also test the Weighted Data Fusion (WDF) method developed by Sun et al. (2023). Although these methods are neither the latest nor the most advanced, they are the most tried-and-tested and transparent methods for this problem, giving us greater control over our results.

3 Results and discussion

Two cases were studied: a simulated case, necessary to obtain accurate comparative data to test the pipeline, and an experimental case to conclude on the replicability of our method. The experimental case focused on the city of Rouen, France, where a recent survey is available to assess the results. The simulated case was developed on the city of Rodez, France, a simple and small monocentric city to use as a sandbox.

Simulated case

To obtain the simulated data, we derived the expected OD flows and AMPD from a set of simulated trips in the Rodez area (Aveyron, France), divided into 28 zones. We generated synthetic trips for N=63,473 persons (population of the area in 2023), for a weekday morning, based on the workplaces and places of study of the French Census. Although these databases only contain commuting trips, they can give a plausible approximation of the flows in the area on a morning, which we can use for the simulation.
The simulation is first run 25 times. As can be seen in figure 1, obtaining the *cellcounts* x_i in (S1) of the pipeline gives very good results, with a quasi-null dissimilarity index and a very small RMSN=0.13 ± 0.01.

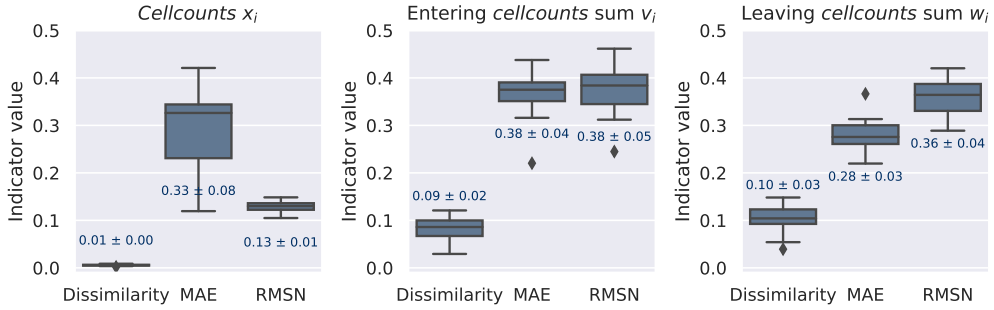

Figure 1: Distribution of the indicators obtained for the 25 iterations of the simulation in Rodez.

Once the *cellcounts* are obtained, we run 10 simulations applying each of the methods listed in the methodology. Some of these methods require prior knowledge of the OD flows that are updated, or a sample matrix to use as a constraint : we used an OD matrix derived from the AMPD using the gravity model method, that requires no parameters or assumptions. For the methods depending on constrained optimization processes with constraints on the counts, we added tests with constraints corresponding to the occupancy o_i and to the information we have on the margins of the MOD (the total of persons entering or leaving a zone i, i.e. the *cellcounts* sums v_i and w_i). Several tolerances $\xi_{margins}$, ξ_{counts} and ξ_{pop} were tested for these constraints. Finally, some methods depend on parameters, for which different values were tested : weights of the datasets w_{MOD} and w_{counts} in the WDF method, distribution forms of the variables in the ML method. For each test, we compute the RMSN and MAE between the computed OD flows and the expected OD flows (see figure 2 and table 1).

Results are consistent from one run of the simulation to the other, as the points are fairly grouped for each test reference. The methods with $\xi_{margins} = 0.0$ in the optimization process generally do not converge. This was to be expected, as the computed margins v_i and w_i are not exact (see figure 1), and should therefore not be used in strict constraints. The Maximum Likelihood approach is the least performing, which could be explained by the fact that the "sample" matrix is the Gravity Model matrix, already biased. The weighted data fusion method from Sun et al. (2023) gives better results. The best performing method seems to be the Entropy Maximization, with parameters $\xi_{counts} = 0.1$, $\xi_{margins} = 0.2$, $\xi_{pop} = 0.05$. These parameters seem quite logical with the rest of the results: the largest tolerance is for the *cellcounts* sums, and the tolerance for the *cellcounts* themselves is smaller and around the RMSN value obtained in figure 1. Finally, the tolerance for the total population is quite small, as this is the most certain data.

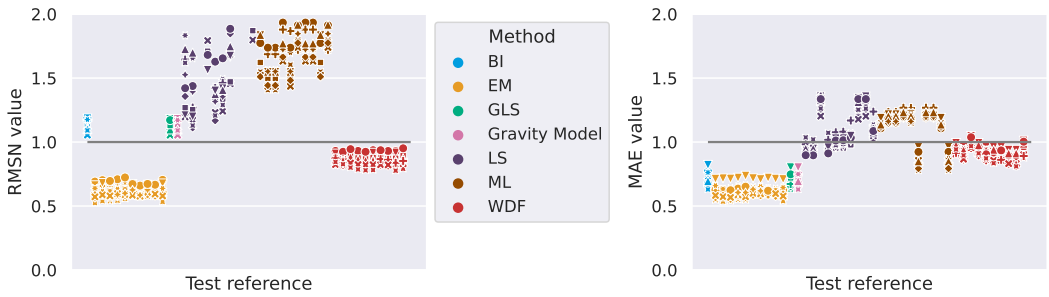

Figure 2: Results for 10 simulations, using the gravity model matrix as an initial/samples matrix. Each marker represents a different simulation. For frameworks with multiple parameters, only the best 10 sets of parameters (on average) were plotted.

Table 1: Results for the 10 simulations plotted in figure 2. Only the best set of parameters for each methods (on average among the simulations) is shown.

Framework	**RMSN**	**MAE**	ξ_{pop}	$\xi_{margins}$	**Other parameters**
Gravity Model (Initialization)	1.133	0.701			
BI	1.137	0.712			
EM	0.607	0.600	0.05	0.2	ξ_{counts}= 0.1
GLS	1.133	0.701			
LS	1.332	0.943	∞	∞	
ML	1.542	1.186	0.2	0.5	Sample distribution: Poisson Counts distribution: MVN
WDF	0.852	0.901	∞	∞	w_{MOD}: 1.0 w_{counts}: 0.25

To improve these first results, we then implement sequences of methods, starting with the GM, then the EM with $\xi_{counts} = 0.1$, $\xi_{margins} = 0.2$, $\xi_{pop} = 0.05$, then other methods. We repeat these sequences 10 times again. Results are shown in figure 3 and are compared with the previous steps (table 2). There are still errors, which is consistent with the under-specification of the problem, but the results are improved for all methods compared to the previous experiments.
This improvement is very important for frameworks like GLS and BI, which mostly depend on the previous or initial matrix. A longer sequence would probably not improve anymore the results, as they are all close to the EM matrix. The best RMSN is obtained with the WDF method. The MAE is larger than that of the initialization, which means that there are relatively more small errors and fewer large errors. Figure 4 shows that, for this sequence, the points are grouped around the identity line and no big deviation is observed: in general, the results are in the right order of magnitude.

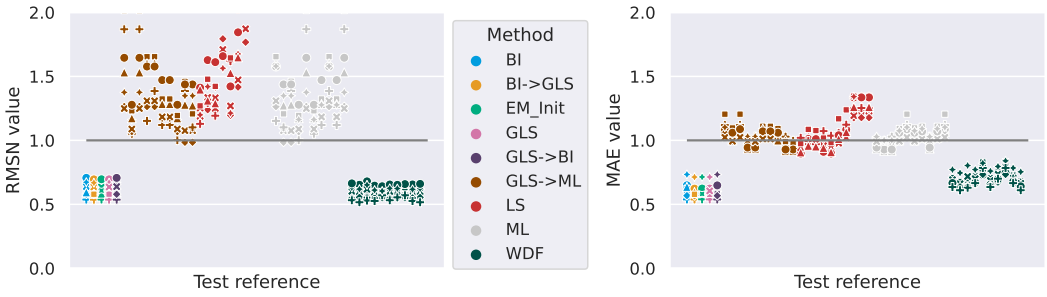

Figure 3: Results for 10 simulations, using the best Entropy Maximization matrix as an initial/samples matrix. Each marker represents a simulation. For frameworks with multiple parameters, only the best 10 sets of parameters (on average) were plotted.

Table 2: Results for the 10 simulations presented in figure 3. For frameworks with multiple parameters, only the best set of parameters (on average among the simulations) is shown.

Framework	**RMSN**	**MAE**	ξ_{pop}	$\xi_{margins}$	**Other parameters**
EM (Initialization)	0.6112	0.6009	0.05	0.2	ξ_{counts}= 0.1
BI	0.6156	0.6178			
BI then GLS	0.6109	0.6028			
GLS	0.6113	0.6032			
GLS then BI	0.6152	0.6174			
GLS then ML	1.1381	1.0366	∞	0.5	Sample distribution: Poisson Counts distribution:MVN
LS	1.2616	0.9482	∞	∞	
ML	1.1381	1.0366	∞	0.5	Sample distribution: Poisson Counts distribution:MVN
WDF	0.5791	0.6994	0.1	∞	w_{MOD}: 0.75 w_{counts}: 0.25

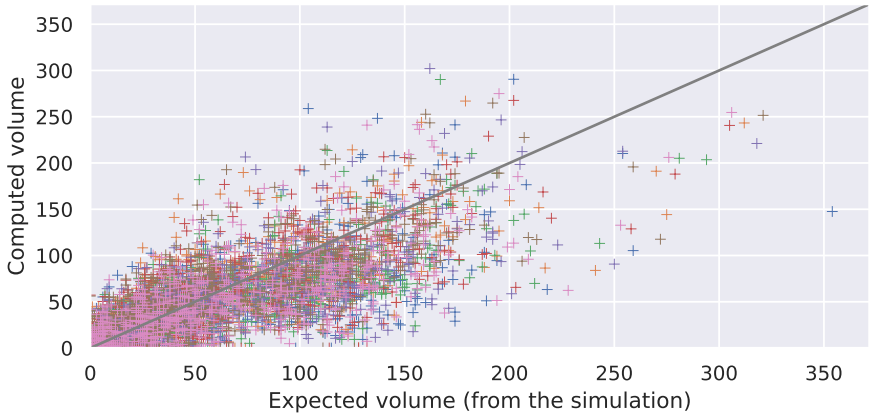

Figure 4: Results for the sequence Gravity Model, then Entropy Maximization, then Weighted Data Fusion. Each color corresponds to a run of the simulation.

Experimental case

Aggregated Mobile Phone Data used for this part were collected in the area of Rouen (Normandy, France) (table 3), divided into 484 zones, which is much larger than our simulated case. Mobile Phone Events were collected on Tuesdays from 2023/11/06 to 2023/12/17, a period with school holidays, and have an average inter-event time of 122±479 s, which ensures good tracking of the devices. They were compiled and extrapolated by Orange using its local market share, socio-demographic characteristics, as well as spatial and temporal corrections. The results of our pipeline are compared with the 2017 Rouen mobility survey (EMD), conducted in the same area on 2.27% of the residents.

Indicator	**Values** mean ± std (median)
Total number of persons present in the studied area (sum of the occupancies in the studied area)	599,330
Occupancies per zone (o_i)	1371.5 ± 1100.6 (1093)
Presences per zone (n_i)	7109.8 ± 6693.8 (4923.5)
Entrances per zone (e_i)	5097.6 ± 5046.4 (3471.8)
Cellcounts per zone (x_i)	3285.1 ± 3535.3 (2100.3)
Cellcounts sum of entering flows per zone (v_i)	1812.5 ± 2114.3 (1061.7)
Cellcounts sum of leaving flows per zone (w_i)	1683.5 ± 1791.3 (1045.6)

Table 3: Characteristics of the AMPD dataset and the derived variables.

The best method identified with the simulation is run with the AMPD of the studied area of Rouen : a Gravity Model matrix for the initialization, then an Entropy Maximization method with parameters $\xi_{counts} = 0.1$, $\xi_{margins} = 0.2$, $\xi_{pop} = 0.05$. From an aggregated point of view, the total number of interzonal trips is similar between both datasets : 68,305 for the survey and 82,053 for the AMPD. The higher value in the OD matrix extracted from the AMPD could be explained by the increase in the population between 2017 and 2023, and by the restriction to residents in the survey. Indeed, the AMPD include everyone who was present in the area, including non-residents, tourists and people who travelled through the city. Although this information may seem like noise, having data on non-residents is an interesting improvement over traditional surveys. On average, there was 609.9 ± 924.3 trips in the survey compared to 732.6 ± 1086.5 in the AMPD, a similar order of magnitude.
Comparing OD flows from the survey and from the OD matrix dervied from the AMPD, we obtain a Pearson coefficient of 0.93420 with a p-value less than 10^{-6}, showing a good correlation between the datasets. The error indicators are $RMSN = 0.6803$ and $MAE = 0.5264$, close to those obtained in the simulated experiment, and quite satisfactory. Finally, a linear regression between the two vectors gives a slope $a = 1.098$ and an intersection $b = 62.9$, with $r^2 = 0.87278$, which shows the overestimation of the OD flows extracted from the AMPD compared to the OD flows from the survey but also the very good correlation between the two data sources.
Finally, the graphical comparison shown in figure 5 shows that the points are grouped around the identity line, as obtained with the simulation experiment. The overestimation of the OD flows extracted from the AMPD compared to the EMD flows is also visible, and shows that it does not induce a major deviation for a given OD pair, but rather a constant ratio.

Discussion

Our results are satisfactory at this proof-of-concept stage, showing good correlation and order of magnitude in both the simulated and experimental results. Results could be improved by using more recent and advanced frameworks used for traffic counts data analysis, such as those based on machine learning or artificial intelligence. In the experimental case, data had to be aggregated to the zoning of the survey, which is much less precise than the zoning for the AMPD. Although the correlation is good at the aggregated level, we have no information about the quality of our results at the disaggregated level. Further research could include a comparison with yet another data source to improve this validation. Other interesting steps could include adding other data sources to our methods, such as a precompiled OD matrix from mobile phone data that uses thresholds to define the trips, to see if this improves our results.

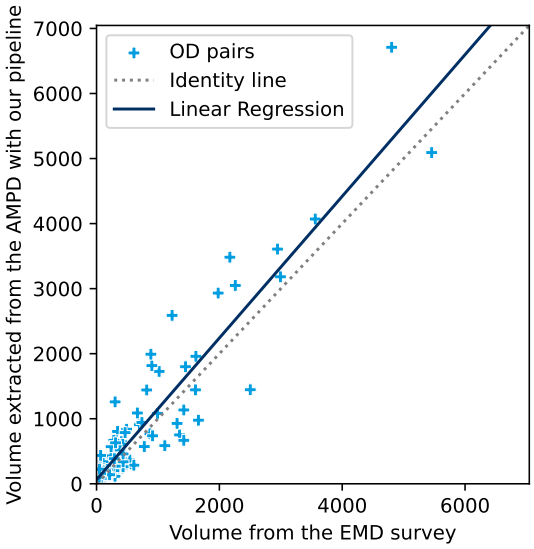

Figure 5: Experimental results of the sequence in the Rouen area.

4 Conclusions

In this paper, we have shown that it is possible to derive *cellcounts* indicators similar to traffic counts data from an Aggregated Mobile Phone Data dataset containing only Presence, Entrance and Occupancy indicators. These *cellcounts* can be used to derive OD flows matrices using the methods developed for traffic counts data. Several methods were tested for this purpose, the most relevant being a sequence of Gravity Model followed by Entropy Maximization, then a Weighted Data Fusion process developed by Sun et al. (2023), which results in an average RMSN of 0.579 in a simulated case. The simulation results are consistent from run to run. From a more qualitative point of view, the obtained OD flows are in the same order of magnitude as the expected ones, which is a crucial characteristic for transport planners to develop adequate transport offer. We proposed to transfer our simulated results to an empirical case in the city of Rouen, France.The comparison between a computed OD matrix and the flows obtained from the survey shows that the OD flows extracted from the AMPD, once aggregated to the survey zoning, give coherent values for the morning time interval, with a RMSN of 0.6803. These results tend to support the hypothesis that the framework is transferable from a simulated case study to an empirical case study, and therefore that Aggregated Mobile Phone Data can provide useful insights into mobility. The information obtained is complementary to traditional surveys: although we do not have any socio-economic or purpose information for the trips, AMPD are available for any time of the day, day of the week and period of the year and could allow for longitudinal studies, and are not restricted to a predefined administrative area. Finally, although we have focused here on MPD, our framework could theoretically be applied to any type of passive data that can provide presence data, entrance data and occupancy data, such as Bluetooth or Wifi counters, or even satellite images. Another way forward would, therefore, be to test our framework with other types of data sources.

Acknowledgements

We thank Transdev for providing us access to the Orange Mobile Phone Data. We thank the Métropole de Rouen for providing the survey data.

References

Bera, S., & Rao, K. V. K. (2011). Estimation of origin-destination matrix from traffic counts: the state of the art. *European Transport*, *49*.

Casassa, E., Côme, E., & Oukhellou, L. (2024, September). Detected or undetected, which trips

are seen in mobile phone OD data? A case study of the Lyon region (France). In *TRC-30*. Crete, Greece.

Dypvik Landmark, A., Arnesen, P., Södersten, C.-J., & Hjelkrem, O. A. (2021, October). Mobile phone data in transportation research: methods for benchmarking against other data sources. *Transportation*, *48*(5), 2883–2905.

Sun, W., Vij, A., & Kaliszewski, N. (2023, September). A flexible and scalable single-level framework for OD matrix inference using IoT data. *Transp. Res. A*, *175*, 103775.