Estimating commuting matrix and error mitigation – A complementary use of aggregate travel survey, location-based big data and discrete choice models

Wan, L., Yang, T., Jin, Y., Wang, D., Shi, C., Yin, Z., Cao, M. and Pan, H.

Estimating commuting matrix and error mitigation: A complementary use of aggregate travel survey, location-based big data and discrete choice models

Abstract

The prevalence of location-based big data has opened up a new research frontier for estimating origin–destination commuting matrices for cities where granular flow data are not yet available from official sources. However, so far there have been few investigations into estimation errors and potential correction methods in the literature. To address the research gap, this paper first compares the performance of two estimated commuting matrices for Shanghai, derived by two distinct matrix estimation methods, namely a big-data approach using mobile phone signalling data and a discrete choice model for simulating the residential location of commuters. The empirical results indicate that there is an outstanding analytical complementarity between the two approaches. A novel method is then proposed for mitigating the errors associated with the big-data approach. The proposed method features a selective blending of the big-data based flow estimation and the model-based estimation. By comparing the blended flow estimation with benchmark travel statistics, we find that the proposed method would significantly reduce the estimation errors and hence improve the robustness of the estimated matrix. It is expected that the proposed method will set a new standard for correcting potential errors in big-data based flow estimation.

Keywords: matrix estimation; commuting; location-based big data; mobile phone data; transport modelling

1 Introduction

The journey-to-work matrix, also known as the origin–destination commuting flow, is a crucial data input for modelling urban land use and transport systems in support of policymaking. Prior to the emergence of novel data sources, model-based approaches have been the mainstay for estimating and simulating origin–destination matrices based on travel surveys. Widely adopted models have included the gravity and the entropy maximisation models of spatial interaction (Batty & Mackie, 1972; Wilson, 1967, 1970) as well as discrete choice models based on random utility theory (Anas, 1983; McFadden, 1973). Among these model applications, the land-use and transport interaction models have played a pivotal role in practical policy appraisals (Lowry, 1964; Batty, 1976, 2009; Echenique et al., 1969; Jin et al., 2013; Wegener, 2004).

However, detailed commuting matrices from official sources are often not accessible, even for academic purposes, particularly for cities in developing countries (Hu et al., 2019). Recent studies have attempted to overcome this data deficiency by employing various location-based big data sources, including smart-card data (Long et al., 2012), GPS tracking data (Papinski & Scott, 2013; Shen et al., 2013; Ta et al., 2016), social media data (McNeill et al., 2017), location-based services (Ahas & Mark, 2005; Wan et al., 2017) and mobile phone data (Bonnel et al., 2015; Kung et al., 2014; Yang et al, 2019; Yuan et al., 2012; Zhou et al., 2018). Location-based big data have untapped potential to produce a robust commuting matrix to support policy endeavours. However, the emerging data sources tend to contain case-specific sampling biases or measurement errors (Chen et al., 2016; Liu et al., 2015; Yang, 2020), which suggests that the derived commuting matrix would require further refinement before it could be used to appraise policies.

This paper proposes a novel method for mitigating the errors in estimated commuting matrices based on limited official travel statistics. The method is informed by an empirical comparison of matrix estimations for Shanghai, which includes a matrix derived from mobile phone signalling data and a matrix estimated by a purpose-built discrete choice model for residence location choice. The complementarity of the big-data approach and the modelling approach is discussed. The results demonstrate that the proposed method significantly reduces estimation errors based on benchmark statistics. Considering the increasing amount of geospatial data available to researchers and policy analysts, the proposed method could be applied to provide a best-match commuting matrix estimation for cities that currently have no official flow data. The proposed error mitigation strategy can improve the robustness of the estimated flow and provide essential data support for the wider transport modelling community.

The paper is structured as follows. Section 2 reviews existing methods for estimating commuting matrices, with a focus on error identification and correction. Section 3 introduces the case study area (Shanghai), data input and workflow. Section 4 discusses the matrix estimation methods and reveals the comparison results. Section 5 explains the new approach to error mitigation and presents an assessment of the refined matrix. Conclusions, limitations and suggestions for future research are provided in Section 6.

2 Literature Review

A vast body of literature exists estimating origin–destination commuting matrices using emerging data sources such as mobile phone (MP) data, with several review papers readily available (Chen et al., 2016; Y. Liu et al., 2015; Steenbruggen et al., 2013; Wang et al., 2018). As most emerging data sources are not intended for transport research, the origin–destination of trips would have to be inferred, commonly through a rule-based algorithm. This is based on the observation that the locations and trip duration of regular commuters are likely to exhibit a high level of regularity.

The specification of the rule-based algorithm is usually done on a case-by-case basis, hence the associated estimation errors. A study by the UK Office for National Statistics (ONS, 2017) using MP data to estimate commuting flows found that there is a good correlation between the MP matrix and the census flow data for relatively long-distance journeys. However, the MP matrix would largely overestimate short-distance journeys. The error is likely to be associated with the rule-based algorithm for location inference. The rule-based algorithm has a methodological limitation in differentiating between home-workers, commuters who travel very short distances and non-commuters with regular travel patterns (e.g. housewives who visit nearby shopping areas twice a week might be identified as commuters). Regarding the processing of MP data, census and mid-year population estimates from ONS are applied to scale the MP data to a full population flow. Wan et al. (2017) used location-based services (LBS) data to infer commuting flows for Beijing and found that a rule-based algorithm tends to perform better at identifying residence locations than employment locations. Yang (2020) further noted that the estimation errors from a rule-based algorithm can differ significantly across locations. Iqbal et al. (2014) estimated the commuting matrix for Dhaka using Call Detail Records and found that the estimation errors (against actual traffic count) also vary significantly across different times of the day.

Despite sampling biases and estimation errors, the existing literature has focused on demonstrating the ‘richness’ of big data regarding the spatial-temporal dynamics of movements. Relatively few papers, however, have addressed the errors or validation of the estimated matrix (Chen et al., 2014; Liu et al., 2016; Wang et al., 2018), or how the identified errors could be mitigated to improve the representativeness of the matrix based on benchmark statistics.

To validate the estimated commuting matrix, three methods have been adopted in the existing literature: (1) comparison against observed data from travel surveys or census; (2) cross-validation with multi-source big data; and (3) comparison against model simulations. As an example of the first approach, Long and Thill (2015) explored commuting patterns in Beijing using bus smart-card data, whereby the estimated matrix was validated with the observed local travel statistics (travel volume, average travel time and distance). Alexander et al. (2015) compared the estimated commuting matrix for Boston with both local and national travel survey data, and identified a relatively large estimation error on travel volume for short-distance, low-speed and low-volume trips. In the case of the second approach, Calabrese et al. (2013) validated the commuting matrix derived from MP data with census data as well as vehicle odometer readings in Singapore.

Using the third approach, Iqbal et al. (2014) compared their estimated MP matrix with results from a micro-simulation model for selected trip origins and also with limited travel count data. Wan et al. (2017) and Yang (2020) validated the derived matrix from LBS and MP data with results from a land-use and transport interaction model as well as aggregate census data. Chen et al. (2014) examined estimation errors associated with a rule-based algorithm based on a simulated MP data set that synthesises information from a structured survey and raw MP traces. Their findings suggested that a satisfactory goodness of fit could be achieved through a location inference algorithm and the model perform better at detecting home locations than that of workplaces.

Among papers that explicitly include a validation exercise, scant few consider how the identified errors could be mitigated. Wismans et al. (2018) published one of the few papers that corrects the MP matrix by substituting trips within a certain distance with a model-based matrix, demonstrating discernible improvement. Another widespread approach to error correction is to apply scaling factors to the sample matrix, such that the scaled matrix could reflect the total population. We argue that this scaling exercise is, in fact, a part of the matrix estimation process as in conventional transport modelling (e.g. through iterative proportional fitting). The scaling may in part address the sampling error of the MP data in the sense that the scaled matrix would correspond to observed trip-end totals, but it does not address the estimation errors associated with the rule-based algorithm as well as the scaling exercise.

One of the main reasons that error correction is not widely considered in the literature is perhaps the lack of ‘ground-truth’ data, which is why an estimated matrix is required. However, even with limited travel survey data, the biases and errors in the commuting matrix estimated with big data should be identified and corrected. In particular, if the derived matrix is to be used for practical policy use, the identified errors (as opposed to limited benchmark data) would undermine the credibility of the model-based policy assessment.

Prior to the prevalence of big data, established methods in the field of transport modelling were used for estimating a trip matrix. Two broad strands of estimation methods can be identified: (1) matrix adjustment through iterative proportional fitting (IPF) (Deming & Stephan, 1940), also known as the Fratar–Furness method, named after Fratar (1954) and

Furness (1965); and (2) matrix generation through the application of spatial interaction models (Batty & Mackie, 1972; Wilson, 1967, 1970) and subsequently discrete choice models (Anas, 1983; McFadden, 1973). The key difference between the two strands is that the former requires an *a priori* seed matrix, while the latter does not. In the case of the first strand, a classic application was presented by Ben-Akiva et al. (1985). Recent applications include Gordon et al. (2018) and Ji et al. (2015). One of the key features of the IPF method is that the adjustment would preserve the trip patterns as in the seed matrix, suggesting that the quality of the seed matrix, in effect, determines the quality of the adjusted matrix (de Dios Ortuzar & Willumsen, 2011). Ben-Akiva (1987) demonstrates a unique solution of the IPF method if the seed matrix contains no zeros. If a large number of zeros exist in the seed matrix (i.e. a sparse matrix), which is not uncommon for sample data, the IPF method may encounter a convergence issue for certain locations. For instance, in extreme yet relatively rare cases where no trip is captured for a workplace or place of residence, that is the cells along one row or column in the matrix are all zeros, the IPF method would fail to meet the observed constraints for that particular row or column. Nonetheless, those locations could be removed from the seed matrix and the IPF method could be applied to the rest of the matrix.

By contrast, generating a matrix using spatial interaction or discrete choice models does not require an *a priori* seed matrix. These models might be a useful alternative for cities where a quality seed matrix remains inaccessible. However, the formulation of spatial interaction and discrete choice models does require considerable additional data inputs, for instance, the measurement of mass and distance for gravity models and the location utility components in discrete choice models. Acquiring such data from conventional sources is usually no less demanding than conducting a small-size sample survey. However, emerging online data sources open up a new frontier for utilising these models. For example, door-to-door travel time and distance are now readily available from online map services, which facilitates the calculation of transport costs by origin–destination pair. Housing rental data from property listing websites enable the cost of housing to be accurately measured (in terms of spatial location) and incorporated into location choice models. The predictive power of some model applications has been empirically verified (Miller et al., 2012; Wan & Jin, 2017). However, few papers have explored how these established models may help to identify and potentially correct the errors in the matrices derived from big data.

This paper aims to address the research gap by empirically comparing a sample matrix derived from MP data for Shanghai, a Fratar-adjusted matrix based on the sample matrix, and a matrix generated by a discrete choice model using online open data, against observed travel statistics. A new error-correction strategy is then proposed, which is informed by the complementarity of the MP matrix and the model-estimated matrix.

3 Study Area, Data and Workflow

3.1 Geographical specification and official travel statistics

The study area is Shanghai in China. Despite being one of the most developed city regions in China, transport statistics from official sources are rather limited. The published data from the official Shanghai Comprehensive Travel Survey are all at an aggregate level. Access to microdata or detailed origin–destination data is virtually impossible even for academic use.

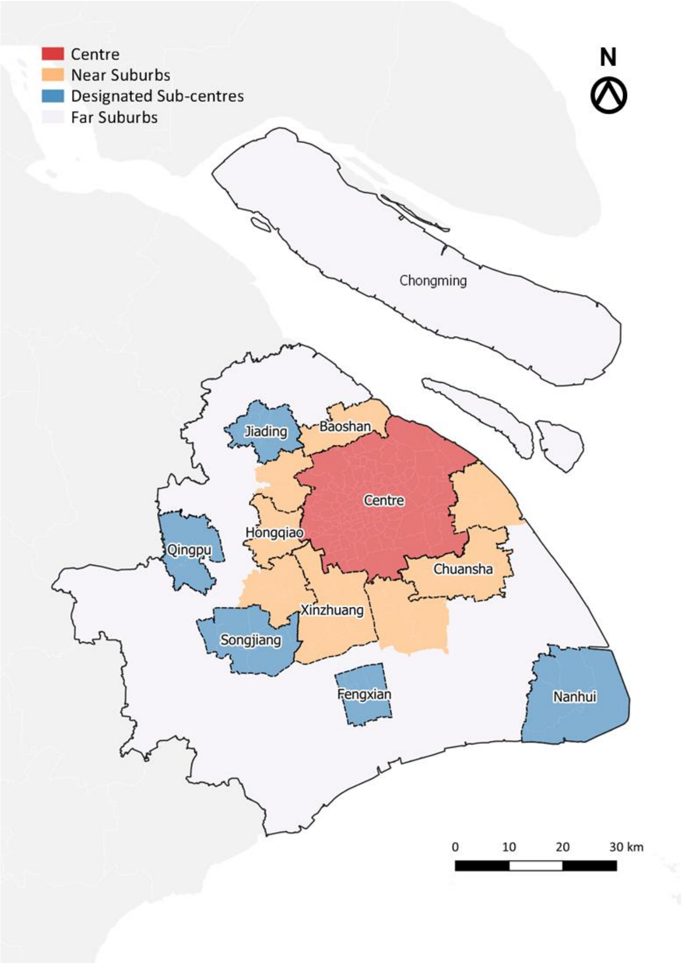

Figure 1 Zonal categorisation

A three-level zoning system was developed to represent the study area (see Figure 1). At the *micro-zonal* level, we followed the sub-district administrative boundary (i.e. *Jiedao*, which is the smallest unit of census geography in China). A total of 233 micro-zones were established, the area of which ranges from approximately one km² in the city centre to 148 km² in the far suburbs, with an average of 29 km². At the *meso-zonal* scale, we aggregated micro-zones into several clusters according to the local Comprehensive Plan (2017–2035). At the *macro-zonal* level, we further aggregated the meso zones into four categories: Centre (Ct), Near Suburbs (NS), Designated Sub-centres (DS) and Far Suburbs (FS).

The 2009 Shanghai Comprehensive Travel Survey is used as the benchmark for comparing alternative matrix estimations, which is the official survey closest to the observation year (2011) of the MP data. The released data do not include detailed origin–destination information and are mostly aggregate statistics, which typically include the city-level average travel distance, modal share and the percentage share of intra-zonal trips (i.e. origin and destination both within the same macro-/meso-zone) and share of to-Centre trips. Note that some of the location-specific survey data are sourced from a domestic academic article (Zhou & Chen, 2015). The travel survey data are presented in the matrix comparison section.

3.2 Mobile phone data

Anonymised cellular signalling records from a leading domestic operator (with a market coverage of over 50% in Shanghai) are used in this paper. Following the methods developed by Yang et al. (2019) and Yang (2020), we estimated the employment–residence location pair of MP users by identifying the most frequently connected signal towers during the daytime (9:00 a.m. to 5:00 p.m.) and night-time (midnight to 6:00 a.m.) over a period of two weeks in March 2011. This approach yielded 3.81 million regular commuters, which is approximately 33% of the total employment figure in 2011 Shanghai (see Table 1). A higher capture rate is observed for Designated Sub-centres for both place of residence and workplace, while the capture rate for the Centre is relatively low. This may be attributed to the more competitive market among mobile network operators (MNOs) in the city centre, and the fact that MNOs tend to introduce different market promotion strategies across locations. Note that the raw origin–destination matrix derived from the MP data features a sparse matrix containing a large number of zeros (27,488 zeros out of 233*233 = 54,289 origin–destination pairs). The different capture rates across geographical areas and the appearance of zeros in zone-pairs reflect the importance of validation in using location-based big data. Aggregated statistics from raw data would lead to biased observations and skewed results towards more represented groups.

Table 1 Identified number of commuters using mobile phone (MP) data (2010) versus observed data (2010) by macro-zone

	Commuters at home (Employed residents)			**Commuters at workplace (Employment)**		
	Mobile phone raw data (2011)	Census data[1] (2010)	Capture rate	Mobile phone raw data (2011)	Census data (2010)	Capture rate
Centre	1,294,632	5,045,878	25.7%	1,397,234	5,636,588	24.8%
Near Suburbs	1,005,688	2,747,090	36.6%	897,046	2,035,377	44.1%
Designated Sub-Centres	630,089	1,138,491	55.3%	632,029	1,219,467	51.8%
Far Suburbs	879,731	2,632,638	33.4%	883,831	2,672,664	33.1%
Total	3,810,140	11,564,096	33.0%	3,810,140	11,564,096[2]	33.0%

[1] The total number of employed residents and employment was obtained from the local census for 2010.

[2] The total number of employees has been adjusted to correspond to the total number of employed residents for producing the commuting matrix – commuters to/from outside Shanghai are hence omitted.

3.3 Network distance and time

The network distance and travel time for both intra- and inter-zonal trips are required for the model-based matrix estimation. We obtained the network distance and time by travel mode through online map services. The detailed method for processing the intra- and inter-zonal commuting distance and time is provided in the Supplementary Material.

3.4 House rents and average wages

To facilitate the development of the discrete choice model, data on house rents and workplace wages in Shanghai were also collected. Historical fine-scale house rents for 2010 are not available from conventional sources. Instead, we sourced the housing rental data from a domestic property listing website (lianjia.com) in December 2017, which includes a total of over 790,000 transaction records from 2015 to 2017 covering the entire modelling area. Zonal average annual house rent per unit was processed and weighted by the house type (i.e. number of bedrooms) for each micro-zone. To estimate zonal house rents for 2010, we first calculated the average house rent at the city level in 2010 (based on household expenditure data from the census) and then adjusted the 2017 house rents pro rata to match the 2010 city average. The average wage was estimated based on district-level household wage income retrieved from the census data (2010). The use of these data in the discrete choice model is discussed shortly.

3.5 Workflow

The following sections introduce the process of matrix estimation and blending based on the prepared data sets mentioned above. As shown in Figure 2, the differences between an

MP raw matrix, a mobile-phone-adjusted (MPA) matrix, and a model estimated (ME) matrix were compared using official statistics as a benchmark. The MPA and ME matrix were then blended based on their respective strengths, leading to a final (blended) matrix.

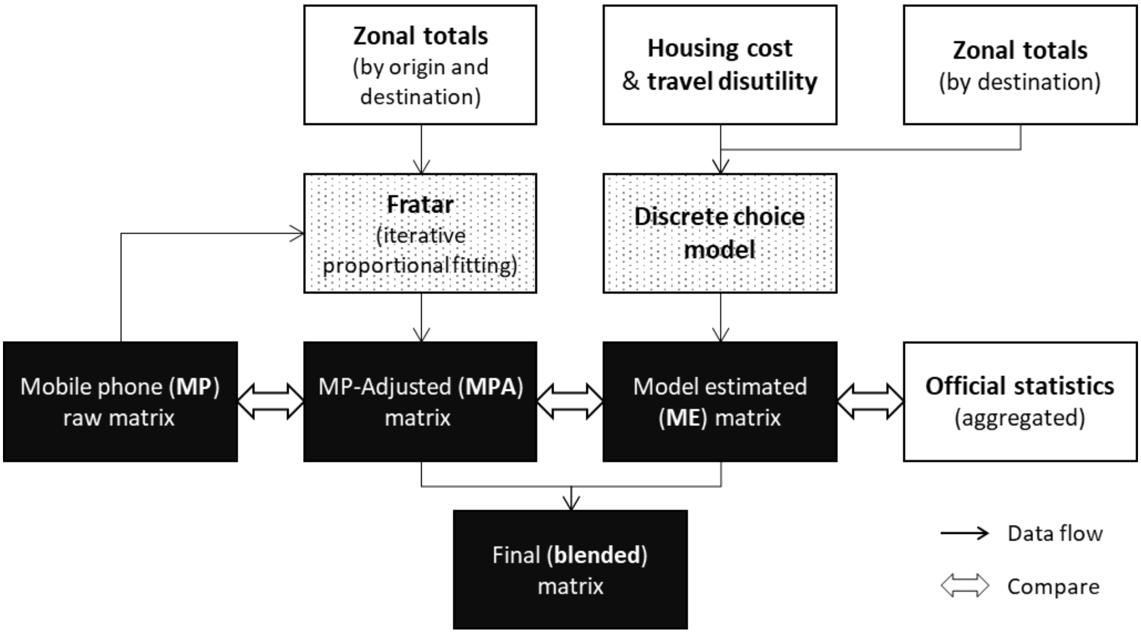
Zonal totals (by origin and destination)
Housing cost & travel disutility
Zonal totals (by destination)
Fratar (iterative proportional fitting)
Discrete choice model
Mobile phone (MP) raw matrix
MP-Adjusted (MPA) matrix
Model estimated (ME) matrix
Official statistics (aggregated)
Final (blended) matrix
Data flow
Compare

Figure 2 Workflow of matrix estimation and blending

4 Matrix estimation

In this section, we explain the method used for scaling the MP matrix and the model-based method respectively. This is followed by an investigation of estimation errors in the respective matrices by comparing the two estimated matrices against observed travel statistics.

4.1 Scaling the mobile phone matrix

Note that the methodological focus of this section is on scaling the sample matrix derived from the MP data to a full-population matrix. The MP matrix requires scaling because it does not reflect the full population, that is, the total number of commuters aggregated by the workplace and place of residence does not correspond to the observed totals (see Table 1). To carry out the scaling, we followed the standard Fratar–Furness method (Fratar, 1954; Furness, 1965), using the MP sample matrix as the seed matrix. The total zonal number of commuters by origin (i.e. employed residents) and by destination (i.e. workplace employment) is retrieved from the local census and used as the double constraints for the Fratar adjustment.

The formulation of the Fratar method was discussed in detail by de Dios Ortuzar (2011, p. 180).

The Fratar method is subject to one complication. If no trip is captured for a workplace or place of residence, that is, the cells along one row or column in the seed matrix are all zeros, the Fratar method would fail to meet the observed constraints for that row or column. We found that there were 11 such zones in our seed matrix, two of which are located in the Centre and three in each of the other macro-zones. We deemed that this is probably due to a sampling error in the original MP data. The Fratar-adjusted matrix is presented shortly, together with the method for deriving the ME matrix.

4.2 Model-based matrix estimation

The Fratar method requires an *a priori* seed matrix. When such an observed matrix is not available, a model-based method is often used to generate a commuting matrix that is compatible with observed travel statistics. In this paper, we propose a discrete choice model for simulating the residential location of workers in Shanghai. The discrete choice model incorporates emerging data sources such as travel time and distance, from online map services and housing rent data from property listing websites. The commuting flow from location i to j (M_{ij}^{Mod}) is defined as:

$$M_{ij}^{Mod} = E_j \frac{S_i e^{\lambda(-d_{ij}-\ln(C_{i|j})+Z_i)}}{\sum_k S_k e^{\lambda(-d_{kj}-\ln(C_{k|j})+Z_k)}} \qquad \text{Eq. 1}$$

where E_j is the observed zonal number of employees at workplace j; S_i is the total number of housing units in zone i; λ is a dispersion parameter to be calibrated; d_{ij} is the travel disutility for commuting between zone i and j; $C_{i|j} = r_i/w_j$ is the ratio between average housing rent at residence zone i (r_i) and the average wage at workplace j (w_j); and Z_i is a residual term for place of residence i, which is estimated subject to λ input.

The formulation of the proposed model includes two important features. First, the housing cost variable ($C_{i|j}$) addresses the trade-off between transport and housing in relation to residential location choices (Alonso, 1964; Mills, 1967). This is enabled by the emerging real estate listing website which provides open, geocoded housing rent data. The empirical results are presented shortly to demonstrate that the incorporation of $C_{i|j}$ would improve the goodness of fit of the model.

The second feature relates to the formulation of the travel disutility function (d_{ij}). The commuting catchment area of large city regions tends to have a radius of 50 km or more (BCT & BTRC, 2015; Vine et al., 2017). A travel disutility function that is linear to distance or time would have difficulty representing the variable cost elasticity of travel demand across a large city region such as Shanghai. A log-linear transformation (Fox et al., 2009; Jin et al., 2013) is therefore applied:

$$d_{ij} = \beta t_{ij} + (1-\beta)\ln t_{ij} - \beta \qquad \text{Eq. 2}$$

where t_{ij} is the average travel time from location i to j; $\beta = 0.01$ is a transformation coefficient, the value of which is informed by our own sensitivity tests. The log-linear transformation would produce a lower price elasticity for long-distance travel than a linear formulation (see Figure 3).

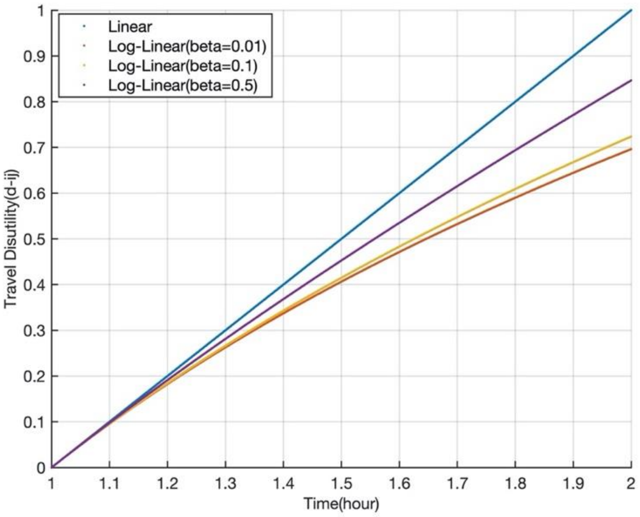

Figure 3 Log-linear versus Linear Travel Disutility (the line representing the linear function has been shifted vertically for readability)

Regarding model calibration, the dispersion parameter λ was estimated such that the estimated matrix reproduces the observed average commuting distance ($\widehat{ACD}$) as in the travel survey (8.2 km in 2009). The problem is defined as:

$$\min_{\lambda}(\widehat{ACD} - \frac{M_{ij}^{Mod} D_{ij}}{\sum_{m,n} M_{mn}^{Mod}}) \qquad \text{Eq. 3}$$

where D_{ij} is the network distance (km) between zone i and j. For a given λ, the zonal residual term Z_i can be estimated. The optimisation problem is defined as: $\min_{Z_i}(\widehat{R_\iota} - \sum_j M_{ij}^{Mod})$, where $\widehat{R_\iota}$ is the observed number of employed residents at place of residence i. The problem can be solved using a standard Newton–Raphson algorithm. The calibrated model can then produce a commuting matrix compatible with the observed ACD and remains doubly constrained by the zonal commuter totals.

Table 2 demonstrates the benefit of incorporating the housing cost variable ($C_{i|j}$) in terms of the goodness of fit of the location choice model. The combination of d_{ij} and $C_{i|j}$ can explain up to 76.7% of the variance in the observed data ($\widehat{R_\iota}$).

Table 2 Model comparison – with/without the housing affordability variable ($C_{i|j}$)

	R^2	Adjusted R^2	Root mean squared error	Ave. commuting distance (km)
Model without $C_{i\|j}$	0.699	0.698	0.0014	8.2
Model with $C_{i\|j}$	0.768	0.767	0.0013	8.2

Linear regression model: $y = ax + b$

y: observed percentage share of employed residents at place of residence i, given workplace j ($\widehat{R_{\iota|J}} / \sum_m \widehat{R_{m|J}}$)

x: modelled percentage share of employed residents at place of residence i, given workplace j ($M'^{Mod}_{i|j} / \sum_m M'^{Mod}_{m|j}$)

Note that the difference between M'^{Mod}_{ij} and M^{Mod}_{ij} is that the former is calculated without the residual term (i.e. $Z_i = 0$), aiming to test the explanatory power of endogenous variables d_{ij} and $C_{i|j}$. The M^{Mod}_{ij} is calculated with the calibrated residual term, hence $\sum_j M^{Mod}_{ij} = \widehat{R_\iota}$.

Note that d_{ij} and $C_{i|j}$ are of the same unit (hours) in our formulation. The combination also requires that the numeric range of d_{ij} and $C_{i|j}$ are mutually compatible; otherwise, one variable would completely dominate the outcome of the function. Table 3 presents the descriptive statistics of the two variables, which confirm the range compatibility.

Table 3 Descriptive statistics of transport disutility (d_{ij}) and housing affordability ($C_{i|j}$)

	Count	Max	Min	Mean	S.d.
Travel disutility (d_{ij})*	54,289	36.4	12.8	17.3	4.39
housing affordability (ln($C_{i\|j}$))	233	8.8	6.2	7.6	0.58

* No logarithm on d_{ij} because the log-linear transformation has been applied.

4.3 Comparison with observed data

In this section, we compare the raw matrix derived from the MP data, the Fratar-adjusted (MPA) matrix and the ME matrix against the observed travel statistics. We first compare the city-level ACD. The MP raw matrix and the MP-adjusted (MPA) matrix produce a much shorter ACD (5.1 and 6.5 km respectively) than the observed distance (8.2 km). The ME matrix uses the observed ACD as the input for calibration, and is therefore a perfect match.

The travel survey data also includes some macro-/meso-zone level trip statistics, although these are incomplete (see ‘travel survey’ column in Table 4). The comparison in Table 4 includes the raw MP sample matrix, the MPA matrix and the ME matrix. As the raw MP matrix does not conform to the observed trip-end constraints, the following discussion focuses on the MPA matrix and the ME matrix only. The table clearly shows that no method has absolute analytical superiority over another, as the direction and magnitude of errors vary across trip origins and destinations. For trips from the Centre, both the MPA matrix and the ME matrix perform well. For trips from the Near Suburbs (NS), the MPA matrix fits better with the survey data than the ME matrix, except for the NS_Xinzhuang-to-Centre trips. The ME matrix tends to overestimate the NS-to-Centre trips by a large margin.

For trips from the Designated Sub-centres (DS), the MPA and ME matrices show distinct error patterns. The MPA matrix tends to underestimate the DS-to-Centre trips and the magnitude of error (measured by the percentage difference based on the observed statistics) varies significantly across DS locations. The reason for such error patterns remains unclear. We have investigated whether the errors are related to the average distance to the Centre or the sample rate of the MP data in those locations, but no significant correlation was identified. By contrast, the ME matrix outperforms the MPA matrix for DS_Jiading and the Far Suburbs (FS) by a significant margin. The ME matrix also fits better for DS_Fengxian and DS_Nanhui, albeit the advantage is marginal in absolute terms.

The comparison between the MP and MPA matrix shows that the Fratar method may not only reduce estimation errors for certain locations (e.g. trips from the Designated Sub-centres) but also incur additional errors (e.g. trips from the Centre and Near Suburbs). The sources of errors and how to mitigate the additional errors incurred by the matrix manipulation remains a research gap in the literature. Comparing the two methods, our case study shows that the MPA matrix may underestimate long-distance trips (above 30 km, from the Designated Sub-centres and Far Suburbs to Centre), while the ME matrix tends to overestimate trips from the Near Suburbs to the Centre (around 20 km). However, additional data are required to identify whether the respective estimation errors are caused by location-specific or distance-specific factors.

Table 4 Comparison of estimated matrix against observed travel survey data – by trip origin and type

Trip by macro- & meso-zones as place of residence	**Travel survey**[1]		**Mobile phone (MP) sample matrix**				**MP-Adjusted (MPA) matrix**[2]				**Model estimated (ME) matrix**			
	%intra	%to-Ct	%intra	%to-Ct	Error %intra	Error %to-Ct	%intra	%to-Ct	Error %intra	Error %to-Ct	%intra	%to-Ct	Error %intra	Error %to-Ct
Centre (Ct)	93.1%	93.1%	94.4%	94.4%	1%	1%	95.4%	95.4%	2.4%	2.4%	91.4%[3]	91.4%	-1.9%	-1.9%
Near Suburbs (NS)	-	-	79.1%	15.5%			64.9%	2.8%			62.0%	4.9%		
NS_Baoshan	72.0%	18.0%	70.6%	19.1%	-2%	6%	71.3%	16.8%	-1.0%	-6.7%	52.3%	36.4%	-27.4%	102.1%
NS_Hongqiao	-	-	77.8%	15.6%			57.2%	29.9%			58.8%	29.5%		
NS_Xinzhuang	75.8%	13.8%	77.8%	17.3%	3%	26%	63.5%	32.3%	-16.3%	134.1%	59.7%	29.8%	-21.2%	115.7%
NS_Chuansha	-	-	74.9%	12.0%			57.7%	27.6%			59.5%	33.2%		
NS_Others	-	-	70.6%	13.1%			55.3%	28.3%			58.5%	29.0%		
Designated Sub-centres (DS)	-	-	88.0%	1.1%			85.2%	0.5%			85.8%	1.3%		
DS_Jiading	78.5%	5.2%	83.6%	1.4%	7%	-74%	87.1%	0.8%	10.9%	-84.8%	77.5%	7.5%	-1.3%	44.0%
DS_Qingpu	86.6%	2.9%	88.2%	0.9%	2%	-70%	87.2%	2.0%	0.6%	-32.3%	97.0%	0.9%	12.0%	-69.9%
DS_Songjiang	85.1%	4.1%	90.2%	1.2%	6%	-71%	84.6%	3.3%	-0.6%	-19.6%	83.8%	5.2%	-1.6%	26.6%
DS_Fengxian	82.5%	2.0%	85.2%	1.0%	3%	-50%	80.3%	1.8%	-2.6%	-8.2%	82.6%	1.9%	0.1%	-2.7%
DS_Nanhui	82.0%	3.8%	94.7%	0.4%	15%	-90%	86.2%	2.5%	5.1%	-35.1%	92.2%	2.8%	12.4%	-25.8%
Far Suburbs (FS)	87.0%	5.4%	89.8%	1.4%	3%	-74%	90.0%	2.4%	3.5%	-55.6%	89.2%	5.1%	2.5%	-6.4%
FS_Others	-	-	70.6%	1.5%			88.2%	2.8%			87.3%	5.9%		
FS_Chongming	-	-	99.8%	0.1%			99.6%	0.2%			99.6%	0.3%		

Note:

[1] Data by origin is incomplete.

%intra: percentage share of intra-trips (origin and destination within the same macro-/meso-zone) as a proportion of total trips from the same origin.

%to-Ct: percentage share of trips to the Centre as the workplace as a proportion of total trips from the same origin.

[2] The MP-Adjusted (MPA) matrix is the Fratar-adjusted matrix based on the MP sample matrix. Note that the MPA matrix includes a total of 11 zones that have zero commuters; the observed constraints for these zones thus are not applicable.

4.4 Identifying the complementarity between the MPA and ME matrix

One interesting finding from the comparison in Table 4 is that the direction of error (i.e. over- or under-estimation) of the MPA and ME matrix tends to be opposite, that is, an overestimation error by the MPA matrix is often associated with an underestimation error by the ME matrix. The opposite direction of error implies an analytical complementarity between the two methods. This complementarity seems discernible in two aspects of our study. First, the MPA matrix, although it matches the observed data well for some trip origins, contains 11 locations that do not meet the observed zonal constraints. This is likely to be caused by the inherent sampling errors in the raw MP data. The ME matrix thus provides an alternative source for correcting the sampling errors.

Second, the MPA matrix appears to better capture trips from the Centre and the Near Suburbs than the ME matrix, while the ME matrix outperforms the MP matrix for trips from some Designated Sub-centre origins and the Far Suburbs. The complementarity of the two methods seems evident – a combination of the two may help reduce the respective error and improve the representativeness of the estimated matrix, which neither method can achieve alone. The new method is discussed in the next section.

5 Combining the MPA matrix and ME matrix – a new approach

5.1 Method for matrix blending

Despite the complementarity discussed above, there is no generic method for combining the two matrices because the errors are case-specific. It is, therefore, reasonable to propose an error mitigation method that reflects the merits of the respective matrices. The procedure for blending the two matrices is explained as follows.

<u>Step 1</u>: blend the two matrices for selected trips, using the following formula:

$$M_{ij}^{Mix} = \alpha M_{ij}^{MPA} + (1-\alpha) M_{ij}^{ME} \qquad \text{Eq. 4}$$

where M_{ij}^{Mix} is the blended commuting flow from location i to j; M_{ij}^{MPA} and M_{ij}^{ME} are the Fratar-adjusted MP matrix and the ME matrix, respectively, and α is the mixing rate. For $0 < \alpha < 1$, the mixing rate denotes a blending of the two matrices, and $\alpha = 0 \; or \; 1$ indicates the use of a single matrix without blending. According to the error pattern of the respective matrix, we differentiated α values across origins and between *intra-trips* and *trips to Centre*. The initial value of α is informed by the comparison in Table 4. For example, for trips from the Centre, the MPA (ME) matrix produces a higher (lower) trip share than the observed data, and an initial value of $\alpha = 0.5$ is therefore set for blending the selected trips of the two matrices. For trips where the MPA matrix significantly outperforms the ME matrix, and the errors in the two matrices are of the same direction, an initial value of $\alpha = 1.0$ is set, indicating the use of the MPA matrix alone. Note that for trip origins that have no travel survey data, we follow the MPA matrix by default ($\alpha = 1.0$). This acknowledges the empirical nature of the MPA matrix. It should also be noted that, for the 11 zones that have no captured trips in the MPA matrix, trips from/to those zones are replaced by the ME matrix. The initial set of a is summarised in Table 6 (see values parentheses).

The blended matrix M_{ij}^{Mix} is, however, interim for two reasons. First, the blending would cause M_{ij}^{Mix} to no longer meet the observed zonal constraints. Second, the ACD of the M_{ij}^{Mix} is also likely to change due to the blending. Therefore, the M_{ij}^{Mix} matrix needs to be further adjusted to correspond to the observed statistics.

Step 2: apply a multinomial logit model to adjust the interim matrix M_{ij}^{Mix} according to the observed constraints, which include the total number of commuters by home location ($\widehat{R_i}$) and workplace ($\widehat{E_J}$) as well as the observed ACD ($\widehat{ACD}$). The logit model is defined as:

$$M_{ij}^{Mix\_Adj} = \widehat{E_J} \frac{e^{\lambda(\ln\left(M_{ij}^{Mix}\right)+Z_i)}}{\sum_k e^{\lambda(\ln\left(M_{kj}^{Mix}\right)+Z_k)}} \qquad \text{Eq. 5}$$

where $M_{ij}^{Mix\_Adj}$ is the adjusted commuting flow from location i to j; λ is a dispersion parameter to be calibrated; and Z_i is a residual term for place of residence i, which is to be calibrated conditional to λ. The optimisation problem of solving λ is the same as the one defined in Eq. 3 and can be solved accordingly. The proposed model is capable of adjusting the matrix to meet both observed zonal constraints and the ACD (due to the incorporation of λ), while the Fratar method can only achieve the former. If $\lambda == 1$, the proposed model would serve the same function as the Fratar method.

Step 3: refine the mixing rates (α) to produce a best-match $M_{ij}^{Mix\_Adj}$ matrix according to the observed statistics. Step 3 uses $M_{ij}^{Mix\_Adj}$ from Step 2 as the input, hence Step 2 is embedded into Step 3. Due to the embedded optimisation problem, the simultaneous estimation of α values for all trip origins and destinations is analytically difficult. We thus propose a trial-and-error approach, (1) introducing a unit change in α based on the initial value; (2)

examining the relative change in the matrix error; and (3) if the relative error is reduced, continuing to adjust the α until the matrix error is minimised for the trip category of interest; if not, reversing the unit change to a different direction and then doing step (2). The method is then applied to other trip categories sequentially.

It should be noted that the trial-and-error approach is subject to two complications. First, trips from different origins are interdependent due to exogenous double constraints. Adjusting the mixing rate for one trip origin may affect trips from other origins. The estimation process hence needs to be iterated several times until an overall optimum can be approximated. Our own tests suggest that starting from the trip origin with a larger number of commuters will often ease the iteration requirement. Second, the interdependence between *intra-trips* and *trips to Centre* suggests that an increase in *intra-trips* often leads to a reduction in *trips to Centre* and vice versa, albeit that the elasticity varies across origins. Our experience suggests that starting from the trip type with a relatively larger error may be helpful. In the next section, the performance of the final matrix is discussed.

5.2 Discussion of results

The performance of the final (blended) matrix is examined in Table 5, where both the survey data and the MPA matrix statistics are provided for reference. Regarding matching the observed data, the proposed approach delivers notable improvements – the relative error in the blended matrix is significantly reduced for most trip origins. One exception is the *intra-trips* for DS_Songjiang, where the error in the blended matrix is marginally higher than for the MPA matrix, although the absolute change is minimal.

Table 5 Comparison of trip statistics – MPA matrix vs Blended matrix

Trip by macro- & meso-zones as place of residence	**Travel survey[1] (2009, all trips)**		**MP-Adjusted (MPA) matrix**				**Blended matrix**			
	%intra	%to-Ct	%intra	%to-Ct	Error		%intra	%to-Ct	Error	
					%intra	%to-Ct			%intra	%to-Ct
Centre (Ct)	93.1%	93.1%	95.4%	95.4%	2.4%	2.4%	94.5%	94.5%	1.5%	1.5%
Near Suburbs (NS)	-	-	64.9%	2.8%			66.2%	2.8%		
NS_Baoshan	72.0%	18.0%	71.3%	16.8%	-1.0%	-6.7%	72.7%	17.9%	0.9%	-0.6%
NS_Hongqiao	-	-	57.2%	29.9%			48.3%	34.4%		
NS_Xinzhuang	75.8%	13.8%	63.5%	32.3%	-16.3%	134.1%	74.8%	14.4%	-1.3%	4.6%
NS_Chuansha	-	-	57.7%	27.6%			50.4%	30.7%		
NS_Others	-	-	55.3%	28.3%			52.1%	32.6%		
Designated Sub-centres (DS)	-	-	85.2%	0.5%			84.5%	1.0%		

Trip by macro- & meso-zones as place of residence	**Travel survey[1] (2009, all trips)**		**MP-Adjusted (MPA) matrix**				**Blended matrix**			
	%intra	%to-Ct	%intra	%to-Ct	Error		%intra	%to-Ct	Error	
					%intra	%to-Ct			%intra	%to-Ct
DS_Jiading	78.5%	5.2%	87.1%	0.8%	10.9%	-84.8%	82.0%	5.4%	4.4%	4.0%
DS_Qingpu	86.6%	2.9%	87.2%	2.0%	0.6%	-32.3%	86.7%	2.6%	0.1%	-10.3%
DS_Songjiang	85.1%	4.1%	84.6%	3.3%	-0.6%	-19.6%	84.4%	4.8%	-0.8%	17.7%
DS_Fengxian	82.5%	2.0%	80.3%	1.8%	-2.6%	-8.2%	83.5%	1.9%	1.2%	-2.7%
DS_Nanhui	82.0%	3.8%	86.2%	2.5%	5.1%	-35.1%	84.4%	3.8%	2.9%	1.2%
Far Suburbs (FS)	87.0%	5.4%	90.0%	2.4%	3.5%	-55.6%	88.5%	5.2%	1.8%	-4.4%
FS_Others	-	-	88.2%	2.8%			86.6%	6.0%		
FS_Chongming	-	-	99.6%	0.2%			99.0%	0.8%		

Note:
[1] Data by origin is incomplete.
%intra: percentage share of intra-trips (origin and destination within the same macro-/meso-zone) as a proportion of total trips from the same origin.
%to-Ct: percentage share of trips to Centre as the workplace as a proportion of total trips from the same origin.

Table 6 The mixing rate (α) – Adjusted values and initial values (in parentheses)

Trip by macro- & meso-zones as place of residence	**Trip type**			
	Intra-trips		Trips to Centre	
	MPA matrix[1] (α)	ME matrix[1] ($1-\alpha$)	MPA matrix (α)	ME matrix ($1-\alpha$)
Centre	0.95[2] (0.5)[3]	0.05 (0.5)	0.95 (0.5)	0.05 (0.5)
Near Suburbs (NS)[4]				
NS_Baoshan	- (1.0)	1.0 (-)	1.0 (0.5)	- (0.5)
NS_Xinzhuang	1.0 (1.0)	- (-)	1.0 (0.5)	- (0.5)
Designated Sub-Centres (DS)				
DS_Jiading	0.5 (0.5)	0.5 (0.5)	0.85 (0.5)	0.15 (0.5)
DS_Qingpu	0.5 (1.0)	0.5 (-)	1.0 (1.0)	- (-)
DS_Songjiang	- (1.0)	1.0 (-)	1.0 (0.5)	- (0.5)
DS_Fengxian	0.1 (0.5)	0.9 (0.5)	0.1 (-)	0.9 (1.0)
DS_Nanhui	1.0 (1.0)	- (-)	1.0 (-)	- (1.0)
Far Suburbs (FS)	- (-)	1.0 (1.0)	0.9 (-)	0.1 (1.0)

[1] MPA: Fratar-adjusted mobile phone data; ME: model estimation. [2] Adjusted values. [3] Initial values. [4] For trip origins with no observed data, mixing rates are not estimated, hence are omitted.

The values of the mixing rates are summarised in Table 6. For certain trip origins/types, the adjusted α may differ significantly from the initial value. One such extreme case is trip origin NS_Baoshan, where a reversed blending scheme (i.e. switching the use of a single matrix from MPA to ME) is seen for *intra-trips*. The explanation for this is that the initial set of α values are derived purely based on the horizontal comparison of the two matrices (MPA matrix and ME matrix) for the same trip origin without considering the interdependences between trip origins and types. The initial values of α are, therefore, only provisional.

Finally, it is arguable that the improvement delivered by the blended approach is only numerical but not substantive. It is indeed difficult without additional observed data to confirm that the improvement is substantive. However, as revealed by the comparison, the scaling exercise using the Fratar method does lead to a change in error patterns from the original MP matrix. These errors, if not corrected according to the aggregate travel survey data, would limit the policy use of the estimated matrix. The proposed method aims to mitigate the errors from two sources: (1) sampling errors in the raw MP data (e.g. zero-value cells); and (2) estimation errors incurred by the scaling process. Given the limited benchmark statistics, the new strategy is shown to significantly reduce such errors. Thus, we argue that the proposed method has its own methodological merits.

6 Conclusions

Despite extensive literature on using location-based big data for inferring origin–destination matrices, relatively few papers have compared the estimation errors between methods, nor how the identified errors could be mitigated to improve the representativeness of the estimated matrix. This paper represents an early step towards addressing this research gap. We compared a sample commuting matrix derived from MP data for Shanghai, a scaled matrix using the Fratar method, and a matrix generated by a discrete choice model against observed travel statistics. The comparison showed a complementarity between the MP matrix and the ME matrix. A new error mitigation strategy was thus proposed, namely blending the two matrices according to the respective error patterns, which leads to a significant reduction in estimation errors.

Our empirical findings suggest that the complementary use of big data and urban modelling techniques may lead to breakthroughs that neither could achieve alone. It is expected that the proposed method would establish a new standard for producing robust commuting matrix estimations based on aggregate travel survey data for cities where detailed flow data are not yet available. For fast-growing cities where the need for timely land-use and transport planning interventions is most pronounced, the proposed method can help to address the acute deficiency in commuting flow data from official sources. It is expected that the proposed method will set a new standard for correcting potential errors in big-data based flow estimation.

The research has several limitations. First, we do not consider the socioeconomic background or travel mode of the commuters. While the lack of socioeconomic information remains an inherent drawback of the data set, travel mode detection has seen much progress in the literature (Bachir et al., 2019; Huang et al., 2019), which could be explored when more mode-specific data become available. Second, the lack of additional survey data for Shanghai makes it difficult to identify the source of errors for the respective estimation method. However, access to detailed survey data is not insurmountable for some cities in China (see Hu et al., 2019). More empirical research is thus expected to investigate not only the travel patterns but also the source, magnitude and distribution of errors associated with estimated trip matrices and corresponding error mitigation methods.

7 References

Ahas, R., & Mark, Ü. (2005). Location based services - New challenges for planning and public administration? *Futures*, *37*(6), 547–561. https://doi.org/10.1016/j.futures.2004.10.012

Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, *58*, 240–250. https://doi.org/10.1016/j.trc.2015.02.018

Alonso, W. (1964). *Location and land use*.

Anas, A. (1983). Discrete Choice Theory, Information Theory and the Multinomial Logit and Gravity Models. *Transportation Research Part B: Methodological*, *17*(1), 13–23. https://doi.org/10.1016/0191-2615(83)90023-1

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., & Puchinger, J. (2019). Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, *101*(July 2018), 254–275. https://doi.org/10.1016/j.trc.2019.02.013

Batty, M. (1976). *Urban modelling*. Cambridge University Press Cambridge.

Batty, M. (2009). Urban modeling. In *International Encyclopedia of Human Geography. Oxford, UK: Elsevier*.

Batty, M., & Mackie, S. (1972). The calibration of gravity, entropy, and related models of spatial interaction. *Environment and Planning A*, *4*(2), 205–233.

BCT, & BTRC. (2015). *Beijing Travel Survey*. Retrieved from http://www.bjtrc.org.cn

Ben-Akiva, M. E. (1987). Methods to combine different data sources and estimate origin-destination matrices. *Transportation and Traffic Theory*.

Ben-Akiva, M., Macke, P. P., & Hsu, P. S. (1985). *Alternative methods to estimate route-level trip tables and expand on-board surveys.*

Bonnel, P., Hombourger, E., Olteanu-Raimond, A. M., & Smoreda, Z. (2015). Passive mobile phone dataset to construct origin-destination matrix: Potentials and limitations. *Transportation Research Procedia*, *11*, 381–398. https://doi.org/10.1016/j.trpro.2015.12.032

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, *26*, 301–313. https://doi.org/10.1016/j.trc.2012.09.009

Chen, C., Bian, L., & Ma, J. (2014). From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies*, *46*, 326–337. https://doi.org/10.1016/j.trc.2014.07.001

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, *68*, 285–299. https://doi.org/10.1016/j.trc.2016.04.005

de Dios Ortuzar, J., & Willumsen, L. G. (2011). *Modelling transport*. John wiley & sons.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*(4), 427–444.

Echenique, M., Crowther, D., & Lindsay, W. (1969). A spatial model of urban stock and activity. *Regional Studies*, *3*(3), 281–312.

Fox, J., Daly, A., & Patruni, B. (2009). Improving the treatment of cost in large scale models. *European Transport Conference, 2009*.

Fratar, T. J. (1954). Vehicular trip distribution by successive approximations. *Traffic Quarterly*, *8*(1).

Furness, K. P. (1965). Time function iteration. *Traffic Engineering and Control*, *7*(7), 458–460.

Gordon, J. B., Koutsopoulos, H. N., & Wilson, N. H. M. (2018). Estimation of population origin–interchange–destination flows on multimodal transit networks. *Transportation Research Part C: Emerging Technologies*, *90*(January), 350–365. https://doi.org/10.1016/j.trc.2018.03.007

Hu, L., Yang, J., Yang, T., Tu, Y., & Zhu, J. (2019). Urban spatial structure and travel in China. *Journal of Planning Literature*, *35*(1), 6–24. https://doi.org/10.1177/0885412219853259

Huang, H., Cheng, Y., & Weibel, R. (2019). Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*, *101*(January), 297–312. https://doi.org/10.1016/j.trc.2019.02.008

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, *40*, 63–74. https://doi.org/10.1016/j.trc.2014.01.002

Ji, Y., Mishalani, R. G., & McCord, M. R. (2015). Transit passenger origin-destination flow estimation: Efficiently combining onboard survey and large automatic passenger count datasets. *Transportation Research Part C: Emerging Technologies*, *58*, 178–192. https://doi.org/10.1016/j.trc.2015.04.021

Jin, Y., Echenique, M., & Hargreaves, A. (2013). A recursive spatial equilibrium model for planning large-scale urban change. *Environment and Planning B: Planning and Design*, *40*(6), 1027–1050. https://doi.org/10.1068/b39134

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PloS One*, *9*(6), e96180.

Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 134–142.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., … Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, *105*(3), 512–530. https://doi.org/10.1080/00045608.2015.1018773

Long, Y., & Thill, J.-C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems*, *53*, 19–35.

Long, Y., Zhang, Y., & Cui, C. (2012). Identifying commuting pattern of Beijing using bus smart card data. *Acta Geographica Sinica*, *67*(10), 1339–1352.

Lowry, I. S. (1964). *A model of metropolis*. Rand Corporation, Santa Monica, CA, USA.

McFadden, D. (1973). *Conditional logit analysis of qualitative choice behavior*.

McNeill, G., Bright, J., & Hale, S. A. (2017). Estimating local commuting patterns from geolocated Twitter data. *EPJ Data Science*, *6*(1), 24.

Miller, E. J., Farooq, B., Chingcuanco, F., & Wang, D. (2012). Historical validation of integrated transport-land use model system. *Transportation Research Record: Journal of the Transportation Research Board*, *2255*(1), 91–99. https://doi.org/10.3141/2255-10

Mills, E. S. (1967). An aggregative model of resource allocation in a metropolitan area. *The American Economic Review*, *57*(2), 197–210.

Office for National Statistics. (2017). Using mobile phone data to estimate commuting flows. Retrieved March 11, 2020, from https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/populationcharacteristics/researchoutputsusingmobilephonedatatoestimatecommutingflows#conclusions-and-next-steps

Papinski, D., & Scott, D. M. (2013). Route choice efficiency: An investigation of home-to-work trips using GPS data. *Environment and Planning A*, *45*(2), 263–275. https://doi.org/10.1068/a44545

Shen, Y., Kwan, M. P., & Chai, Y. (2013). Investigating commuting flexibility with GPS data and 3D geovisualization: A case study of Beijing, China. *Journal of Transport Geography*, *32*, 1–11. https://doi.org/10.1016/j.jtrangeo.2013.07.007

Steenbruggen, J., Borzacchiello, M. T., Nijkamp, P., & Scholten, H. (2013). Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, *78*(2), 223–243.

Ta, N., Zhao, Y., & Chai, Y. (2016). Built environment, peak hours and route choice efficiency: An investigation of commuting efficiency using GPS data. *Journal of Transport Geography*, *57*, 161–170. https://doi.org/10.1016/j.jtrangeo.2016.10.005

Vine, S. Le, Polak, J., & Humphrey, A. (2017). *Commuting trends in England 1988 - 2015*. Retrieved from www.gov.uk/dft

Wan, L., Gao, S., Wu, C., Jin, Y., Mao, M., & Yang, L. (2017). Big data and urban system model - Substitutes or complements? A case study of modelling commuting patterns in Beijing. *Computers, Environment and Urban Systems*, (October), 0–1. https://doi.org/10.1016/j.compenvurbsys.2017.10.004

Wan, L., & Jin, Y. (2017). Assessment of model validation outcomes of a new recursive spatial equilibrium model for the Greater Beijing. *Environment and Planning B: Urban Analytics and City Science*, 239980831773257. https://doi.org/10.1177/2399808317732575

Wang, Z., He, S. Y., & Leung, Y. (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, *11*, 141–155. https://doi.org/10.1016/j.tbs.2017.02.005

Wegener, M. (2004). Overview of land-use transport models. *Handbook of Transport Geography and Spatial Systems*, *5*, 127–146.

Wilson, Alan G. (1967). A statistical theory of spatial distribution models. *Transportation Research*, *1*(3), 253–269.

Wilson, Alan G. (1970). Inter-regional commodity flows: Entropy maximizing approaches. *Geographical Analysis*, *2*(3), 255–282.

Wilson, Alan Geoffrey, & Senior, M. L. (1974). Some relationships between entropy maximizing models, mathematical programming models, and their duals. *Journal of Regional Science*, *14*(2), 207–215.

Wismans, L. J. J., Friso, K., Rijsdijk, J., de Graaf, S. W., & Keij, J. (2018). Improving a priori demand estimates transport models using mobile phone data: A rotterdam-region case. *Journal of Urban Technology*, *25*(2), 63–83. https://doi.org/10.1080/10630732.2018.1442075

Yang, T., Jin, Y., Yan, L., & Pei, P. (2019). Aspirations and realities of polycentric development: Insights from multi-source data into the emerging urban form of Shanghai. *Environment and Planning B: Urban Analytics and City Science*, *47*(8), 1440–1455.

Yang, T. (2020). Understanding commuting patterns and changes: Counterfactual analysis in a planning support framework. *Environment and Planning B: Urban Analytics and City Science*, *46*(7), 1264–1280.

Yuan, Y., Raubal, M., & Liu, Y. (2012). Correlating mobile phone usage and travel behavior - A case study of Harbin, China. *Computers, Environment and Urban Systems*, *36*(2), 118–130. https://doi.org/10.1016/j.compenvurbsys.2011.07.003

Zhou, Xiang, & Chen, X. (2015). Relieving urban congestion: Polycentric development from a transportation perspective (in Chinese). *Shanghai Urban Planning Review*, *121*(2), 49–55.