Issue 20(4), 2020, pp. 1-21
https://doi.org/10.18757/ejtir.2020.20.4.5303

TU Delft OPEN

EJTIR

ISSN: 1567-7141
http://ejtir.tudelft.nl/

# Optimal synthesis of tours from multi-period origin-destination matrices using elements from graph theory and integer programming

**Haris Ballis**[1]
Department of Civil and Environmental Engineering, University of Cyprus, Cyprus.

**Loukas Dimitriou**[2]
Department of Civil and Environmental Engineering, University of Cyprus, Cyprus.

Nowadays, mobility modelling at individual level is receiving significant attention. Moreover, the technological advances in the field of travel behaviour analysis have supported and promoted the modelling paradigm shift to disaggregate methods such as agent/activity-based modelling Nonetheless, such approaches usually require significant amounts of detailed and fine-grained data which are not always easily accessible. The methodology presented in this paper aims to generate individual home-based trip-chains (i.e. tours) utilising aggregated sources of information, primarily, typical Origin-Destination matrices (ODs) and secondarily travel surveys. A suitable framework able to optimally identify 'hidden' tours in typical ODs is proposed and evaluated through its application on a set of multi-period OD matrices, covering an urban area of realistic size. This novel methodological framework synthesises the individual tours by combining and elevating advanced graph theory and integer programming concepts. The performance of the proposed methodology proves particularly encouraging since high estimation accuracy (greater than 85%) was achieved even for the most challenging examined test-case. The presented results provide positive evidence that information regarding travel behaviour on an individual level can be produced based on aggregated data sources such as OD matrices. This element is particularly valuable towards the analysis of mobility at the person-level, especially within the framework of agent-based modelling.

***Keywords***: *Origin-Destination Matrices, Tours, Graph Theory, Integer Programming*

[1] A: 1 Panepistimiou Avenue , 2109 Aglantzia, Nicosia T.: +357 22894000 E: ballis.theocharis@ucy.ac.cy
[2] A: 1 Panepistimiou Avenue , 2109 Aglantzia, Nicosia T.: +357 22894000 E: dimitriou.loukas@ucy.ac.cy

# 1. Introduction

*1.1 Motivation*

In the current era of information availability and the emergence of new modelling paradigms (e.g. agent-based, activity-based, data-driven modelling), travel demand modelling faces new opportunities. Moreover, advances in transport modelling, have enabled modelers to develop realistic and detailed representations of the mobility landscape accounting for the complex and dynamic interrelationships between people, activities, and the supply of transport services (Matthews *et al.*, 2007; Vogel and Nagel, 2013; Huynh *et al.*, 2014; Adnan *et al.*, 2016). Nonetheless, these emerging transport modelling approaches often require fine-grained information such as very detailed and extensive travel behaviour surveys (Bhat and Koppelman, 1999; Castiglione, Bradley and Gliebe, 2015) or extended location traces which may not be always accessible.

The traditional mean to organise travel demand information takes the form of Origin-Destination (OD) matrices. These matrices represent mobility as the total volume of trips between pairs of locations or areas, often referred as zones. Moreover, in many applications, OD matrices segregate trips according to their characteristics such as the time period of trip departure (e.g. AM-peak, PM-peak) or their trip-purpose (Home-Based-Work, Non-Home-Based-Shopping, etc.). Over the years, transport authorities and operators have allocated significant resources to develop and maintain similar matrices to support a plethora of decisions, related to urban planning, policy evaluation and transport infrastructure investments amongst others. Despite their extensive use, ODs face a serious limitation with regards to the representation of the interdependence between trips. This limitation, hinders their ability to capture travel behaviour in its full context and consequently limits the ODs' usefulness to relevant studies (Mcnally and Rindt, 2008). Transport modelling paradigms such activity-based modelling attempt to counter this limitation by expressing travel demand as sequences of activities linked by sequences of trips , often referred to as trip-chains (Bhat *et al.*, 2004; Pinjari and Bhat, 2011; Chu, Cheng and Chen, 2012). Although this data representation format is more flexible and better suited for the purposes of agent-based modelling (Ben-Akiva *et al.*, 2007), the capturing of the relevant information can prove an expensive, tedious and complex task (Gu, 2004; Ben-Akiva *et al.*, 2007). Advances in Information and Communications Technology (ICT) such as Mobile Network Data (MND) and GPS tracking has paved the way for more efficient and less expensive (passive) observation of individuals' travel behaviour (Gong *et al.*, 2014; Toole *et al.*, 2015; Anda, Fourie and Erath, 2016). Passively observed data, able to capture the trip-chains of individuals, are becoming abundant. Nowadays, the use of similar data sources in the field of transport modelling is gaining considerable ground. On the other hand, significant concerns have been raised regarding aspects such as '*anonymity preservation*' in cases where disaggregate personal data are used for similar studies. The proposed methodology can be utilised as a mechanism to eliminate data privacy concerns by reverse-engineering aggregate ODs to realistic individual tours.

Although, technological advances have enabled the observation and the expression of travel demand at a disaggregate level, the value of aggregated sources such as OD matrices should not be disregarded. Standard OD matrices can still provide a wide range of valuable information ranging from trip distribution patterns and estimates of total transport demand, to insights regarding the mobility motif. Moreover, typical OD matrices are still considered the most straightforward and widely used means to express travel patterns with no sign suggesting their complete replacement by other approaches. On the contrary, their wide use, in conjunction with their continuous development (Zhao, Rahbee and Wilson, 2007; Iqbal *et al.*, 2014; Bonnel *et al.*, 2015; Toole *et al.*, 2015; Tolouei, Psarras and Prince, 2017) indicates that their usability and value will not diminish in the foreseeable future. Despite their wide adoption and use, the aggregate nature of ODs hinders their use for the studying of trip-chaining and trip interdependency phenomena. However, the enhancement of OD matrices to include trip-chaining information could enable their

use in a wider range of applications and could considerably boost their value. Some of the potential use cases for the suggested methodology are presented below:

- Preparation of relevant input for agent-based/microsimulation models. For instance, trip-chaining information including the duration of stay between trips, stands for an essential input for microscopic agent-based simulation platforms as well as traffic assignment models (Kemper, 1990; Maerivoet, Sven; De Moor, 2006).
- Anonymisation of tours deriving from observable data sources (mobile phone/GPS traces). Aggregating these traces, to create ODs and subsequently use them as input to estimate tours could introduce randomness in the dataset without affecting the demand patterns as captured in the original ODs.
- Validation of OD matrices. Poor performance of the here suggested methodology on a set of ODs could be an indicator of inconsistencies related to trip-chaining.

### *1.2 Literature review*

The significance of trip-chaining for travel behaviour analysis has drawn considerable attention over the years (Thill and Thomas, 1987; Goulias and Kitamura, 1991; McGuckin and Murakami, 1999; Yue *et al.*, 2014). Although, the majority of the relevant studies has been based on detailed micro-samples (i.e. travel diaries), more recently urban sensing data sources (e.g. Mobile Network Data, GPS traces, smart card data etc.) have been also utilised for such purposes. As an example, Mobile Network Data (MND) have been utilized to observe daily trips-chains which were subsequently transformed into activity-sequences (Liu *et al.*, 2014). Similarly, smart-card data has been coupled with land-use information to impute activity sequences using continuous hidden Markov models (Han and Sohn, 2016). The use of hidden Markov chain methodologies to assign activity sequences deriving from micro-samples (surveys) to a synthetic populations has been also explored in recent studies (Saadi *et al.*, 2016).

A common factor across travel behaviour related studies is their dependence on fine-grained, personal travel behaviour data except for limited cases. For example, Balmer et al. (2006), proposed a framework capable of combining multiple sources of information, including OD matrices, to generate the demand, in the form of trip-chains, for large scale microsimulation. Preliminary approaches to synthesise sequences of typical (home-work-leisure-home) activity-chains, solely based on ODs, were firstly suggested by Ballis and Dimitriou (2018; 2019). This study presents a complete framework based on advanced graph theory and combinatorial optimisation concepts aiming at the conversion of aggregate ODs to fully tractable and disaggregate home-based trip-chains (i.e. tours).

The following section (Section 2) describes the methodology suggested by this study while Section 3 validates the methodology through a fully realistic case study. The last section of the paper (Section 4) discusses the conclusions and the future steps of the study.

## 2. Methodology

### *2.1 Outline*

This section focuses on the detailed description of a methodological framework aiming to create *tours*, using information obtained from multi-period OD matrices. For the purposes of this study and in order to avoid confusion, a *tour* is defined as a sequence of trips beginning and ending at a home location (Cirillo and Axhausen, 2002). Moreover, the individual trips consisting a tour will be referred to as *legs*. In its most basic form, the methodology can be utilised to convert an OD matrix, containing the total volume of individual trips between locations into multi-leg tours. Every additional dimension in the input OD (e.g. trip-purpose, transport mode, user group, etc.), can be exploited so that more detailed tours can be created. In the current study, the focus has been

placed on multi-period ODs which allowed for the allocation of departure time-period to each leg of the resulting tours. Nonetheless, the framework is generic and flexible enough to exploit every available dimension present in the input ODs.

The main principle behind the proposed tours' synthesis framework is that accurate OD matrices should include all the individual trips taking place in the covered area. Hence, there must exist a combination of trips which perfectly recreates the daily tours completed by the population. This assumption is supported by the observation that most people begin and end their daily activities at their home location (Sicotte, Morency and Farooq, 2017). Therefore, most of the trips in ODs should be able to be incorporated into diurnal tours. This assumption holds particularly true in cases where the OD matrices have derived from observational data sources (e.g. mobile phone/GPS data, etc.). These ODs are usually built by tracking the movements of individual people for consecutive days or even months. Therefore, ODs built from such sources are indeed formed as the aggregation of the trips within trip-chains and tours. Even in cases where OD flows have stemmed from modelling processes (e.g. typical 4-step models), and therefore flows are not entirely based on observed measurements, the fact that most trips within ODs should belong to tours, still holds true.

The followed approach to reconstruct multi-period ODs into individual tours is organised in two modules. The first module is responsible for the identification of all the plausible tours within ODs and is driven by a graph theory-based algorithm. The second module deploys an optimisation algorithm to identify the combination between the previously identified plausible tours which maximises the utilisation of trips in the input ODs. The two main steps involved in the process of tours' identification are presented in the following sub-sections.

### *2.2 Tours' identification module*

Identifying sequences of nodes (paths) originating and ending at the same node has been thoroughly studied in the field of graph theory (Diestel, 2017). These paths are formally known as *cycles* and a plethora of reliable and efficient methodologies has been suggested for their identification (Johnson, 1975). Based on their definitions, it can be observed that tours and cycles share multiple similarities. In fact, tours can be considered as a subset of cycles which follow some logical, spatial and/or temporal conditions. For instance, tours are considered valid only if they originate and end at a home location. Similarly, trips within tours must follow a chronological order. However, the previous conditions are irrelevant for the identification of cycles. For the purposes of this study, suitable algorithms drawn from the graph-theory context are combined with filtering mechanisms to first identify all cycles in an OD matrix and then filter out the non-valid tours within the identified cycles set. Prior to the application of graph theory algorithms, the conversion of the input ODs into a graph (network) is required. This can be straightforwardly achieved by expressing the OD zones as the network's nodes and the trips as the corresponding links. The case of converting multiple OD matrices into a single network can be addressed by multigraphs (Bollobás, 1999), where multiple links with different attributes can be used to connect the same pair of nodes.

Finally, in order to distinguish the graph-theory context form the transport modelling one, cycles will be thereafter referred to as *zone-sequences*. Based on the above-mentioned, the process of tours' identification can be split into the two following procedures:

- Identification of the plausible zone-sequences.
- Identification of the valid tours among all the previously identified zone-sequences.

*Identification of zone-sequences*

As noted earlier, a zone-sequence represents a tour only if the starting and ending zone is a home location. In order to avoid the identification of zone-sequences which do not hold this property,

the following approach is applied iteratively for all the origin zones. The network is stripped away of (a) Home-Based trips not linked to the origin zone and (b) Non-Home-Based trips linked to the origin zone. This renders the formation of invalid zone-sequences impossible and allows for a significant reduction in the number of the valid zone-sequences. The process is repeated iteratively until all the zones in the network have been traversed and consequently all the plausible zone-sequences have been identified. The procedure is illustrated in Figure 1.

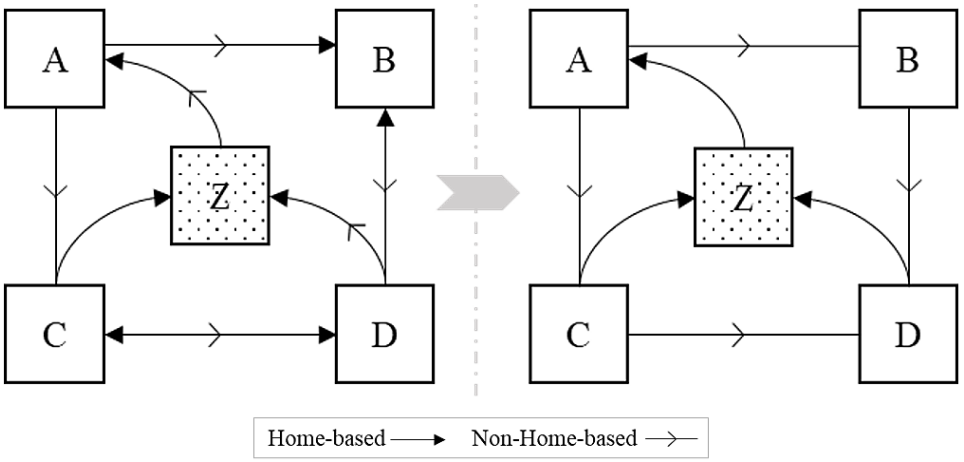


*Figure 1. Network filtering to ensure that only Home-Based zone-sequences originating from Z can be formed.*

*Identification of valid tours*

The outcome of the process described above is a set containing all the plausible zone-sequences which originate and end at a home location. Nonetheless, depending on the number and type of the available ODs, the same zone-pairs can be possibly connected by multiple trips, differentiated by aspects such as trip-purpose, time of departure, transport mode, etc. Therefore, each zone-sequence can result due to the possible permutations of trip attributes, into numerous tours which are not all sensible from a chronological and/or a travel behaviour perspective. Depending on the available information and the particularities of each case, a plethora of tour plausibility checks can be applied. For the purposes of this study, the focus has been placed on assuring the temporal consistency of tours. In simple terms, this is translated to the verification of the chronological order of the trips consisting a tour. As an example, if the initial leg of a tour takes place in the evening, none of the following trips can take place earlier in time (e.g. during the morning). This straightforward and reasonable validation check can significantly reduce the number of valid tours. In general, the methodology is flexible enough to allow for the application of any validation check. As an example, a validation check can be enforced to verify that tours initiated with a private car will also be terminated with the same mode. Similar rules can be examined in cases where certain types of activities (e.g. shopping or education) are not available during certain hours of the day. In general, the framework allows the introduction of any consistency check. The final set of tours surviving these validity checks is referred to as the *candidate tours set* and forms the input to the optimisation process.

### *2.3 Optimisation module*

*Mathematical formulation*

The following section presents the module responsible for the identification of the combination of candidate tours which maximises the use of the available trips in the input ODs. Equivalently, the module is responsible for the identification of the tours' combination which minimises the number of non-utilised trips. The mathematical formulation as well as insights regarding the theoretical extent of the problem domain are presented next.

Let $C$ be the set of candidate tours as identified in the previous step. The optimisation objective Eq. (1) aims to identify the utilisation of each tour in $C$ so that the number of unused trips as expressed in the input ODs is minimised. Inequality Eq. (2) guarantees that trips are not used more than what is described in the input ODs and Eq. (3) that the frequency of use for tours is non-negative. To sum up, the optimisation problem can be formulated as follows:

$$\min Z = \sum_{o}\left(\sum_{p_o}\left(\left|\sum_{c}\left(N_c D_c^{p_o}\right) - T_{p_o}\right|\right)\right) \quad (1)$$

subject to:

$$\sum_{c}\left(N_c D_c^{p_o}\right) - T_{p_o} \leq 0 \;\; \forall\, o \in \mathrm{O}, p_o \in P_o \quad (2)$$

$$N_c \geq 0, \forall\, \mathrm{c} \in \mathrm{C} \quad (3)$$

$$b_c^i = \frac{N_c G_c^i}{\sum_c N_c} \leq \delta_i \;\; \forall\, \mathrm{c} \in \mathrm{C}, i \in I \quad (4)$$

Nomenclature:

| | |
|---|---|
| $O$ | Set of multi-period ODs ($o \in O$) |
| $P_o$ | Set of zone-pairs in each $o$ ($p_o \in P_o$) |
| $C$ | Set of candidate tours ($c \in \mathrm{C}$) |
| $I$ | Set of distribution groups ($i \in I$) |
| $N_c$ | number of times each $c$ is utilised |
| $D_c^{p_o}$ | binary variable indicating whether $p_o$ is part of $c$ |
| $T_{p_o}$ | the number of trips between each $p_o$ |
| $G_c^i$ | binary variable indicating whether $c$ belongs in $i$ |
| $b_c^i$ | the probability of $c$ belonging in $i$ |
| $\delta_i$ | accepted error between the input and the modelled probability for each $i$ |

The above-mentioned optimisation problem is expressed as an integer linear program (ILP) and solved by the CPLEX Branch-and-Bound optimiser (IBM, 2020). Due to the modular nature of the methodology the currently used Branch-and-Bound algorithm can be substituted by other suitable optimisation techniques (e.g. Genetic Algorithms, Simulated Annealing, etc.).

*Optimality of solution*

It is acknowledged that the execution of the previously presented methodology can potentially return multiple combinations of tours which satisfy equally well the objective function. As with many other problems of combinatorial nature (e.g. OD estimation based on network loading information), no uniqueness criteria can be guaranteed. Nonetheless, the enhancement of the problem setting with information regarding the solution properties can guide the estimation towards the desirable direction. In the proposed estimation framework, such information can be incorporated both in the generalized error term used in the objective function as well as in the constraints set. The more detailed the constraints set, the more likely is to obtain a solution which best describes reality.

More specifically, in cases where additional information regarding the distribution of tours' characteristics is available (e.g. tours' length distribution), then this information can be included to refine the output. The introduction of distribution constraints is enabled by assigning to each of the candidate tours the corresponding distribution group ($i$) which best describes them. The application of Eq. (4) assures that the resulting combination of tours will follow the desired distribution and therefore the resulting solution will match more closely the expected one. Although the problem is relatively simple in its formulation, the identification of an optimal

solution can be hindered by the size of the search space (i.e. number of candidate tours). The relevant complexities and potential approaches to restrict the size of the candidate tours set are discussed in the following section.

### *2.4 Search space*

According to the characteristics of the current problem setting in terms of its properties as an integer programming problem, the size of the candidate tours (i.e. search space) mainly depends on two factors, namely:

1. the spatial resolution of the zoning system used to develop the OD matrices, and
2. the length of the tours defined as the maximum number of legs allowed in each tour.

*Effect of spatial resolution*

In order to quantify the above-mentioned spatial resolution, the concept of network density ($g$) is used as a proxy. In graph theory, network density, also known as gamma index (Rodrigue, Comtois and Slack, 2017), is defined as the fraction between the actual connections in the network and the possible ones. The arithmetic value of network density typically ranges from 0% to 100% but it can grow larger for the cases of multigraphs. The formula to calculate network density for directed graphs is presented in Eq. (5) where $l$ is the number of links and $n$ the number of nodes present in the network.

$$g = \frac{l}{n\,(n-1)} \tag{5}$$

A dense network allows the connection between multiple pairs of nodes, leading potentially to a significant increase in the number of plausible tours. As a result, this has a negative effect on the performance of the tours' identification process. To illustrate the effect of spatial resolution to the methodology, two simplified networks, characterised by resolutions (densities) along with their attributes are depicted in Figure 2 and Table 2.

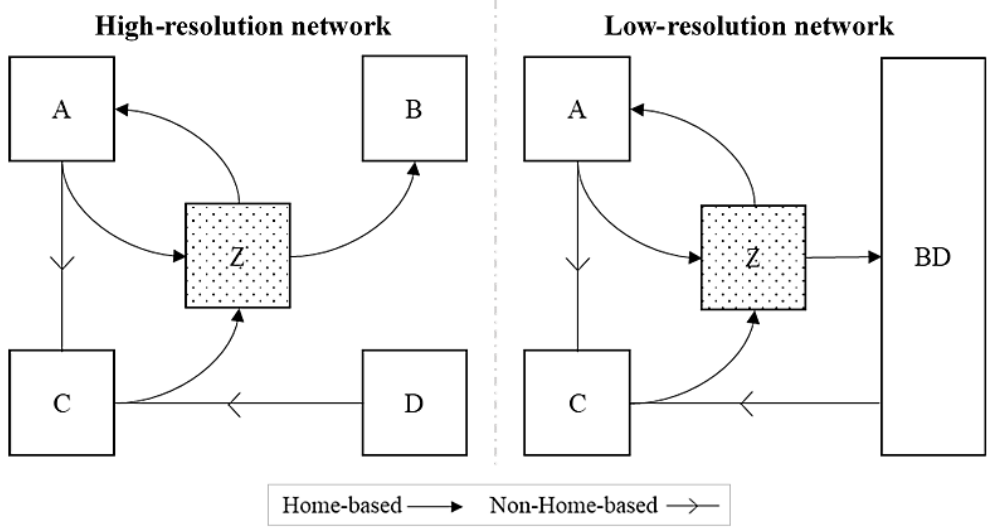


*Figure 2. Representation of the same OD matrix using a high-resolution (left) and low-resolution (right) network.*

**Table 1. Effect of network density on tours' identification process**

| Network resolution | Zones | Links | Density | Tours |
|---|---|---|---|---|
| **High** | 5 | 6 | 30% | (1) [Z, A, Z] (2) [Z, A, C, Z] |
| **Low** | 4 | 6 | 50% | (1) [Z, A, Z] (2) [Z, A, C, Z] (3) [Z, BD, C, Z] |

For this example, a low-resolution network is created by aggregating the zones of a high-resolution one. Although, the number of trips between these two cases remains the same, the effect of the spatial tessellation in the low-resolution scenario is significant. The network densities for the high- and low-resolution networks are 30% and 50% respectively. The most notable effect is that the reduction of spatial resolution (i.e. increase in density) enables the formation of tours which were impossible in the high-resolution case. For instance, in Table 1 it can be observed that the number of tours originating from zone Z increases from two to three. For large-scale networks the implications of using a low-resolution zoning system can be even more significant, leading to a many-fold increase of the number of plausible tours. The effect of network density will be evaluated at a greater extent at the case study section (Section 3).

*Effect of tours' length*

In the case of dense networks and/or of tours consisting of numerous legs (i.e. long tours) the number of potential tours and therefore the size of the search space can grow exponentially. As with many combinatorial optimisation problems, an exhaustive search for the identification of the optimal solution is not always feasible. Attempting to identify all the plausible zone-sequences can result in excessive processing times, especially in the case of dense networks. For example, the number of all plausible tours in a fully connected network (g=1) is calculated using Eq. (6). An example for a network of 140 nodes is also presented in Table 2.

$$Size(\mathrm{t}) = \sum_{l=1}^{L} \frac{(pz)!}{(pz-\mathrm{l})!} \tag{6}$$

where:

- $z$ the number of zones in the network
- $p$ the number of available time periods in the network
- $L$ the maximum number of legs allowed in tours

***Table 2. Number of possible tours in a fully connected network of 140 nodes***

| Maximum tour length (legs) | Possible tours | Cumulative sum |
|---|---|---|
| 2 | 560 | 560 |
| 3 | 313,040 | 313,600 |
| 4 | 174,676,320 | 174,989,360 |
| 5 | 97,294,710,240 | 97,469,386,560 |

As expected, the number of plausible tours grows rapidly. In order to improve the efficiency of the methodology and reduce the required processing times, the function responsible for identification of tours (Johnson, 1975) was modified to include a parameter limiting the maximum number of zones within a tour. This modification is also sensible from a travel behaviour perspective since relevant travel diaries and studies can verify that most travellers do not complete tours with very high number of legs. For instance, in the National Travel Survey of UK (Department for Transport, 2017) no tour exceeded thirteen legs while the majority (≥ 97%) of tours ranged between two to five.

*Search space reduction*

Despite the advances in processing power and in integer programming optimisation algorithms, such a vast number of candidate tours cannot be efficiently handled by the currently available optimisation tools. Nonetheless, steps can be taken to drastically reduce the number of plausible tours, such as the:

- exploitation of the connectivity between zones as expressed in the ODs
- filtering of cycles not originating at a home location and
- application of chronological consistency checks.

Based on the results obtained from a relevant test case presented in Section 3, the application of the three above-mentioned filters can reduce the 5-leg tours from almost 97.5 billion to approximately 2 million unique tours. Regardless of this significant reduction, a figure of that size can still prove challenging to the currently available optimisation solvers. Nonetheless, the available search space can be further reduced by exploiting information regarding the expected characteristics of tours. As an example, the maximum expected duration of tours can be utilised to exclude long tours from the set of candidates. The same principle can be applied to tours belonging into any distribution group with insignificant or small frequency. If for example, a small fraction of tours consists of *n* legs, then this group can be omitted without a considerable effect on the accuracy of the result. In real applications, observed data can suggest the exclusion of multiple distribution groups, leading to a substantial reduction of the problem domain.

## 3. Case study

### *3.1 Experimental design*

The objective assessment of the proposed framework's performance requires a fully observed and controlled test-case. For that reason, a realistic set of tours (referred to as *original tours*) is prepared based on information obtained from a relevant travel survey (Department for Transport, 2017). For the purposes of this study, the original tours are designed to replicate travel behaviour of individuals residing in large urban areas. It is important to stress that the *original* tours are solely developed to form the validation set which will be later used to compare with the methodology's results. After their creation, the *original* tours are aggregated into a set of multi-period OD matrices (referred to as *original ODs***)**. At this stage the direct reverse-engineering of the *original* ODs to the *original* tours is not possible since all the trip-sequencing information is lost due to aggregation. The *original* OD matrices, stripped of trip-chaining information, are fed to the methodology with the purpose to estimate a set of *modelled tours*, able to represent the travel demand as described in the *original* ODs. Ultimately, the individually *modelled* tours are compared side-by-side with the *original* tours.
To sum up, the experimental design can be summarised by the following steps:

1. Preparation of the original tours set (here, based on survey data)
2. Aggregation of tours into multi-period ODs (referred to as *original ODs*)
3. Application of the suggested methodology to the *original* ODs
4. Estimation of the *modelled* tours
5. Validation between the *observed* and the *modelled* results (both for ODs and tours).

### *3.2 Evaluation scenarios*

*Explored parameters*

This section presents a set of scenarios designed to fully assess the applicability of the methodology as well as the accuracy of the results. The scenarios are differentiated by the following parameters:

- the spatial resolution of the zoning system, and
- the enforcement or not of distribution constraints

The following evaluation scenarios will attempt to quantify this effect by applying the methodology to the same set of ODs expressed a) in a high-resolution zoning system and b) in a low-resolution zoning system. Another element which can substantially affect the overall process is the constraints relevant to the observed tours' distribution. For this purpose, two additional parameters have been included to further enrich the evaluation scenarios, namely:

1. the enablement of the joint distribution constraints and
2. the number and type of the joint distribution's dimensions.

The inclusion of the first parameter aims to examine the effect of the joint distribution on the resulting tours. The focus is to evaluate whether an unconstrained solution, in terms of distribution, describes the *original* OD matrix more accurately at the expense of identifying unrealistic tours (e.g. high distribution of long tours). The second parameter examines the potential benefits of increasing the information included in the provided distribution by including more dimensions to the joint distribution. In the first instance, a distribution describing the tours in terms of their number of legs and their total travel time is evaluated. In the second instance, the previous joint distribution is enriched with information regarding the time period of departure for each of the legs in tours. The purpose is to evaluate the effect of providing additional, in this case temporal, information to the solution.

*Summary of the evaluation scenarios*

The parameters forming the exploration space are presented below:

- Spatial resolution: High and Low
- Distribution constraints: Enabled or Disabled
- Joint distribution dimensions:
    1. Legs: Number of legs in tours
    2. Travel time: Tour's duration grouped in bins of 15-mins
    3. Time periods: Time periods of departure for each tour's legs

Based on these dimensions, the defined eight scenarios are presented in Table 3.

**Table 3. Summary of scenarios' parameters**

| Scenario # | Spatial Resolution | Distribution dimensions | Distribution constraints |
|---|---|---|---|
| 1 | High | Legs / Travel time | Disabled |
| 2 | High | Legs / Travel time | Enabled |
| 3 | High | Legs / Travel time / Time periods | Disabled |
| 4 | High | Legs / Travel time / Time periods | Enabled |
| 5 | Low | Legs / Travel time | Disabled |
| 6 | Low | Legs / Travel time | Enabled |
| 7 | Low | Legs / Travel time / Time periods | Disabled |
| 8 | Low | Legs / Travel time / Time periods | Enabled |

### *3.3 Data inputs*

*The zoning system*

The first step of the evaluation process entails the definition of a zoning system which will be used to express the sequence of zones followed by each tour. As it has already been pointed out, the zoning system can significantly affect the tours' resolution as well as the complexity of the overall problem (Section 2.4). To simplify the process, the required for the evaluation zoning systems were developed based on UK standard census geographic boundaries. As of 2011 UK is divided in 34,753 Lower layer Super Output Areas (LSOAs) with an average population of 1,500 in each. The

aggregation of LSOAs to larger groups results in a coarser census geographic boundary called Middle layer Super Output Areas (MSOAs) with a mean population of around 7,200. Since MSOAs emerge as pure aggregation of LSOAs therefore a direct mapping between them does exist.
For the purposes of our analysis, a *high-resolution* zoning system was created by sub selecting 470 LSAOs. Aggregating these 470 LSOAs to the corresponding MSOAs, resulted in a *low-resolution* zoning system of 140 zones. The conversion from the *high-* to the *low-resolution* zoning system led to a 70% reduction in the number of zones with a subsequent eightfold increase of the network density (Table 4).

**Table 4. Summary of zoning-systems used for the synthesis of the original tours**

| Spatial Resolution | Based on | Zones | Links | Network density (%) |
|---|---|---|---|---|
| **High** | LSOAs | 470 | 7,596 | 3.20 |
| **Low** | MSOAs | 140 | 6,474 | 25.2 |

*Original tours*
As mentioned before, a set of realistic tours has been used to create the input OD matrices. We refer to the respective tours and ODs as *original*. These *original* tours were synthesised in accordance to relevant information obtained from the UK National Travel Survey (Department for Transport, 2017). The relevant information in the survey were expanded to create an adequately large number of tours, suitable for the creation of considerably sized multi-period ODs. The distributions of the produced tours in terms of (a) number of legs (b) total travel-time and (c) time period of departure for each leg are presented in Figure 3 and Figure 4. Four The time periods were used for the purposes of this analysis are presented in Table 5.

**Table 5. Definition of available time periods**

| Time period | Covered period |
|---|---|
| AM | 07:00 – 10:00 |
| IP | 10:00 – 16:00 |
| PM | 16:00 – 19:00 |
| OP | 19:00 – 07:00 |

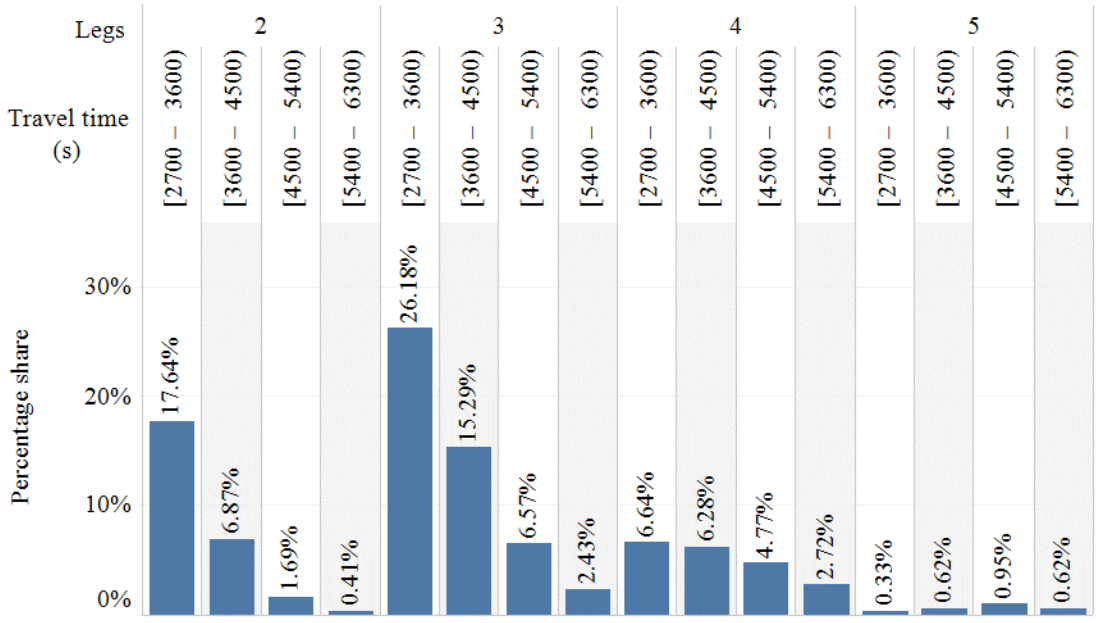


*Figure 3. The applied joint distribution for the synthesis of tours by number of legs and total travel time.*

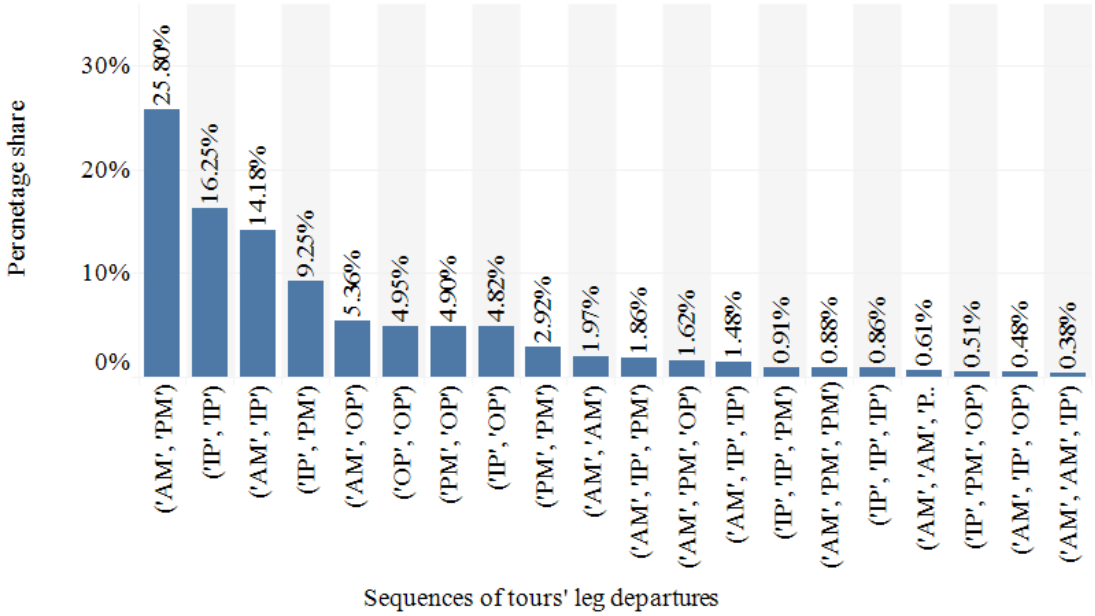


*Figure 4. The applied temporal distribution of legs' departures of the original tours*

The evaluation of the effect of spatial resolution to the methodology was achieved by expressing the original tours in two different sets depending on the resolution used to define them. The first set of original tours is based on the *high-resolution* network zoning system and therefore all zone-sequences are expressed as sequences of LSOAs. The second set is based on the *low-resolution* zoning system and is formed by replacing the LSOAs in each of the previously identified zone-sequences by the corresponding MSOAs. Following the previously described methodology returns two sets of tours, differentiated by the zoning system used to define them. A summary of these sets is presented in Table 6.

**Table 6. Summary of the original tour sets**

| Tour set | Spatial Resolution | Total tours | Unique tours | Ratio |
|---|---|---|---|---|
| **1 (LSOAs)** | High | 3,659 | 2,864 | 1.27 |
| **2 (MSOAs)** | Low | 3,323 | 2,696 | 1.23 |

*Original OD matrices*

Aggregating the legs of the *original* tours results in a set of *original* OD matrices (Table 7). Due to this aggregation, all information regarding trip-chaining is lost. However, the *original* ODs still enclose information regarding the origin, destination, trip-purpose and time period of departure for all the individual trips completed in the study area. Two sets of multi-period OD matrices differentiated by the underlying zoning system were developed. The previous steps present the procedure to obtain the required input to the methodology. The application of the methodology to the *original* OD matrices results in the set of *modelled* tours.

**Table 7. Summary of the original OD matrices**

| Trip-purpose | Home-Based (HB) | | | | Non-Home-Based (NHB) | | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| Time period | AM | IP | PM | OP | AM | IP | PM | OP | |
| **High-resolution** | 4,654 | 4,164 | 3,574 | 2,200 | 450 | 2,244 | 1,142 | 242 | 18,670 |
| **Low-resolution** | 3,962 | 4,000 | 3,174 | 2,248 | 310 | 1,886 | 1,038 | 222 | 16,840 |

### *3.4 Validation*

The following section presents the results from the application of the methodology for the evaluation scenarios. All scenarios were completed on an 8-core i7@2.6GHz CPU with 16GB RAM while the required processing time for each scenario is presented in Table 8. As it can be noticed, the use of low spatial resolution (i.e. increase of network density) leads to considerably higher identification processing time requirements (tenfold increase). On the other hand, the increase of the required optimisation processing time does not exceed 60%. Nonetheless, the overall processing time is still reasonable (less than 1.5 days) and can be reduced by using computing systems with additional cores.

**Table 8. Processing time requirements per scenario**

| Scenario No | Spatial Resolution | Identification processing time (sec) | Optimisation processing time (sec) | Total processing time (sec) |
|---|---|---|---|---|
| **1** | High | 9,690 | 1,150 | 10,840 |
| **2** | High | 9,690 | 960 | 10,650 |
| **3** | High | 9,690 | 950 | 10,640 |
| **4** | High | 9,690 | 860 | 10,550 |
| **5** | Low | 105,530 | 1,820 | 107,350 |
| **6** | Low | 105,530 | 1,760 | 107,290 |
| **7** | Low | 105,530 | 1,650 | 107,180 |
| **8** | Low | 105,530 | 1,480 | 107,010 |

The assessment of the methodology is completed based on a series of comparisons between the expected (*original)* and the estimated (*modelled)* results. Firstly, the evaluation at aggregate level is performed by comparing the *original* and *modelled* OD matrices, in terms of their overall resemblance. Secondly, the comparison of the distributions describing the characteristics of the *original* and the *modelled* tours is completed. Finally, the individual *original* and *modelled* tours are compared at a disaggregate level allowing for an in-depth appraisal of the methodology's accuracy.

*Comparison of ODs*

The first level of assessment is related to the methodology's ability to accurately reproduce the *observed* OD matrices. As it has already been explained, the *original* and the *modelled* OD matrices derive from the aggregation of the *original* and *modelled* tours respectively. The most straightforward way to achieve such an assessment is to compare the total number of trips, possibly segmented in different categories (e.g. trip-purpose). A summary of the total trip differences is presented in Table 9. Based on these results, it can be observed that the methodology manages to perfectly reproduce the *original* OD matrices in six out of eight scenarios while the difference is less than 0.3% for the rest.

**Table 9. Comparison between the original and the modelled OD total trips' difference**

| Scenario No | Spatial Resolution | Distribution Dimensions | | Distribution Constraints | Percentage Difference (HB trips) | Percentage Difference (NHB trips) |
|---|---|---|---|---|---|---|
| **1** | High | Legs/Travel time | | Disabled | 0% | 0% |
| **2** | High | Legs/Travel time | | Enabled | -0.04% | -0.04% |
| **3** | High | Legs/Travel periods | time/Time | Disabled | 0% | 0% |
| **4** | High | Legs/Travel periods | time/Time | Enabled | 0% | 0% |
| **5** | Low | Legs / Travel time | | Disabled | 0% | 0% |
| **6** | Low | Legs / Travel time | | Enabled | 0% | 0% |
| **7** | Low | Legs/Travel periods | time/Time | Disabled | 0% | 0% |
| **8** | Low | Legs/Travel periods | time/Time | Enabled | -0.29% | -0.29% |

Apart from the comparison concerning the total trips between the *original* and the *modelled* OD matrices, a pairwise comparison of the ODs' cells is presented in Figure 5. In this graph a scatter plot comparing the *original* versus the *modelled* trips for each of the compared ODs' cells is presented. Each point on the graph represents a trip:

- between two specific zones (e.g. Zone1 to Zone2)
- with a specific trip-purpose: (e.g. Home-Based)
- executed in a single time period (e.g. AM).

As it can be observed, the goodness-of-fit is excellent for all scenarios with the $R^2$ ranging from 1 to 0.99 indicating an almost perfect correlation. This correlation can be explained by the fact that the algorithm managed to utilise almost all trips present in the input ODs, therefore the pairwise differences between the *observed* and the *modelled* OD should be minimal.

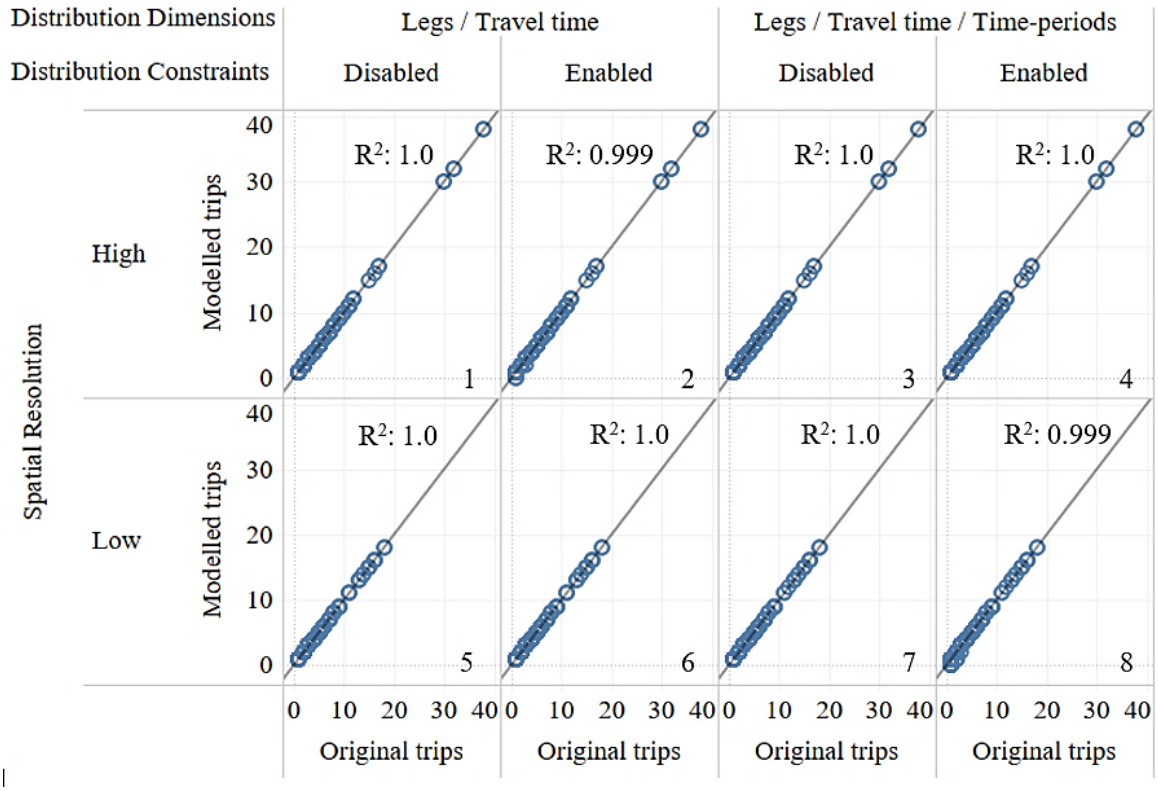


*Figure 5. Pairwise comparison of the original and the modelled OD cells*

*Comparison of tours' distributions*

The next set of comparison focuses on the distribution of attributes between the *original* and *modelled* tours. Tours are classified into different distribution groups with respect to their attributes. For scenarios number 1, 2, 5 and 6 the attributes under consideration were the number of legs as well as the total travel time of each tour, while for scenarios number 3, 4, 7 and 8 the dimension of legs' departure time period was also introduced. For each scenario, the share of *modelled* tours in each distribution group was compared with the respective share for the *original* tours. For brevity, only the first set of scenarios (scenarios number 1, 2, 5, 6) and only the groups with a share greater than 1% are presented. This is due to the large number of resulting distribution groups (33 for scenarios number 1 to 4, and 325 groups for scenarios number 5 to 8). Finally, the comparison examines the distribution groups separately for the *high-resolution* (Figure 6) and *low-resolution* (Figure 7) scenarios.

The results for the *high-resolution* scenarios indicate a particularly close fit between the *original* and the *modelled* distributions even in the cases where the distribution constraints are not enforced. As it can be observed, none of the compared distribution groups deviates more than 0.7% from the respective distribution of the *original* tours. The minor differences can be explained by the *high-resolution* of the *original* OD which consequently limits the problem domain. In other words, due to the *high-resolution* of the *original* ODs, the combinations of tours which can reproduce the initial ODs are limited and as a sequence the attributes of the *modelled* tours are very similar to those of the *original* ones. Therefore, the optimisation routine manages to identify a (near) optimum solution both in terms of total trips used as well as in terms of tours' characteristics without the enforcement of additional constraints.

Similar conclusions can be drawn for the *low-resolution* scenarios although in this case the benefits of enforcing the distribution constraints are more obvious. The solution space in these cases is wider and thus the available solutions can deviate more compared to the *high-resolution* scenarios. For instance, the difference for distribution group 8 between the *original* and *modelled* results in Scenario 5 is 2.2%. Even if the figure is still relatively low, it's considerably larger than the respective figure for the high-resolution scenario (Scenario 1). It is expected that the differences, in cases of low-resolution scenarios, will be greater and thus the inclusion of distribution constraints to limit the available search space is recommended.

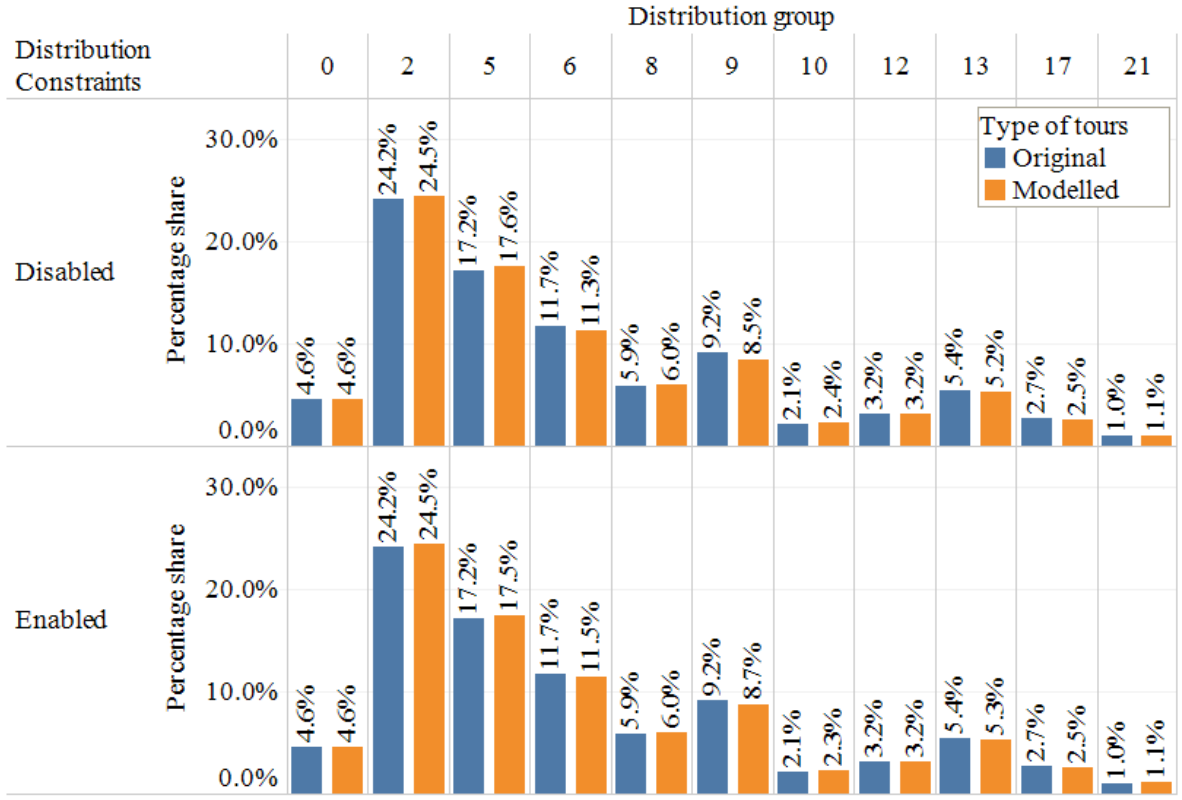


*Figure 6.* *Comparison of distributions between original and modelled tours with a share greater than 1%. (Scenarios 1-2)*

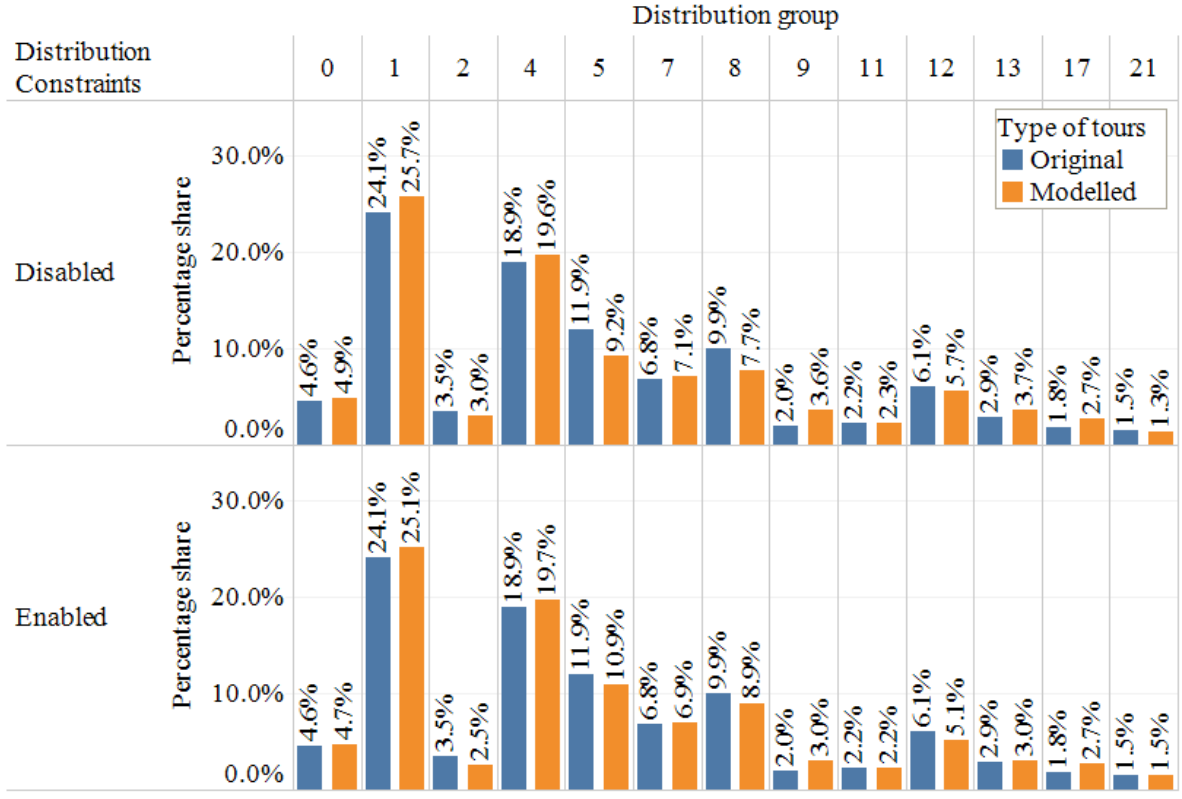


*Figure 7.* *Comparison of distributions between original and modelled tours with a share greater than 1%. (Scenarios 5-6)*

*Comparison of tours' characteristics*

A final stream of analysis entails the comparison between *original* and *modelled* tours at an individual level. The first type of assessment compares the zone-sequences between the *original* and *modelled* tours. Each point in Figure 8 represents a unique zone-sequencing (e.g. *[Zone-1, Zone-2, Zone-1]*) and the values on x and y axes represent the number of times each zone-sequence has been used by the *original* and the *modelled* tours respectively. As it can be seen, the correlation is high for both the *high-* and the *low-resolution* scenarios. Moreover, the effect of network's spatial resolution to the solution is once again noticeable. Although, the correlation is still high for the *low-resolution* cases (scenarios numbers 5-8), the respective $R^2$ is significantly lower compared to the *high-resolution* cases (scenarios numbers 1-4). Including the distribution constraints seems to have a positive outcome since an increase of the resulting goodness-of-fit is noticed.

The second and final type of assessment compares the time period departure profiles of trips within the *original* and the *modelled* tours. A time period departure profile is expressed as the sequence of the time periods of departure for the trips consisting a tour (e.g. *[AM, IP, OP]*). The scatter plot presented in Figure 9 depicts the number of observations for all the different time period departure profiles between the *original* and *modelled* tours. As it can be observed, the fit is very close for all scenarios, with the $R^2$ ranging between 0.999 and 0.960.

Based on the results obtained from the comparison of both the zone-sequences and the time period departure profiles, it can be deducted that the error related to the matching of *original* and *modelled* tours is mostly influenced by the misalignment of zone-sequences. This is supported by the fact that the corresponding goodness-of-fit for the time period of departure profiles (Figure 9) is higher compared to the one for zone-sequences (Figure 8).

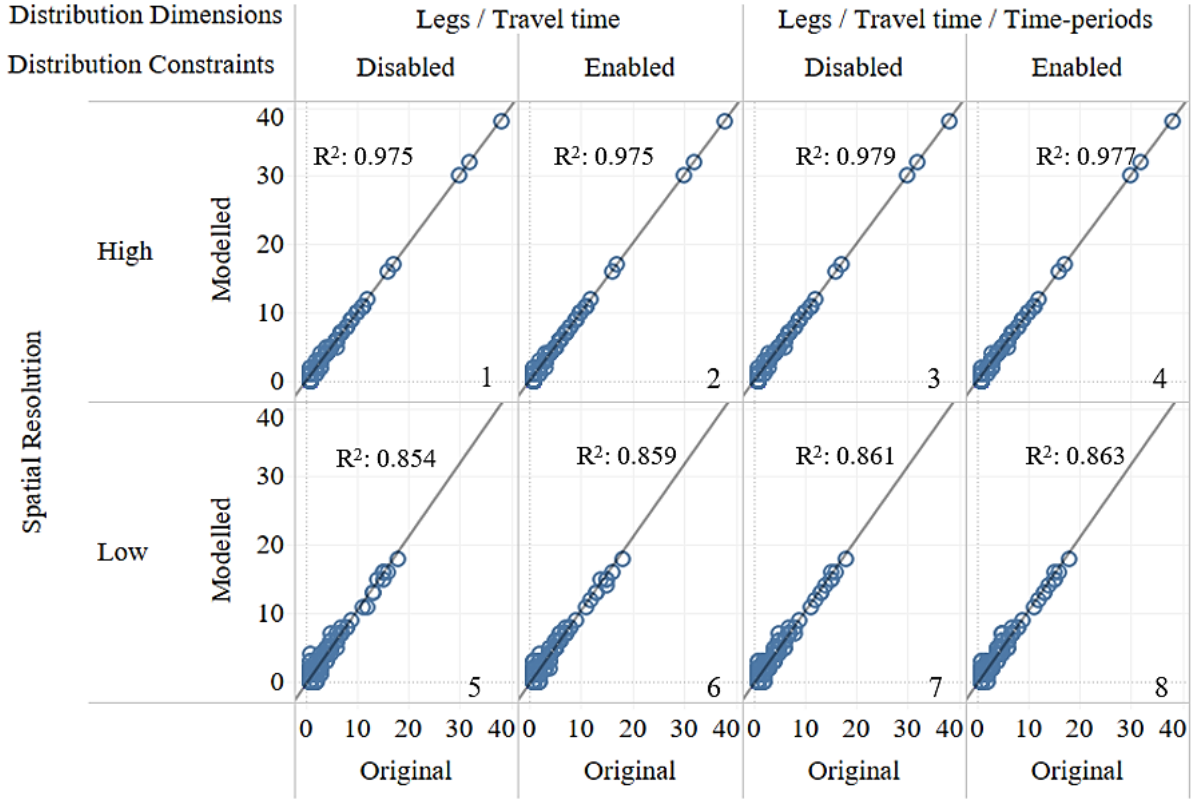


*Figure 8. Comparison of zones-sequencing between original and modelled tours*

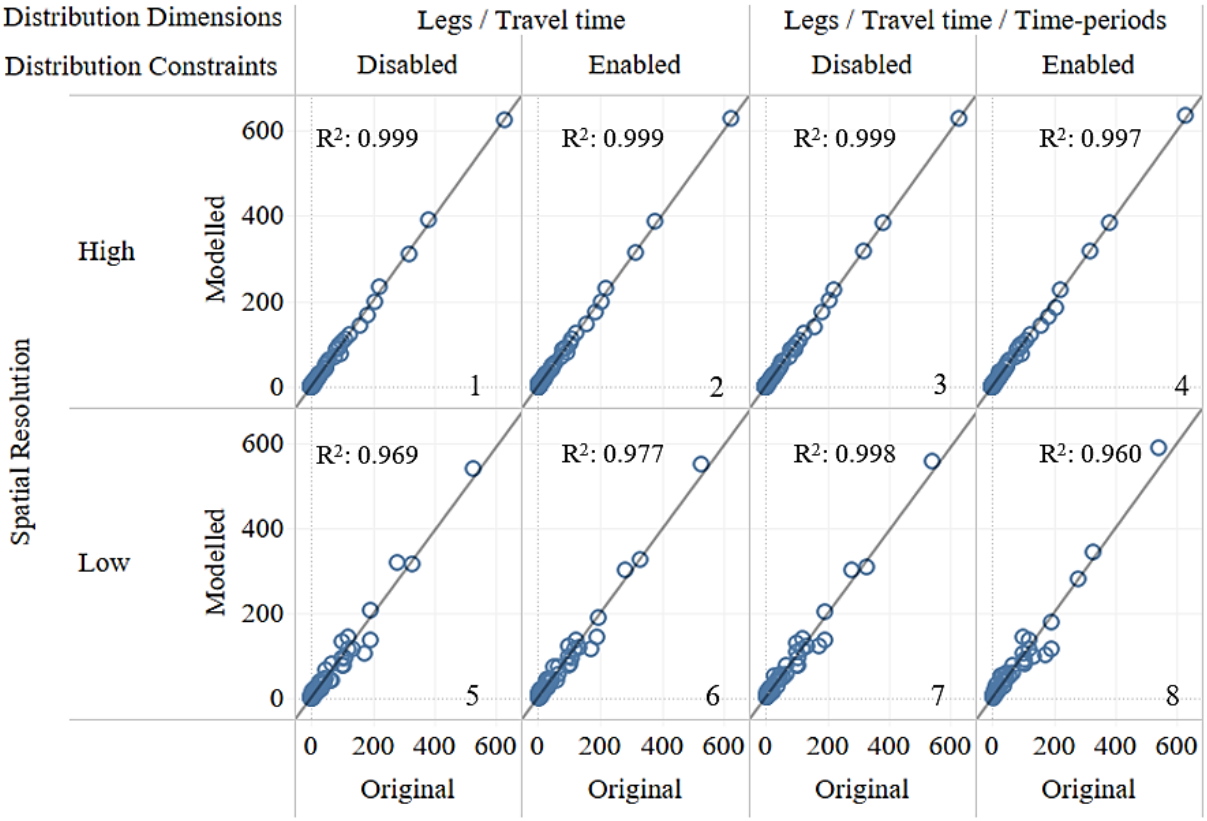


*Figure 9. Comparison of departures time period between original and modelled tours*

## 4. Conclusions and future steps

In this paper, an efficient methodology to convert multi-period Origin-Destination (OD) matrices into sequences of trips originating and ending at a home location (*tours*) is presented. The main motivation behind this study is to enhance the abundant information regarding travel behaviour in typical OD matrices with information about the interdependency between trips.

The process to achieve so takes place in two steps. Firstly, a specifically designed module exploits the connectivity information between zones in the ODs to retrieve all the plausible tours (*candidate tours set*) within the input (*original)* ODs. To achieve so the conversion of the ODs into a graph (network) is required. The conversion of ODs into a network allows for the application of graph theory algorithms, such as cycles identification, which can aid the identification of sequences of zones wherein a zone is reachable from itself (i.e. tour). Secondly, an integer linear programming program is assigned to identify the optimal combination of tours which maximises the utilisation of trips from the *original* OD matrices. The output of this methodology is a set of tours which once aggregated reconstructs the original ODs as accurately as possible. The approach aligns well with the data-driven modelling paradigm, where data is placed in the centre of the model development (Antoniou, Dimitriou and Pereira, 2019). In an era where information regarding mobility is becoming abundant, the authors envisage that the frequent flow of aggregated mobility data could be utilised to dynamically infer behavioural information at the person-level.

In this study, the parameters affecting the complexity of the given problem are thoroughly examined and discussed. Namely the explored parameters are (a) the spatial resolution of the network (network density) and (b) the maximum allowed number of legs in tours. Since transport networks are usually dense, the number of plausible tours can quickly exceed the capabilities of

the currently available optimisation solvers, therefore numerous approaches to limit this number to practical limits are presented and evaluated. These approaches are mainly based on the exploitation of observed information regarding travel behaviour and trip-chaining.

Finally, this novel tours' synthesis framework is tested on a set of realistic cases exhibiting its performance in preparing tours out of standard multi-period ODs. An extensive number of scenarios is prepared and evaluated for the thorough assessment of the accuracy of the presented methodology. The test scenarios are differentiated in terms of (a) the spatial resolution of the input OD matrices, (b) the presence or not of observed information regarding the expected characteristics of the resulting tours and finally (c) the level of detail of this information. The results of the evaluation are particularly encouraging since the suggested methodology managed to accurately reconstruct the original OD matrices into tours without considerable loss of information.

Despite the presented accuracy of the suggested framework, a range of enhancements is still required to improve the methodology and to make it applicable to fully realistic scenarios. Some aspects requiring further research are presented below.

- Further research is required to assess the additional time-processing requirements when more complex travel behaviour patterns are to be included in the solution space (e.g. tours with many legs or tours including subtours).
- Exploration of the possibility to convert tours into activity sequences (i.e. personal activity plans) in the case of purpose dependent OD matrices. The resulting activity plans could be also coupled with a population synthesiser in order to infuse demographic characteristics into the activity sequences.

Arguably, there are still open questions regarding the performance of the presented 'mechanistic' approach to combine standard travel information into detailed travel behaviour patterns. Nonetheless, the suggested methodology has already yielded encouraging results in terms of exploiting aggregated sources of travel demand (i.e. OD matrices) to synthesise travel behaviour information at an individual level.

## References

Adnan, M. *et al.* (2016) 'SimMobility□: A Multi-Scale Integrated Agent-based Simulation Platform', *Transportation Research Board 95th Annual Meeting*, (January), pp. 1–18.

Anda, C., Fourie, P. and Erath, A. (2016) 'Transport Modelling in the Age of Big Data', *Singapore - ETH Centre: Future Cities Laboratory*, Work Repor(June).

Antoniou, C., Dimitriou, L. and Pereira, F. (2019) *Mobility patterns, big data and transport analytics□: tools and applications for modeling*. Elsevier.

Ballis, H. and Dimitriou, L. (2019) 'Optimal population of trip chains synthesis from multi-period origin-destination matrices', in *Proceedings of Transportation Research Board 98th Annual Meeting, Washington D.C.*

Ballis, H., Dimitriou, L. and Ballis, A. (2018) 'A preliminary preparation of trip chains from origin-destination matrices for supporting activity-based models', in *Proceedings of Transportation Research Board 97th Annual Meeting, Washington D.C.*

Balmer, M., Axhausen, K. W. and Nagel, K. (2006) 'A Demand Generation Framework for Large Scale Micro Simulations', in *Transportation Research Board (TRB) Annual Meeting*.

Ben-Akiva, M. *et al.* (2007) 'Towards Disaggregate Dynamic Travel Forecasting Models', *Tsinghua Science and Technology*, 12(2), pp. 115–130. doi: 10.1016/S1007-0214(07)70019-6.

Bhat, C. R. *et al.* (2004) 'Comprehensive Econometric Microsimulator for Daily Activity-Travel Patterns',

*Transportation Research Record: Journal of the Transportation Research Board*, 1894, pp. 57–66. doi: 10.3141/1894-07.

Bhat, C. R. and Koppelman, F. S. (1999) 'Activity-Based Modeling of Travel Demand', (January 2003), pp. 35–61. doi: 10.1007/978-1-4615-5203-1_3.

Bollobás, B. (1999) 'Modern graph theory', *Computers & Mathematics with Applications*, 37(3), p. 136. doi: 10.1016/S0898-1221(99)90429-7.

Bonnel, P. *et al.* (2015) 'Passive mobile phone dataset to construct origin-destination matrix: Potentials and limitations', in *Transportation Research Procedia*. Elsevier, pp. 381–398. doi: 10.1016/j.trpro.2015.12.032.

Castiglione, J., Bradley, M. and Gliebe, J. (2015) *Activity-Based Travel Demand Models: A Primer*, *SHRP 2 Report*. doi: Strategic Highway Research Program (SHRP 2) Report S2-C46-RR-1.

Chu, Z., Cheng, L. and Chen, H. (2012) 'A Review of Activity-Based Travel Demand Modeling', in *CICTP 2012*, pp. 48–59. doi: 10.1061/9780784412442.006.

Cirillo, C. and Axhausen, K. W. (2002) 'Mode choice of complex tours', *Arbeitsbericht Verkehrs- und Raumplanung*, (126), pp. 9–11.

Department for Transport (2017) *National Travel Survey: England 2016*, *National Travel Survey*.

Diestel, R. (2017) *Graph Theory*. Springer. doi: 10.1007/978-3-662-53622-3.

Gong, L. *et al.* (2014) 'Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies', *Procedia - Social and Behavioral Sciences*, 138, pp. 557–565. doi: 10.1016/j.sbspro.2014.07.239.

Goulias, K. G. and Kitamura, R. (1991) 'Recursive Model System for Trip Generation and Trip Chaining', *Transportation Research Record*, (40), pp. 59–66.

Gu, Y. (2004) 'Integrating a Regional Planning Model (TRANSIMS) With an Operational Model (CORSIM)'.

Han, G. and Sohn, K. (2016) 'Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model', *Transportation Research Part B: Methodological*, 83, pp. 121–135. doi: 10.1016/j.trb.2015.11.015.

Huynh, N. *et al.* (2014) 'An Agent Based Model for the Simulation of Transport Demand and Land Use', *Eighth International Workshop on Agents in Traffic and Transportation*, (2005), pp. 1–7.

IBM (2020) *CPLEX Optimizer | IBM*. Available at: https://www.ibm.com/analytics/cplex-optimizer (Accessed: 28 July 2018).

Iqbal, M. S. *et al.* (2014) 'Development of origin-destination matrices using mobile phone call data', *Transportation Research Part C: Emerging Technologies*, 40, pp. 63–74. doi: 10.1016/j.trc.2014.01.002.

Johnson, D. B. (1975) 'Finding All the Elementary Circuits of a Directed Graph', *SIAM Journal on Computing*, 4(1), pp. 77–84. doi: 10.1137/0204007.

Liu, F. *et al.* (2014) 'Building a validation measure for activity-based transportation models based on mobile phone data', *Expert Systems with Applications*, 41(14), pp. 6174–6189. doi: 10.1016/j.eswa.2014.03.054.

Matthews, R. B. *et al.* (2007) 'Agent-based land-use models: a review of applications'. doi: 10.1007/s10980-007-9135-1.

McGuckin, N. and Murakami, E. (1999) 'Examining Trip-Chaining Behavior: Comparison of Travel by Men and Women', *Transportation Research Record: Journal of the Transportation Research Board*, 1693, pp. 79–85. doi: 10.3141/1693-12.

Mcnally, M. G. and Rindt, C. (2008) 'The Activity-Based Approach', in Hensher, D. A. and Button, K. (eds) *Handbook of Transport Modelling*. 2nd edn. Emerald Group Publishing Limited.

Pinjari, A. R. and Bhat, C. R. (2011) 'Activity-based Travel Demand Analysis', in de Palma, A., Lindsey, R., and Quinet, E. (eds) *A Handbook of Transport Economics*. Edward Elgar Publishing Ltd., pp. 213–248. doi: https://doi.org/10.4337/9780857930873.00017.

Rodrigue, J.-P., Comtois, C. and Slack, B. (2017) *The geography of transport systems*. Routledge.

Saadi, I. *et al.* (2016) 'Forecasting travel behavior using Markov Chains-based approaches', *Transportation Research Part C: Emerging Technologies*, 69, pp. 402–417. doi: 10.1016/j.trc.2016.06.020.

Sicotte, G., Morency, C. and Farooq, B. (2017) 'Comparison Between Trip and Trip Chain Models: Evidence from Montreal Commuter Train Corridor'.

Thill, J. -C and Thomas, I. (1987) 'Toward Conceptualizing Trip-Chaining Behavior: A Review', *Geographical Analysis*, 19(1), pp. 1–17. doi: 10.1111/j.1538-4632.1987.tb00110.x.

Tolouei, R., Psarras, S. and Prince, R. (2017) 'Origin-Destination Trip Matrix Development: Conventional Methods versus Mobile Phone Data', in *Transportation Research Procedia*. Elsevier, pp. 39–52. doi: 10.1016/j.trpro.2017.07.007.

Toole, J. L. *et al.* (2015) 'The path most traveled: Travel demand estimation using big data resources', *Transportation Research Part C: Emerging Technologies*, 58, pp. 162–177. doi: 10.1016/j.trc.2015.04.022.

Vogel, A. and Nagel, K. (2013) 'Multi-Agent Based Simulation of Individual Traffic in Berlin', *Journal of Chemical Information and Modeling*, 53(9), pp. 1689–1699. doi: 10.1017/CBO9781107415324.004.

Yue, Y. *et al.* (2014) 'Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies', *Travel Behaviour and Society*, 1(2), pp. 69–78. doi: 10.1016/j.tbs.2013.12.002.

Zhao, J., Rahbee, A. and Wilson, N. H. M. (2007) 'Estimating a rail passenger trip origin-destination matrix using automatic data collection systems', *Computer-Aided Civil and Infrastructure Engineering*, 22(5), pp. 376–387. doi: 10.1111/j.1467-8667.2007.00494.x.