*sustainability*

MDPI

*Article*

# Rapidex: A Novel Tool to Estimate Origin–Destination Trips Using Pervasive Traffic Data

**S. Travis Waller, Sai Chand *, Aleksa Zlojutro, Divya Nair, Chence Niu, Jason Wang, Xiang Zhang and Vinayak V. Dixit**

Research Centre for Integrated Transport Innovation (rCITI), School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW 2052, Australia; s.waller@unsw.edu.au (S.T.W.); a.zlojutro@unsw.edu.au (A.Z.); d.jayakumarnair@unsw.edu.au (D.N.); chence.niu@unsw.edu.au (C.N.); jason.wang1@unsw.edu.au (J.W.); xiang.zhang1@unsw.edu.au (X.Z.); v.dixit@unsw.edu.au (V.V.D.)
* Correspondence: saichand.transport@gmail.com

**Abstract:** A traffic assignment model is a critical tool for developing future transport systems, road policies, and evaluating future network upgrades. However, the development of the network and demand data is often highly intensive, which limits the number of cases where some form of the models are available on a global basis. These problems include licensing restrictions, bureaucracy, privacy, data availability, data quality, costs, transparency, and transferability. This paper introduces Rapidex, a novel origin–destination (OD) demand estimation and visualisation tool. Firstly, Rapidex enables the user to download and visualise road networks for any city using a capacity-based modification of OpenStreetMap. Secondly, the tool creates traffic analysis zones and centroids, as per the user-specified inputs. Next, it enables the fetching of travel time data from pervasive traffic data providers, such as TomTom and Google. With Rapidex, we tailor the genetic-algorithm (GA)-based metaheuristic approach to derive the OD demand pattern. The tool produces critical outputs such as link volumes, link travel times, OD travel times, average trip length and duration, and congestion level, which can also be used for validation. Finally, Rapidex enables the user to perform scenario evaluation, where changes to the network and/or demand data can be made and the subsequent impacts on performance metrics can be identified. In this article, we demonstrate the applicability of Rapidex on the network of Sydney, which has 15,646 directional links, 8708 nodes, and 178 zones. Further, the model was validated using the Household Travel Survey data of Sydney using the aggregated metrics and a novel project selection method. We observed that 88% of the time, the "estimated" and "observed" OD matrices identified the same project (i.e., the rapid process estimated the more intensive traditional approach in 88% of cases). This tool would help practitioners in rapid decision making for strategic long-term planning. Further, the tool would provide an opportunity for developing countries to better manage traffic congestion, as cities in these countries are prone to severe congestion and rapid urbanisation while often lacking the traditional models entirely.

**Keywords:** pervasive traffic data; demand estimation; genetic algorithm; transport network modelling



**Citation:** Waller, S.T.; Chand, S.; Zlojutro, A.; Nair, D.; Niu, C.; Wang, J.; Zhang, X.; Dixit, V.V. Rapidex: A Novel Tool to Estimate Origin–Destination Trips Using Pervasive Traffic Data. *Sustainability* **2021**, *13*, 11171. https://doi.org/10.3390/su132011171

Academic Editor: Armando Cartenì

Received: 6 August 2021
Accepted: 7 October 2021
Published: 10 October 2021



## 1. Introduction

Globalisation, population rise, and the exponential growth in communication and vehicle technology have led to the increased desire and ability to travel longer distances in shorter periods of time. Consequently, transport systems worldwide suffer from congestion, leading to a lack of travel time reliability, reducing safety, intensifying environmental degradation, and ultimately creating economic inefficiency. Traffic congestion is estimated to cost around $88 billion per year in the US, £6.9 billion in the UK, and €2.8 billion in Germany [1]. Transport practitioners have two broad approaches to alleviate these issues: travel demand management and supply management. Demand management strategies

aim to decrease drive-alone trips by redistributing demand to different transport modes, departure times, and routes. On the other hand, supply management strategies aim to build new roads or infrastructure or upgrade the existing ones to reduce congestion [2]. Nevertheless, these approaches may not always improve traffic conditions and, at times, may exacerbate pre-existing issues.

Transport agencies in relatively few cities globally, mainly from developed countries, deploy traffic assignment models to evaluate and test different traffic management strategies and rank infrastructure design projects [3]. Traffic assignment is the final step in the classic four-stage transport modelling approach where transport demand in a network is loaded over an existing road network [4]. The link travel time and delay are evaluated at every iteration, and vehicle assignment is adjusted until user equilibrium is achieved. The traffic assignment models consist of two critical inputs, i.e., demand (travel analysis zones and the travel patterns) and supply (road network data, traffic signals, etc.). These models can realistically capture the performance of a transportation network by "predicting" the travel routes of forecasted trips and thus help in planning, designing, and operating transportation systems efficiently and sustainably. Despite their advantages, relatively few cities deploy detailed traffic assignment models, which can be attributed to the problems (licensing restrictions, bureaucracy, privacy, data availability, data quality, costs, transparency, transferability, etc.) in accessing data.

Most of the previous applications of traffic assignment obtained road network data from various sources—including city, state, and national data repositories [5–7], which relied extensively on aerial imagery, and surveying. In recent years, OpenStreetMap (OSM), a crowdsourced mapping platform, has become a vital source of road network data, owing to its advantages of being free, easily downloadable, having timely updates, feature-rich data, and wide coverage [8]. Researchers have used OSM data for various transportation studies, focusing on traffic assignment [9–11], incident analysis [12], emissions [13,14], road network properties, and multiple city comparisons. There are also various tools that can download, correct topology, analyse, and visualise data from OSM [15–18].

Unlike the road network data, obtaining the travel demand patterns in various cities simultaneously and comparing and contrasting them, has been challenging. Having access to a database/tool that estimates demand patterns and visualises the output in any selected city would enable researchers to attempt many studies that would have been otherwise challenging. For example, researchers can utilise the tool to understand day-to-day and within-day demand changes in a city. They can compare the demand patterns in one city with another. They can analyse the impacts of network changes, e.g., adding or removing lanes, changing speed limits, adding or closing new roads, etc., on critical metrics such as trip length, travel times, and congestion. The impacts of extreme events such as natural disasters, festivals, lockdowns, etc., on travel patterns can be studied as well. This tool would also help practitioners in rapid decision making for strategic planning and operational purposes. Furthermore, the tool would provide an opportunity for developing countries to better manage traffic congestion, as cities in these countries are prone to severe congestion and rapid urbanisation [19].

The main challenge in developing such a tool has been access to quality and timely data. Researchers and practitioners have traditionally relied on origin–destination (OD) demand data obtained through the Household Travel Surveys (HTS) [5,20,21]. These surveys are conducted as one-shot surveys undertaken every few years [22] and are labour intensive, costly, and often outdated. The HTS data are helpful in long-term planning models but not ideal for day-to-day traffic analysis and devising operational strategies.

Because of the challenges associated with HTS data, many researchers have tried to "estimate" the OD patterns using some field data, notably traffic count data obtained from loop detectors. There have been a few studies lately that used travel time data obtained through Bluetooth scanners, floating cars, call data records (CDR), license plate recognition, etc., in addition to count data. Loop detector and Bluetooth scanner data suffer from the problem of spatial coverage, CDRs have the problem of location accuracy, and the

probe vehicle method lacks transferability. Furthermore, relying on survey data is time consuming and expensive and limits its scalability for rapid deployment. Therefore, travel demand data for multiple cities are challenging to obtain as there is no consistent and reliable data source or tool. Most of the studies in the OD estimation space either deal with test networks or small real-world networks with limited network detail, i.e., fewer number of zones and links. Another key limitation of the existing methods is the reliance on the limited number of observations of traffic counts [23]. The number of unknowns (number of OD pairs, typically thousands or hundreds of thousands for large networks) significantly outnumber the known observations (a few hundred thousand link counts), causing under-determinacy.

The current study tries to overcome the problem by utilising easily accessible pervasive traffic data for travel demand estimation. In recent years, pervasive (some studies refer to it as crowdsourced) traffic data sources have become potential options for traffic data collection because of the widespread usage of smartphones [24,25]. People have quick access to features such as Bluetooth, WiFi, and GPS, and are better connected through social media (SM) applications, through which user location data, travel history, travel times, activity behaviour, incidents, and traffic speeds are collected. Commercial traffic and navigation data providers (Google, TomTom, HERE, etc.), social media platforms (Facebook, Twitter, Instagram, etc.), mobility-as-a-service aggregators (Uber, Didi Chuxing, Ola, etc.), and fitness service providers (Strava, Google Fit, etc.) make use of this data. Pervasive traffic data have a higher sample size that reflects the traffic conditions better than the probe vehicles [25]. Furthermore, they have comprehensive spatial coverage and fine temporal resolution and are cost-effective compared to traditional data sources [24–26]. Such data have been utilised in the past for incident duration prediction [27], designing adaptive traffic signals [26], and congestion estimation [24].

This paper demonstrates how we utilised pervasive traffic data and developed Rapidex, a novel OD demand estimation and visualisation tool. Firstly, the tool enables users to download and visualise road network data (roads, road length, number of lanes, speed limits, type of road, etc.) for any city using OpenStreetMap and an external tool called OSMnx [16]. Secondly, the tool creates traffic analysis zones and centroids, where the user can specify the zone size and the maximum/minimum number of centroids per zone. Thirdly, the tool enables the fetching of travel time data from the TomTom and/or Google Maps application programming interfaces (APIs). Finally, the tool predicts the OD demand patterns across the network using a customised genetic algorithm (GA) approach, which minimises the gap between the observed and estimated travel times within a bilevel optimisation framework. After this step, the tool produces critical outputs such as link volumes, link travel times, OD travel times, average trip length and duration, V/C ratio, congestion level, etc. These output values can be used to validate the demand matrix in the case such data for the real world can be accessed from other sources. Finally, the tool enables the user to perform scenario evaluation, where changes to the network and/or demand data can be made, and the subsequent impacts on performance metrics can be identified. Therefore, the main objective of Rapidex is to help practitioners and authorities in quick decision making for strategic long-term planning by utilising pervasive traffic data.

The rest of the article is organised as follows. First, a brief review of different data sources and methods for demand estimation is provided. It is followed by the description of the methodology and the experiment setup. The case study of Sydney is then presented, followed by concluding remarks.

## 2. Literature Review

The four-step travel model is a framework used to forecast how the transportation network behaves so that planners can better understand what is required to support current and future scenarios. The first step is to assign demand to the trips between all the OD pairs within the network, leading to creating an OD matrix. The OD matrix needs to serve as a solid foundation for the rest of the model to build upon. Traditionally, OD

matrices were populated through large roadside or household travel surveys; however, these surveys are costly, unreliable due to reporting errors, and outdated only a couple of years after collection. Thus, researchers naturally turned to create models and frameworks for estimating OD matrices using collected real-world data. In this paper, we present a brief review of a few widely used methods and datasets for OD demand estimation.

### *2.1. Review of Methods for OD Demand Estimation*

Origin–destination estimation models have existed for decades and have used various data sets to fuel their predictions. The earlier models employed link traffic counts and were based on static formulations. In recent years, there has been a growing interest in developing dynamic [28] and quasi-dynamic formulations [29], which can capture the within-day demand variations. The researchers have proposed various mathematical formulations, but most of them treat the OD estimation as a bi-level optimisation problem, which minimises the gap between the observed and modelled link counts. The upper level is concerned with taking the target OD matrix and adjusting it to reproduce the observed traffic counts. The estimated traffic volume that the upper level uses is produced by the lower level, which applies the OD matrix onto the network and solves for user equilibrium. Furthermore, most of the studies tend to "update" the historical OD matrix based on the additional (or latest) traffic data. Due to this, the solution space for the new matrix could be biased or limited. A review of different solution algorithms has been covered in many past studies, most notably by [30,31].

Essentially, there are four broad approaches for OD matrix estimation, from which multiple generalisations have been proposed [23]. These main approaches are minimum information/maximum entropy, Bayesian inference, generalised least squares (GLS), and maximum likelihood. The generalisations/advancements include methodologies such as neural networks [32,33], Kalman filtering [34], Markov chain theory [35], genetic algorithm [36], random forest [37], and data-driven methods [38,39]. A brief overview of the four main approaches is presented below.

Given the most uncertainty, minimum information/maximum entropy can determine the most probable OD matrix and produce this estimate without any bias. It is computationally advantageous as it is faster and convenient. However, it is unable to consider uncertainties in traffic counts and any prior matrices, which can often lead to errors as these uncertainties significantly impact the output. Further, the model requires removing inconsistencies from the data to produce a feasible solution [40]. Bayesian inference provides a framework that introduces extra information based on accumulated statistics and research through the specification of the prior. In this method, providing variability information and probability intervals of traffic flow estimation provides an essential advantage over non-statistical methods. However, uncertainty can arise when estimating the probability choice distribution. Complicated likelihood distribution and non-negligible measurement errors can occur in vehicle counts. Further, computational costs and the extent of applicability to varying traffic conditions is a limitation [41,42].

The generalised least squares method allows the combination of traffic survey and observed traffic counts. It also accounts for the relative accuracies of both data sources. Nevertheless, GLS estimators can be biased when the starting estimators are biased and/or the assignment model is misspecified [43,44]. The maximum likelihood estimation with expectation maximisation can be used to estimate an OD matrix when the proportions or percentages of the traffic demand on the links are known, but the traffic volumes are unknown [45]. The model still works even if the matrix has zero elements in the cells, whereas it would not be feasible in other models such as the maximum entropy model. Even with a small amount of data available, highly accurate results can still be achieved to estimate the OD matrix. However, it can be computationally challenging for large matrices when using the expectation maximisation function.

### *2.2. Datasets for OD Demand Estimation*

#### 2.2.1. Loop Detector Data

Estimating OD matrices using traffic count data obtained through loop detectors has been going on for several decades [40,46]. However, the quality of OD matrices obtained using count data are questionable [47]. Loop detectors are electrically conducting loops that are installed under the pavement. Once a vehicle moves over it, the metal of the vehicle causes a reduction in the inductance of the loop, which is recognised by the unit. They are better than HTS in terms of labour and time savings. However, they depend on metering infrastructure that is expensive to install and maintain [48]. They are prone to provide erroneous counts—for example, one study found that 31% of the loop detectors had biased counts [49]. Furthermore, not all roads in a network are installed with loop detectors and so the OD estimation model may sometimes be working with only partial information.

#### 2.2.2. Mobile Phone Data/Call Data Records (CDRs)

The surge of information and communications technology innovation has blossomed the reliance on mobile phones as a necessary means for communication and as a life management tool. This has opened up the opportunity for transport planners to exploit the data already collected from cellular providers to deliver newfound insights into travel behaviours. CDRs are collected and stored by cellular providers for billing purposes and contain important spatial data generated by mobile phone users. CDRs are useful when the mobile phone user interacts with their phone across different locations and can thereby monitor their physical displacements over time. Trips can then be tracked within a specific timeframe and generate tower-to-tower matrices [50,51]. A primary advantage of CDR is the passive nature of its collection process. CDR can be particularly useful in developing countries, where reliable HTS data may be limited. The continual inflow of mass data makes it convenient for periodic updates.

Despite their seeming advantages, CDRs suffer from various limitations. They are inhibited by the fact that the users must have their phones on hand and must be in operation. This means that for cases where phones are switched off, lost, or faulty, the CDR data being recorded may come up as missing or discontinuous. Incomplete data may misinterpret trip behaviours of individuals due to the absences of true origins or destinations. Additionally, CDRs are subjected to false displacements induced by the unrelated operations of cellular providers. Cell phone operators control call traffic by redistributing calls to different towers based on the call activity level. This results in incorrect displacements to be recorded for redirected calls. Finally, the mobile operators may not always share the CDR data because of privacy concerns.

#### 2.2.3. Bluetooth Data

Bluetooth technology has been used in a few studies to estimate OD demand patterns [34,52,53]. Sensors installed on the roadside collect the unique Bluetooth ID of vehicles. When several sensors are installed in the network, the time-stamped vehicle locations, i.e., trajectories, can be collected. The costs for hardware, software, and installation are low compared to loop detectors. However, these data have some limitations. Some Bluetooth IDs are found to be cloned, i.e., the same ID is shared among multiple vehicles, which may lead to errors [53]. In addition, the sensors located close to each other can have overlapping detection zones. Furthermore, the penetration is only 1% to 5% because the Bluetooth functionality can be turned off in many mobile devices to conserve the battery [54]. In addition, most of the studies relied on Bluetooth data from only a handful of sensors. Therefore, demand patterns have been estimated in smaller or highly aggregated (less number of links, nodes, and zones) networks.

#### 2.2.4. Social Media (SM) Data

Social media is known to have a generally easy access point via online platforms giving rise to extensive user coverage. The emergence of location tagging and regular

posts by users increasingly generates public-registered information that is often linked to the user's real-time position. Several studies utilised geotagged data from Twitter and Foursquare to estimate OD matrices [48,55–57]. The main advantage of SM as a data source is its universal accessibility and broad coverage across users. In addition, the availability of geo-coded capabilities on platforms such as Twitter and Foursquare allow for differentiating location-based activity patterns that are both easier to obtain and more detailed in comparison with HTS results.

SM data are not that helpful in estimating commuting travel demand due to the low reliability of identifying home and workplace [57]. A significant limitation of such data is that the geotagged tweets account only for a small proportion (1–2%) of all tweets [58]. Another drawback is the influence of the type of users correlating with the locations being checked into. For example, locations with a major landmark or historical value were biased towards tourist users [55]. Due to the innate purpose of SM to share unique personal experiences, it is not rare to observe a high frequency of activity geotagged to significant locations. However, as a tourist, their trips recorded from that location and surrounding areas are now inhibited by very different travel behaviours in comparison to the local community captured by HTS. This inhibits the consistency of the activity patterns being recorded, as tourists often visit seasonally, thereby affecting the accuracy of developed ODMs for frequently visited tourist locations. In addition, the datasets can cover many users but take significant time to accumulate enough samples for each individual.

#### 2.2.5. Other Sources

Researchers have also used other sources, such as floating car techniques in the form of taxi trajectories [59,60] and license plate recognition through traffic monitoring cameras [61–63]. Floating cars may not represent the real traffic conditions due to limited penetration. In addition, the operating conditions of taxis are quite diverse and different from that of the typical commuter. On the other hand, traffic monitoring cameras are mainly intended for surveillance and violations, and their spatial coverage is limited because of the high installation and maintenance costs. However, one can identify vehicle classification, lane-wise traffic, and turning proportions with this method, which can be helpful in OD estimation.

#### 2.2.6. Pervasive Data

In a way, data from pervasive navigation platforms is akin to floating cars but with much higher penetration rates. The main advantages of such data are the comprehensive spatial coverage and fine temporal resolution. The data are updated in real-time and account for fluctuations in traffic activity. Speed/travel time information can be obtained for every road link in the network, which is not feasible with other modes of data collection. However, traffic "count" data cannot be routinely obtained from these platforms owing to privacy issues. Nevertheless, speed patterns are observed to coevolve with traffic volume patterns [64,65], and the data has been used in several studies for traffic analysis and developing network fundamental/flow diagrams. However, to the best of the authors' knowledge, such data have not been used for OD demand estimation, despite being easily accessible.

## 3. Methodology

The Rapidex tool has five core modules, for which the detailed methodology is discussed in this section. The key nomenclature is presented in Table 1.

**Table 1.** List of notations and abbreviations.

| | |
|---|---|
| N | Set of nodes |
| A | Set of links |
| $x_a$ | Flow on link a |
| $t_a$ | Travel time on link a |
| $d_{rs}$ | Demand between zonal centroids r and s |
| $\pi_{rs}$ | Set of acyclical paths connecting zone r to zone s |
| $f_k^{rs}$ | Flow on path $k$ on OD pair ($r$, $s$) |
| E | Error function |
| $TT_{rs}^{obs}$ | Observed (from any pervasive platform) travel time between OD pair $r$ and $s$ |
| $TT^f{}_r$ | Observed free-flow travel time between OD pair $r$ and $s$ |
| $k_{rs}$ | Average shortest distance between the OD pair $r$ and $s$ when the network is empty |
| $G_r$ | User-defined proportion value of zone $r$, where $\sum G_r = 1$ |
| $A_s$ | User-defined proportion value of zone $s$, *where* $\sum A_s = 1$ |
| $D$ | Total demand of the network |
| $TT_{rs}^{est}$ | Estimated (from a solution) travel time between OD pair $r$ |
| $N_{OD}$ | Number of OD pairs |
| $t_{ij}^{est}$ | Estimated (from a solution) travel time between link $i$ and $j$ |
| $t_{ij}^{obs}$ | Observed (from any pervasive platform) travel time between link $i$ and $j$ |
| $f_{ij}^{est}$ | Estimated (from a solution) flow between link $i$ and $j$ |
| $f_{ij}^{obs}$ | Observed (from loop detector or other sources) flow between link $i$ and $j$ |
| $N_f$ | Number of links in the network where flow values are known |
| $N_t$ | Number of links in the network where TT values are known |
| $R_i^{est}$ | Estimated (from a solution) travel time along a user defined route/corridor $i$ |
| $R_i^{obs}$ | Observed (from any pervasive platform) travel time along a user defined corridor $i$ |
| $N_R$ | Number of user-defined corridors |
| OD | Origin–destination |
| GA | Genetic algorithm |
| OSM | OpenStreetMap |
| HTS | Household Travel Survey |
| CDR | Call data records |
| API | Application programming interface |
| GLS | Generalised least squares |
| SM | Social media |
| TFM | Travel time free-flow travel time model |
| FDM | Free-flow travel time distance model |
| TDM | Travel time distance model |
| CGM | Custom gravity model |
| CBD | Central business district |
| LOS | Level of service |
| BPR | Bureau of Public Roads |

**Table 1.** *Cont.*

| | |
|---|---|
| TSTT | Total system travel time |
| MAPE-ODTT | Mean absolute percentage error of OD travel times |
| RMSE-ODTT | Root mean square error of OD travel times |
| MAPE-LF | Mean absolute percentage error of link flows |
| RMSE-LF | Root mean square error of link flows |
| RMSE-LTT | Root mean square error of link travel times |
| MAPE-LTT | Mean absolute percentage error of link travel time |
| MAPE-C | Mean absolute percentage error of corridor travel times |

### *3.1. Road Network Extraction and Zoning*

First, the coordinates of a bounding box encompassing the city are input into Rapidex. By default, we select motorway, trunk, primary, and secondary edges. The tertiary and residential type edges are not considered for modelling. Then we select the maximum and minimum grid sizes (5 km and 1.25 km, respectively, in this case study). First, Rapidex downloads the network from OSM using a Python package called OSMnx [16] and divides the entire network into square grids of the specified maximum size. Then the grids with higher node density are further disaggregated into smaller zones (again, squared shape). Next, the tool identifies a specified number of nodes within each zone to act as centroids for that zone, which permit the entering and exiting of vehicles from the network. By default, a maximum of 4 and a minimum of 3 centroids are selected per zone. The capacity values for different road types are then set. All the parameters mentioned here are only the default settings in the tool, and the user will have the flexibility to change these parameters. Furthermore, the user will have the option to import a pre-defined zoning structure (e.g., census levels or statistical areas) in place of the default zoning structure available in Rapidex. At the end of this step, the users can visualise the attributes of every link (speed limit, length, lanes, etc.) and zone (number of nodes, length of roads, etc., within the zone) in the network.

### *3.2. Travel Time Extraction*

Pervasive data aggregators such as Google and TomTom provide both real-time and typical travel times through their APIs. The real-time travel times are prone to fluctuations and may not represent the everyday traffic conditions. On the other hand, the "typical" travel times are calculated as average travel times across multiple days (on the same day of the week), but at the exact departure time. Rapidex allows the user to collect either of these travel times. The default approach in Rapidex is to extract the typical link travel times and "calculate" the OD travel times using the shortest path algorithm. These OD travel times are then used in the error function calculation. One can directly obtain the OD travel times from the pervasive data platforms; however, a change in any of the zoning parameters may result in the previously collected data being futile. In addition, for cities with many zones, the number of queries would increase exponentially and may prove costly to fetch data.

### *3.3. OD Estimation*

#### 3.3.1. Bilevel Optimisation Approach

The bi-level programming approach is typically used for estimating the OD matrix for networks that are congested. It consists of upper and lower-level problems that are solved through iterations until both levels converge simultaneously. Usually, the upper-level problem involves estimating the OD matrix using information about the observed link counts and aims to reduce the errors between the observed and predicted values. The lower level problem is the user-equilibrium assignment at the network level, which describes travellers' interaction with different traffic situations. At each iteration, the lower level returns the traffic properties, including flow and travel time, to the upper level, while the

upper level provides the estimated OD demand as the input of the lower-level program. Demand and assignment are mutually dependent on each other in a bi-level approach, and therefore, they are better able to replicate congested traffic conditions [52,66]. Most of the previous studies have the upper-level problem specified as follows:

$$\min \mathrm{E(f)} = \sum_{ij} \left( f_{ij}^{est} - f_{ij}^{obs} \right)^2$$

where E(f) = error function; $f_{ij}^{est}$ = estimated vehicle count on link $ij$; and $f_{ij}^{obs}$ = observed vehicle count on link $ij$.

Varying from the existing literature, Rapidex offers multiple developed methods for the error function E, from which the user can choose one or a combination of the error functions. More details about the error function are discussed in the sub-section "Error Function". With the defined error function E, we propose the bi-level modelling framework for the OD estimation process as follows.

Upper level:

$$\min \mathrm{E}$$

where E is the error function (refer to the sub-section "Error Function").

Lower-level:

$$z(x(f)) = \min \sum_{a \in A} \int_0^{x_a} t_a(w) dw$$

subject to the flow conservation and non-negativity constraints as outlined below:

$$x_a = \sum_{r \in N} \sum_{s \in N} \sum_{k \in \pi_{rs}} f_k^{rs} \delta_{a,k}^{rs} \; \forall a \in A$$

$$\sum_{k \in \pi_{rs}} f_k^{rs} = d_{rs} \; \forall r, s \in N$$

$$f_k^{rs} \geq 0 \; \forall r, s \in N, \; \forall k \in \pi_{rs}$$

where $t(x)$ is the function for flow-dependent link travel time.

Currently, Rapidex has two types of solution heuristic, i.e., method of successive averages and Frank–Wolfe to solve user equilibrium traffic assignment.

Link Performance Function

Link performance functions measure the level of service (LOS) associated with the links representing an urban network [4]. These can include travel time, safety, cost of travel, stability flows, and others. However, travel time is typically used as the sole measure of LOS as other measurements are highly correlated with travel time. The Bureau of Public Roads (BPR) function is widely used in practice and is displayed in the equation below.

$$T = T_o * \left\{ 1 + \alpha * \left( \frac{v}{c} \right)^{\beta} \right\}$$

where $T$ = link travel time; $T_0$ = link travel time at free flow speed; $\alpha$, $\beta$ = BPR parameters (often set at 0.15 and 4 respectively); $v$ = traffic flow (veh/h); and $c$ = practical link capacity (veh/h). Rapidex uses the BPR function described above to calculate travel times based on the estimated volume.

3.3.2. Solution Approach: The Genetic-Algorithm-Based Metaheuristics

For real-world and large-scale networks, solving for a global optimum using a bi-level approach is difficult because it is usually non-convex [67]. Researchers have used different heuristic approaches to address the problem, but they rely on the target OD matrix and some may still result in a local optimum for the upper level [68]. Some researchers have used the genetic algorithm (GA) approach to overcome this limitation [36,69] as it can

handle any kind of objective function and constraints. GA starts with a set of feasible solutions and produces an optimal solution by iterating sequential steps [36]. The typical GA process follows the procedure as shown in Figure 1. However, in Rapidex, the GA approach has been tailored significantly to estimate OD demand, and the process is outlined in the subsequent sections.

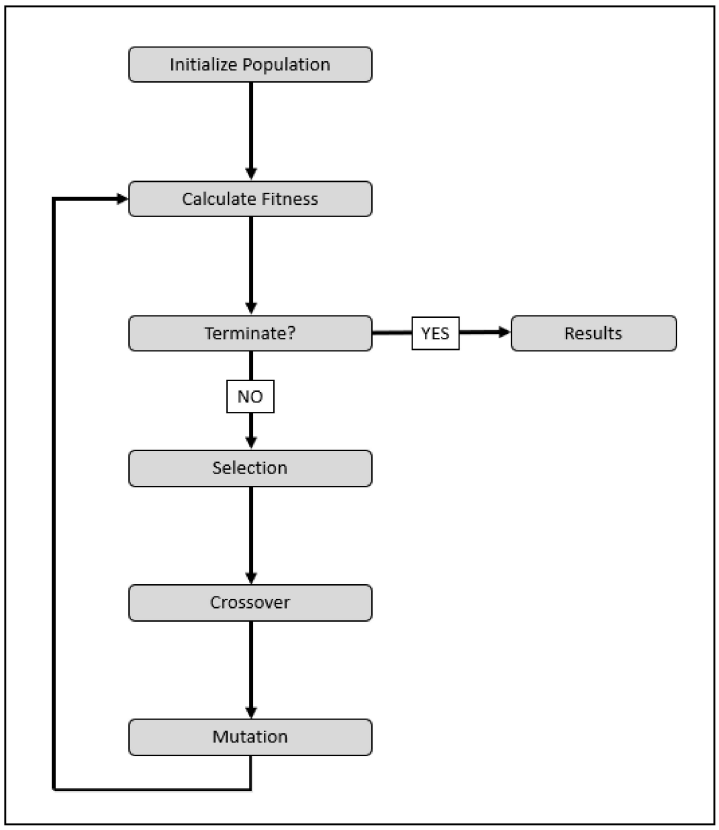



**Figure 1.** Genetic algorithm methodology.

Initial Solutions

OD demand estimation is a highly underdetermined problem, particularly in large networks. Multiple solutions for the OD matrix can exist for the same set of travel times. Therefore, providing accurate initial solutions (matrices) is critical in achieving a reliable final solution. The better the initial solution in terms of closeness to the observable solution, the better the GA estimation will be and the faster it will converge. A list of different initial solutions currently offered by Rapidex is presented in Table 2. It can be noted that all these solutions take the form of a gravity model.

**Table 2.** Initial solutions offered by Rapidex.

| Acronym | Method Name | Governing Equation | Notation |
|---|---|---|---|
| TFM | Travel time—free flow travel time model. | $d_{rs} = \frac{\frac{TT_{rs}^{obs}}{TT_{rs}^{f}}}{\sum_{rs}\frac{TT_{rs}^{obs}}{TT_{rs}^{f}}} \cdot D$ | $TT_{rs}^{obs}$—Observed (from any pervasive platform) travel time between OD pair $r$ and $s$. $TT^{f}{}_{rs}$—Observed free-flow travel time between OD pair $r$ and $s$. $k_{rs}$—Average shortest distance between the OD pair $r$ and $s$ when the network is empty. $G_r$—user-defined proportion value of zone $r$, where $\sum G_r = 1$. $A_s$—user-defined proportion value of zone $s$, *where* $\sum A_s = 1$. |
| FDM | Free flow travel time—distance model. | $d_{rs} = \frac{\frac{TT_{rs}^{f}}{k_{rs}{}^{2}}}{\sum_{rs}\frac{TT_{rs}^{f}}{k_{rs}{}^{2}}} \cdot D$ | |
| TDM | Travel time distance model. | $d_{rs} = \frac{\frac{TT_{rs}^{obs}}{k_{rs}{}^{2}}}{\sum_{rs}\frac{TT_{rs}^{obs}}{k_{rs}{}^{2}}} \cdot D$ | |
| CGM | Custom gravity model. | $d_{rs} = \frac{\frac{G_r A_s}{k_{rs}{}^{2}}}{\sum_{rs}\frac{G_r A_s}{k_{rs}{}^{2}}} \cdot D$ | |

All the initial solution methods listed in Table 2 are subject to:

$$D = \sum_{rs} d_{rs}$$

where $D$ is the total demand of the network and $d_{rs}$ is the demand between a particular OD pair $r$ and $s$. The CGM is the most flexible of the initial solutions as the user can define the values of proportions $G$ and $A$ for each zone. For example, the user may decide that they want to use the proportion of the population (if known) or the proportion of historical trip productions (from a priori OD matrix) as a proxy for trip productions and proportion of historical trip attractions (again, from a priori matrix) as an indicator for the attractiveness of a zone. If such data is not available, Rapidex provides proxies, such as the proportion of residential roads and the proportion of nodes, which can be used in the custom solutions.

The user can decide to use one or a combination of these initial solutions for the first generation. Each initial solution requires a total demand. The user has two options for generating the total demand of a solution. Firstly, the user can decide to have the initial solution's total demand as a random value from the demand range they specified for the network. Alternatively, the demand range can be equally divided for every type of initial solution method being used. These division points are the total demands that the initial solutions are assigned. For example, say that we have 10 solutions, and we elect to use 50% TFM and 50% TDM. For the TFM solutions, the demand range is divided into 5 different demand values, and each of these values is assigned to one of the TFM solutions. Similarly, the TDM solutions will be calculated in the same way.

Error Function

Once each of the OD matrices within the generation has been solved to convergence, i.e., a predefined relative gap, each solution needs to be evaluated by an error function. The error function is the most crucial aspect of the GA as it ranks different solutions and defines their viability. Rapidex offers several error functions (see Table 3), from which the user has the flexibility to choose one or a combination of error functions. For example, if the link flow data at a few locations and the OD travel times in the network are known, then one can use the combined error sums of RMSE-ODTT and RMSE-LF. Weighting can also be incorporated if the user has a preference (or more confidence) for one dataset over another.

**Table 3.** Error function methods in Rapidex.

| Acronym | Method Name | Governing Equation | Notation |
| --- | --- | --- | --- |
| MAPE-ODTT | Mean absolute percentage error of OD travel times. | $E = \sum_{rs} d_{rs} \cdot \frac{\left\lvert TT_{rs}^{est} - TT_{rs}^{obs} \right\rvert}{TT_{rs}^{obs}}$ | • $E$—Error value.<br>• $TT_{rs}^{est}$—Estimated (from a solution) travel time between OD pair $r$ and $s$.<br>• $TT_{rs}^{obs}$—Observed (from any pervasive platform) travel time between OD pair $r$ and $s$.<br>• $N_{OD}$— Number of OD pairs.<br>• $f_{ij}^{est}$—Estimated (from a solution) flow between link $i$ and $j$.<br>• $f_{ij}^{obs}$—Observed (from loop detector or other sources) flow between link $i$ and $j$.<br>• $N_f$—Number of links in the network where flow values are known.<br>• $t_{ij}^{est}$—Estimated (from a solution) travel time between link $i$ and $j$.<br>• $t_{ij}^{obs}$—Observed (from any pervasive traffic platform) travel time between link $i$ and $j$.<br>• $N_t$—Number of links in the network where travel time values are known.<br>• $R_i^{est}$—Estimated (from a solution) travel time along a user defined route/corridor $i$.<br>• $R_i^{obs}$—Observed (from any pervasive platform) travel time along a user defined corridor $i$.<br>• $N_R$—Number of user-defined corridors. |
| RMSE-ODTT | Root mean square error of OD travel times. | $E = \sqrt{\frac{\sum_{rs}\left(TT_{rs}^{est} - TT_{rs}^{obs}\right)^2}{N_{OD}}}$ | |
| MAPE-LF | Mean absolute percentage error of link flows. | $E = \sum_{ij} \frac{\left\lvert f_{ij}^{est} - f_{ij}^{obs} \right\rvert}{f_{ij}^{obs}}$ | |
| RMSE-LF | Root mean square error of link flows. | $E = \sqrt{\frac{\sum_{ij}\left(f_{ij}^{est} - f_{ij}^{obs}\right)^2}{N_f}}$ | |
| RMSE-LTT | Root mean square error of link travel times. | $E = \sqrt{\frac{\sum_{ij}\left(t_{ij}^{est} - t_{ij}^{obs}\right)^2}{N_t}}$ | |
| MAPE-LTT | Mean absolute percentage error of link travel time. | $E = \sum_{ij} \frac{\left\lvert t_{ij}^{est} - t_{ij}^{obs} \right\rvert}{t_{ij}^{obs}}$ | |
| MAPE-C | Mean absolute percentage error of corridor travel times. | $E = \sum_{i} \frac{\frac{\left\lvert R_i^{est} - R_i^{obs} \right\rvert}{R_i^{obs}}}{N_R}$ | |

Selection

To proceed to the subsequent iteration (generation) of GA, new OD matrices must be created from the previous generation. First, solutions from the previous generation are selected and then combined. These selected solutions are referred to as parents. Rapidex offers different methods for parent selection.

The simplest method is randomly selecting two different solutions from the previous generation. The second method, i.e., the tournament selection, is an improvement over the random selection while still maintaining simplicity. In tournament selection, solutions are randomly selected to the desired number (two in our case) and the solution with the best fitness (or least error) is picked [70]. This is repeated once more, resulting in two parents who are guaranteed not to be the worst of the previous generation. To further ensure that the subsequent generation does not degrade, the elitist selection is employed whereby the next generation will be given the best performers of the previous generation. The number of elites that continue onto the next generation is user defined.

Crossover

Once the parent solutions have been decided, they are then combined using a crossover method to produce a new child chromosome. Again, Rapidex provides multiple techniques for "creating" the offspring, i.e., a new solution from the parents.

In the uniform crossover method, either parent's OD demand for any pair is chosen. In the single-point crossover method, the OD demand values for pairs from one parent are copied until a particular OD pair, and then the rest, are copied from the other parent. In the arithmetic method, for each OD pair the demand is the weighted arithmetic mean of the two parent demands [71].

While these general processes will yield results, a tailored crossover approach will provide better performance as characteristics of your problem may be leveraged to crossover parents more efficiently. Thus, a new crossover method was created specifically for our problem. The process is shown in Figure 2 and is outlined below:

1. Calculate the absolute difference between the first parent's estimated travel time and the observed travel time for that OD pair.
2. Repeat step 1 with the second parent.
3. Assign the OD demand of the parent with the lower absolute difference to the child.
4. Repeat for all OD pairs.

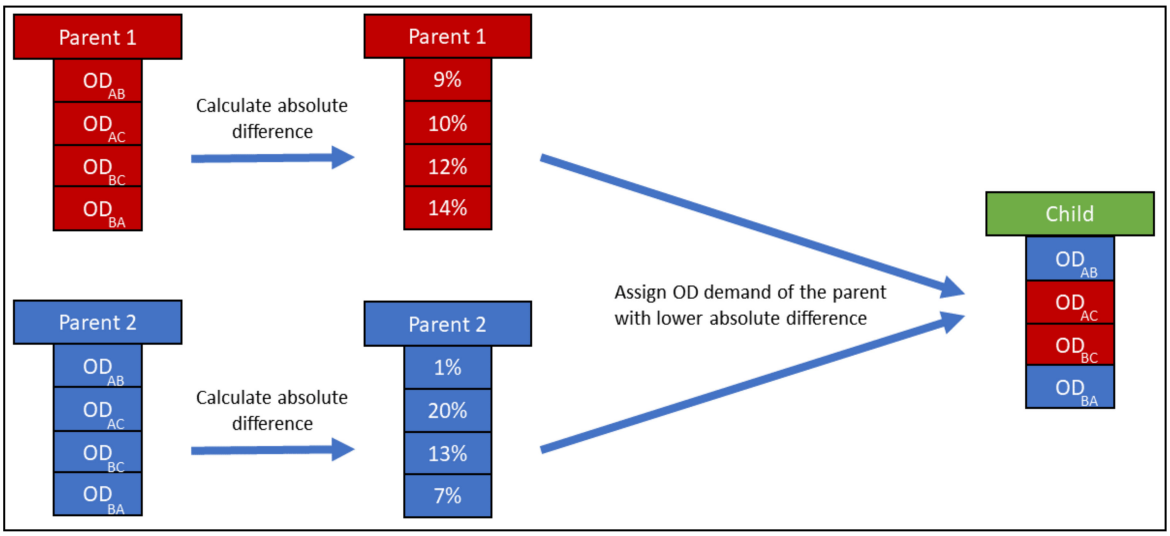



**Figure 2.** Methodology for crossover in the genetic algorithm.

Crossover method execution is subject to a user-defined crossover probability. In the instances where the crossover method is skipped, one of the parents is chosen to become the child.

Mutation

The final step of chromosome creation is a mutation, where the OD pair demand can be changed. The mutation is crucial to ensuring the GA does not concentrate only in a local search space but rather can explore the entire search space. However, too much mutation will cause the GA to be unstable and will stop it from converging. Thus, as with the crossover, there is a user-defined mutation probability that will limit mutations at the generation level. Along with this, there is also user-defined mutation variability which dictates how many of the traits of a child will be mutated. Once again, Rapidex provides multiple mutation methods from which the user has the flexibility to choose.

The swap method randomly chooses two OD pairs and swaps their demands. The bit method randomly chooses an OD pair and flips a bit of the binary number that represents the demand value. Finally, the random method selects an OD pair at random and assigns a value between a specified range.

After mutation, the chromosome is ready to be solved to convergence and have its error calculated. This iterative process continues until the genetic algorithm has achieved enough iterations. The process can be prematurely stopped if the error reaches an adequate value.

Termination of GA

The OD estimation module terminates after attaining the pre-defined error value or the maximum number of generations, whichever it reaches first. At the end of this module, Rapidex provides all the major outputs, namely link volume, link TT, and estimated OD-TT, in addition to the estimated OD matrix in the form of spreadsheets. Rapidex also allows the spatial visualisation of all these metrics on a background map. Additionally, Rapidex produces networkwide average metrics such as trip length, travel time, congestion levels (ratio of travel time to free-flow travel time), vehicle kilometres travelled, etc., which can be used for quick validation.

Due to the potential for having so many chromosomes and very large networks, completing the GA in a reasonable time required parallel computing. Solving user equilibrium is the most time intensive aspect due to the requirement of finding the shortest path tens, if not hundreds or thousands of times (depending on the relative gap) per chromosome. Thus, the Dijkstra algorithm has been implemented to leverage parallel computing to significantly speed up the process.

### *3.4. Scenario Testing*

Here, the user can modify the demand or network input parameters and evaluate the impacts relative to the base case. For example, the user can select a few links (either in isolation or as a corridor) in the network and change their attributes such as speed limit, number of lanes, and capacities. The user can also add a new set of links and nodes to the base network.

## 4. Results: Sydney Case Study

Now, we demonstrate the application of Rapidex on the network of Sydney, Australia. The bounding box coordinates encompassing the city of Sydney are inputted into Rapidex. Then we select the maximum and minimum grid sizes as 5 km and 1.25 km, respectively, in this case study. Overall, the network has 15,646 directional links, 8708 nodes, and 178 zones. Figure 3 shows the road network output.

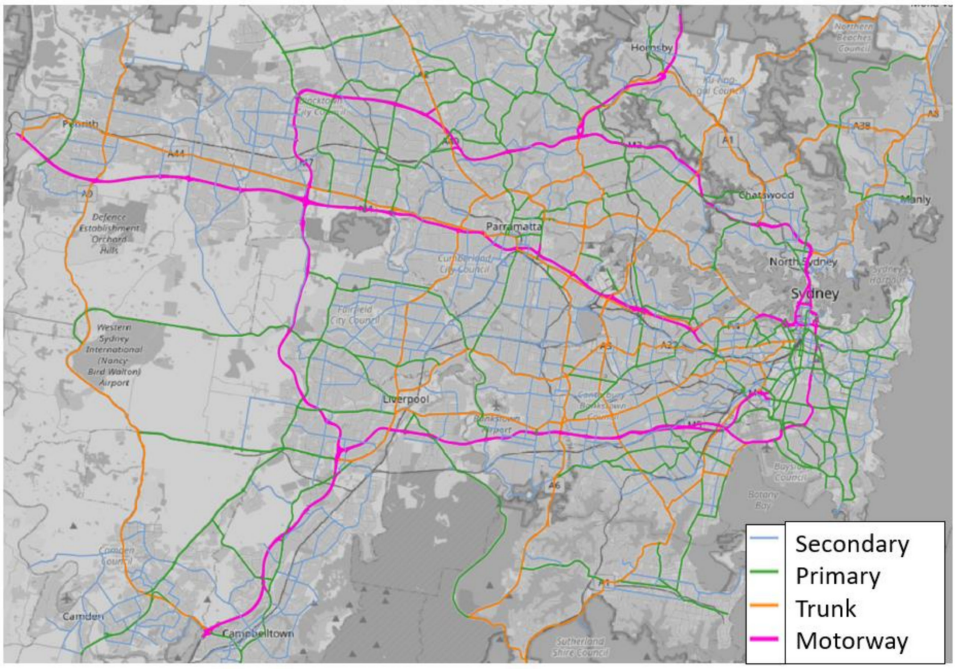



**Figure 3.** Road network output of Sydney, Australia.

For this case study, we collected the link travel times (from Google Maps) by setting the departure time as 8 a.m. on Wednesday, 31 March 2021. At the time of data collection, there were no travel restrictions within Sydney due to COVID-19. For the demand estimation, we defined a large input total demand range (100,000 to 2,000,000) so that the total demand estimated by Rapidex is always within these bounds. Then, we set 40 chromosomes (solutions) per generation, and the GA was set to run until a MAPE-ODTT value of 10% was reached. Therefore, the objective was to minimse the gap between the observed and estimated OD travel times. We used a server with 40 cores, 3.1 GHz processing speed and a memory of 512 GB.

While initialising, we allocated the same percentage of solutions to each initial solution method discussed in Table 2. In the custom gravity initial solution, we considered the proportion of nodes present within a zone as both $G$ and $A$, respectively. In this case study, we considered a relative gap value of 0.01. The Frank–Wolfe method was used to solve user equilibrium.

The networkwide average trip length, travel time, congestion level, and vehicle kilometres travelled were estimated as 11.9 km, 20.73 min, 1.66, and 7,875,392 veh-km, respectively. The total demand estimated was 657,000 vehicles during the morning peak hour. Figure 4 shows the scatterplot between the observed and estimated travel times for 31,506 OD pairs. Figure 5 shows the convergence of the GA algorithm, which took 26 generations to reach the pre-defined MAPE-ODTT error of 10%. This means a total of 1040 solutions (26 generations $\times$ 40 chromosomes per generation) were evaluated. The total processing time was approximately 4.5 h which included 0.5 h for network extraction from OSM and zoning and 4 h for demand estimation. It signifies the main objective of Rapidex, i.e., a tool for quick decision making by authorities. Nevertheless, the processing time could be further reduced with additional computing abilities and better initial solutions. For example, if a census-based zoning structure is used instead of the default grid-based system, the demographic data could be used as initial solutions, which could further speed up the process.

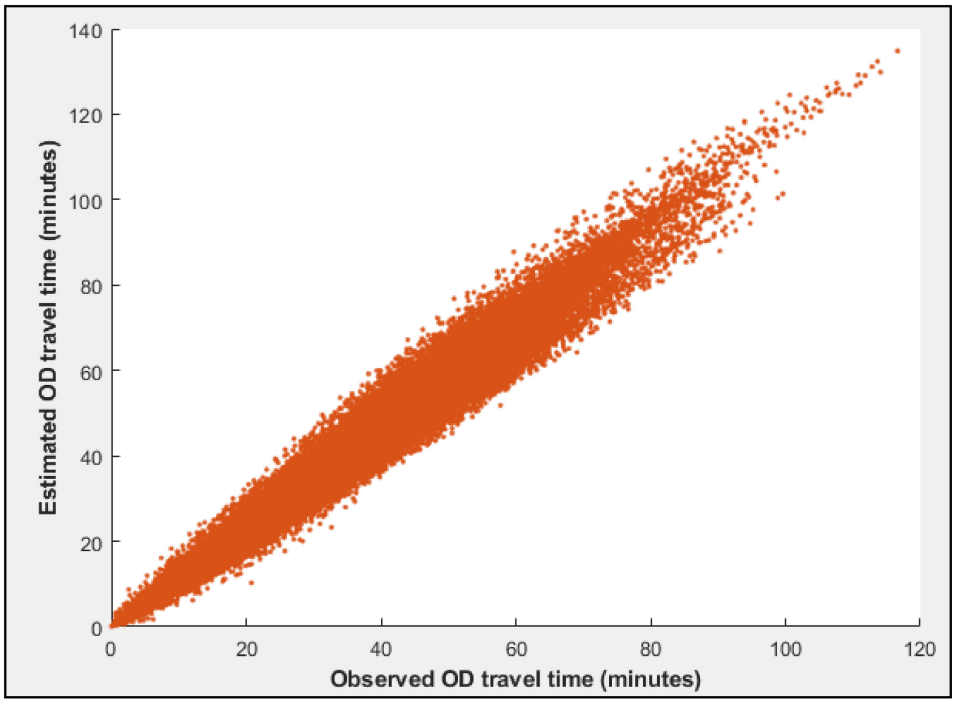



**Figure 4.** Observed vs. estimated OD travel times.

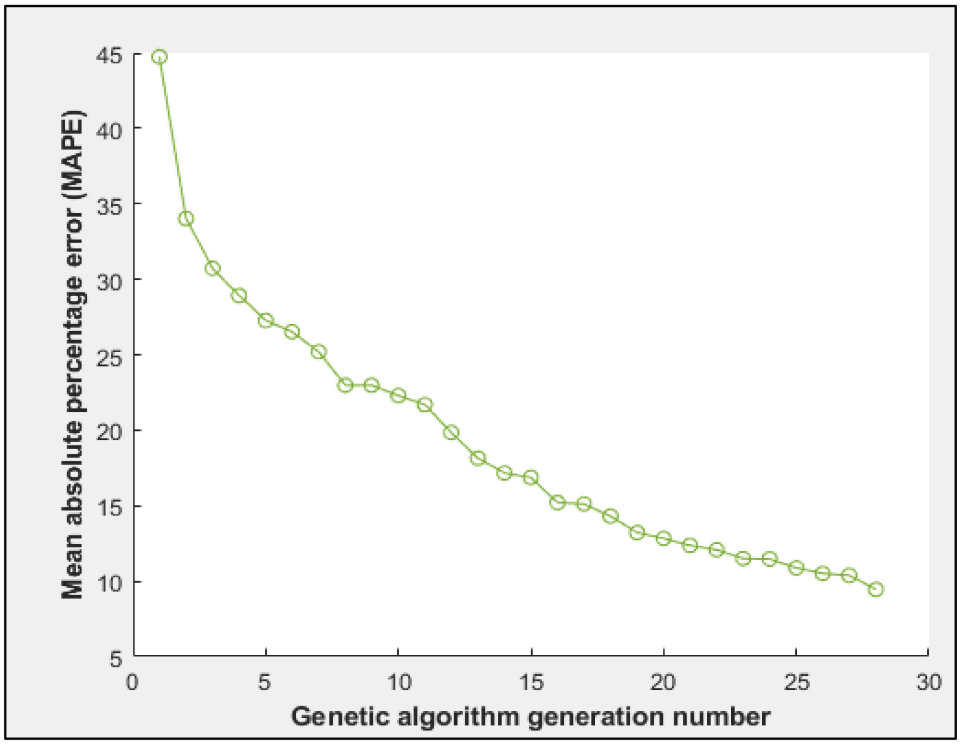



**Figure 5.** Convergence of the genetic algorithm solution.

Figures 6–9 show the zonal structure and the sample output produced by Rapidex. It can be noted that the areas closer to the central business district (CBD) and the ones with higher node density are automatically divided into smaller zones.

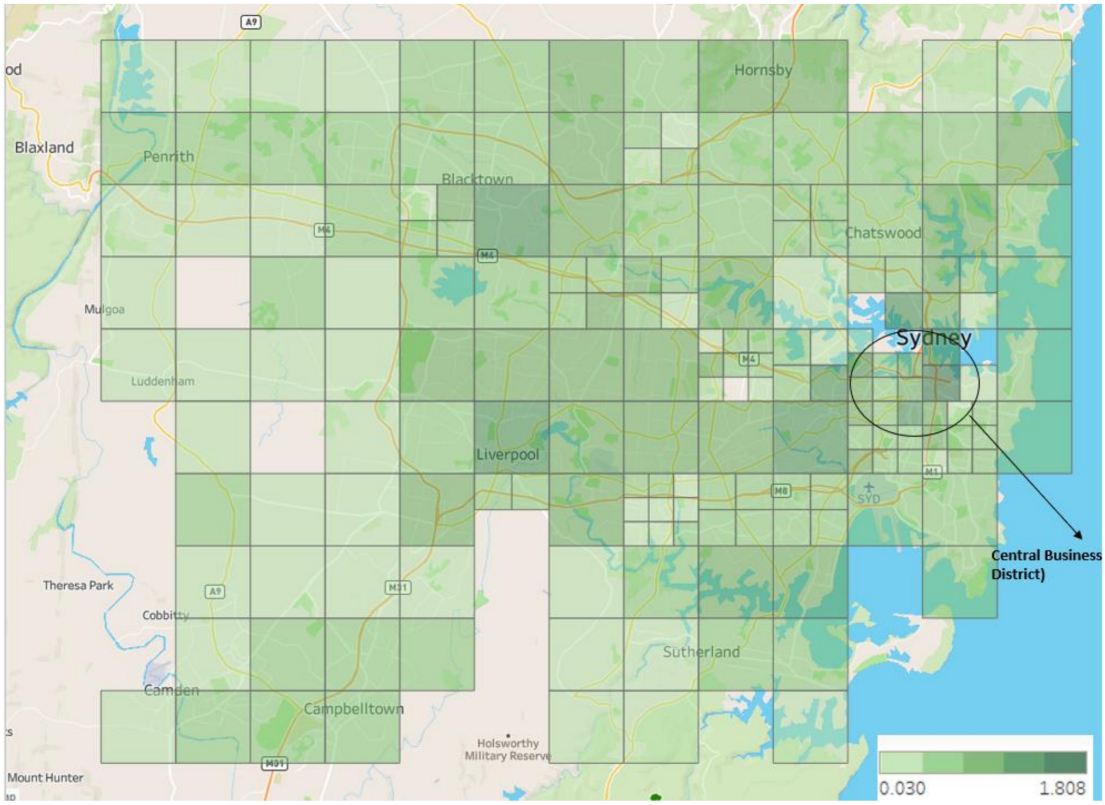



**Figure 6.** Zonal trip attraction (as a percentage of total demand).

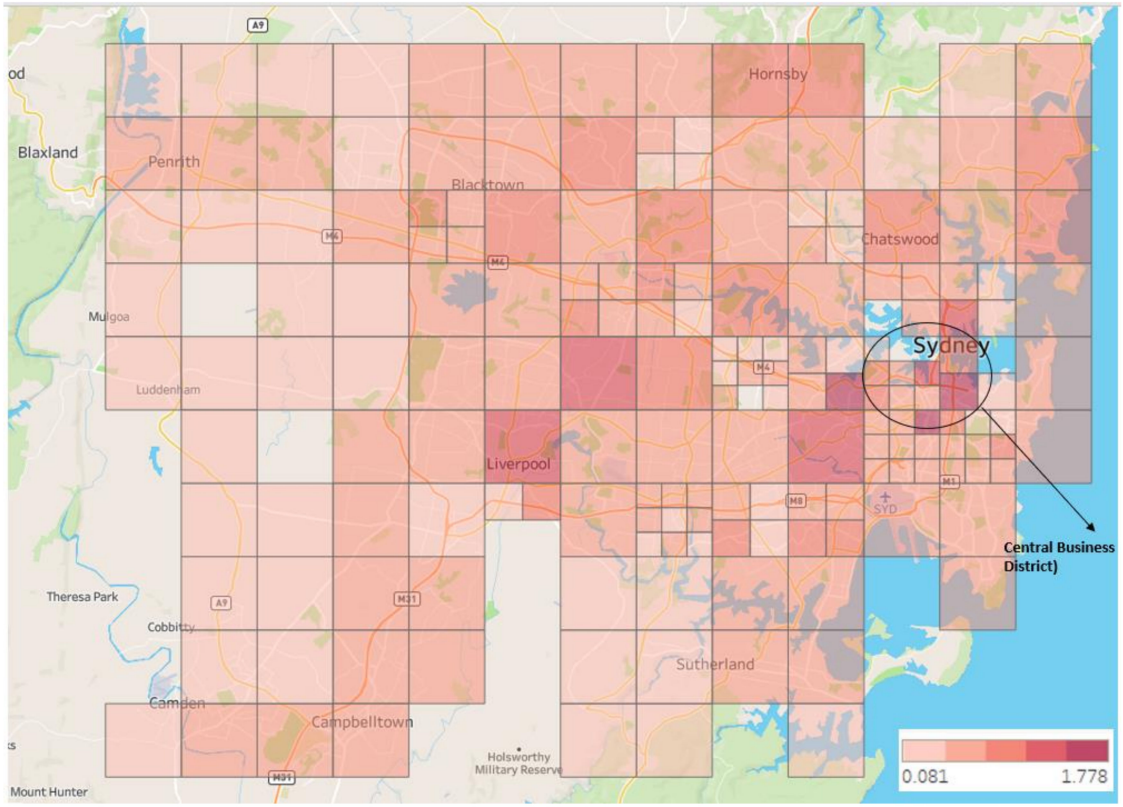



**Figure 7.** Zonal trip production (as a percentage of total demand).

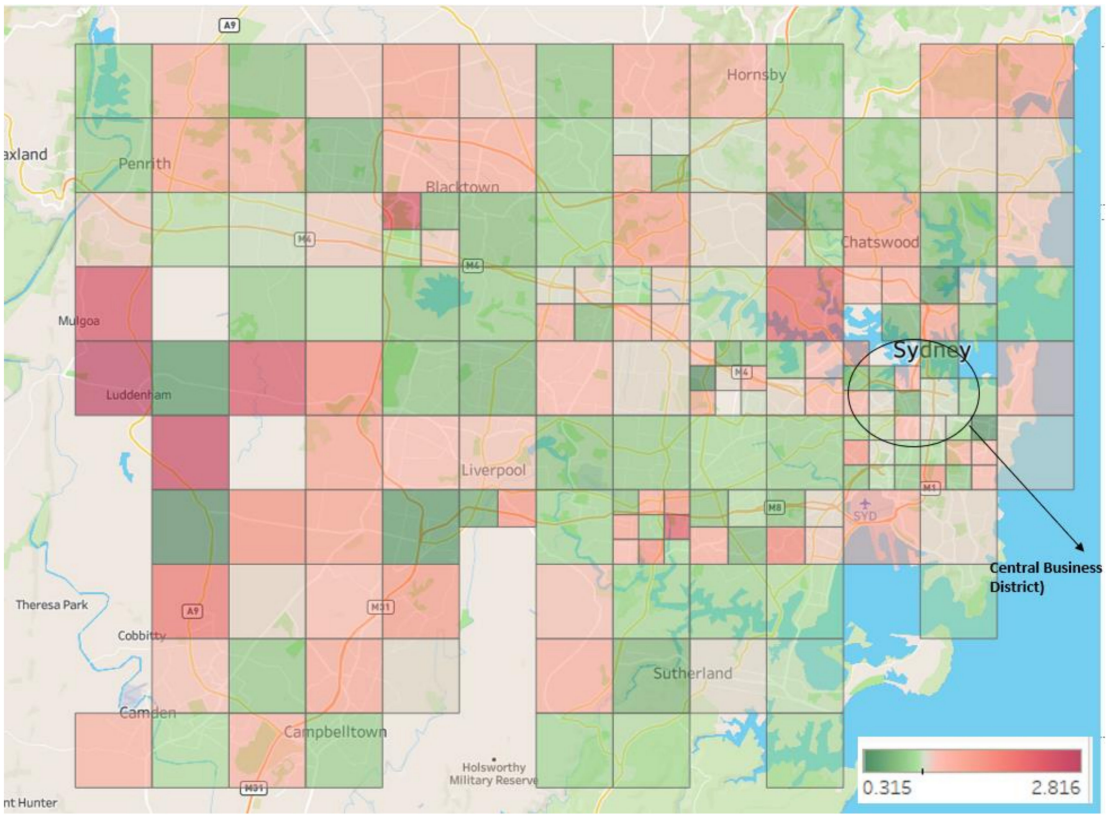



**Figure 8.** Zonal net contribution (the ratio of trip production and attraction).

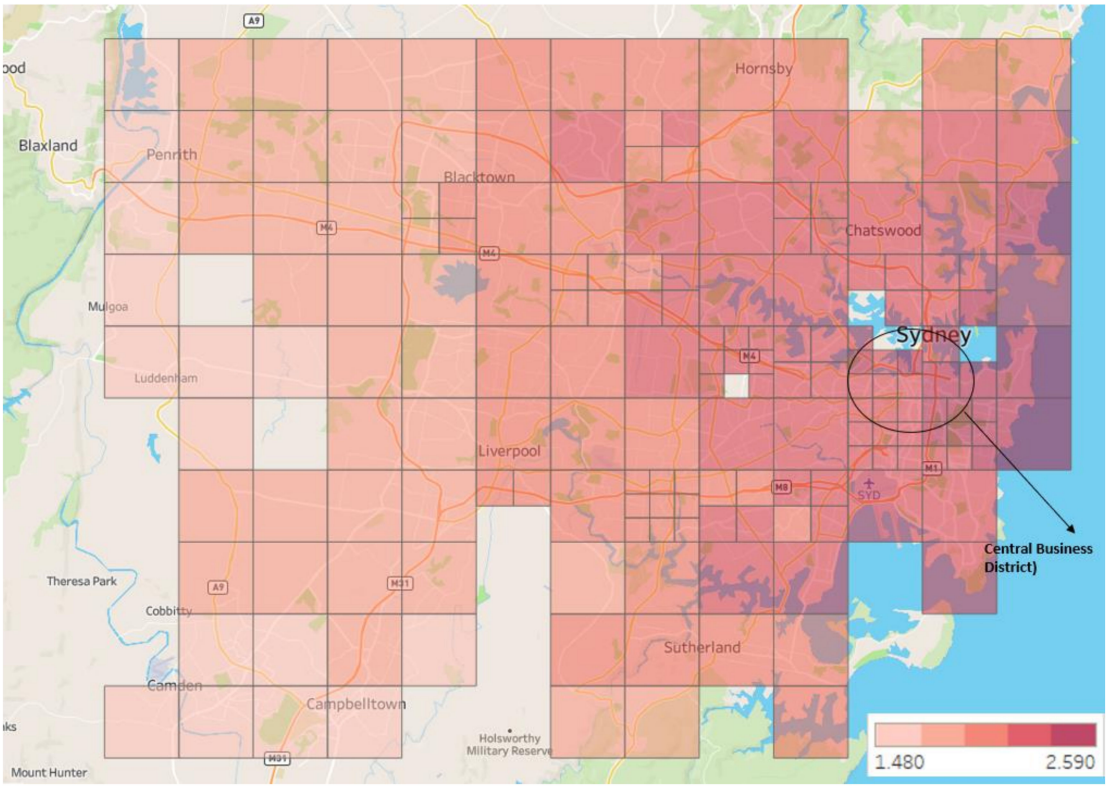



**Figure 9.** Zonal destination congestion index (the ratio of travel time to free-flow travel time to reach the zone from all other zones).

Figure 6 shows the percentage of demand attracted to each zone in the network. The darker the green, the higher the trip attraction of the zone. Figure 7 indicates the percentage of the total demand generated by each zone. In this figure, the darker red signifies that the zone generates a relatively higher amount of trips than the other zones. Figure 8 depicts the net contribution, i.e., the ratio of trip generations and attractions of the different zones. The dark red indicates that the zone produces more trips than it attracts during the morning peak hour. On the other hand, the dark green indicates more trips are ending in these zones than originating. It is evident from this figure that the zones far away from the CBD generate more trips than they attract during the morning peak hour, which is intuitive.

Finally, Figure 9 shows the average level of congestion (the ratio of travel time to free-flow travel time) to reach each zone in the network from every other zone during the morning peak hour. The dark red colour indicates that it takes a significantly longer time to reach such zones than others. It can be seen here that the zones closer to the CBD are a darker red than the ones that are further out. In the morning peak hour, one can expect more trips to be going towards the CBD and hence more congestion.

## 5. Validation

The output validation was performed using three approaches, first using the GEH (named for Geoffrey E. Havers) statistic, then using the aggregated output metrics and finally using a novel project selection method. The details are outlined below.

The Household Travel Survey (HTS) data, obtained at Statistical Area Level 2 (SA2) for Sydney for the year 2016, has been considered for validation. Therefore, the SA2 zoning system was used for validation purposes to align with the HTS data instead of the zoning structure of Rapidex. Figure 10 shows the SA2 zoning structure used for validation purposes. First, we solved the user equilibrium using the HTS data and recorded the OD travel times, which are treated as the "observed values". Using these values as input, we "estimated" the OD matrix using Rapidex.

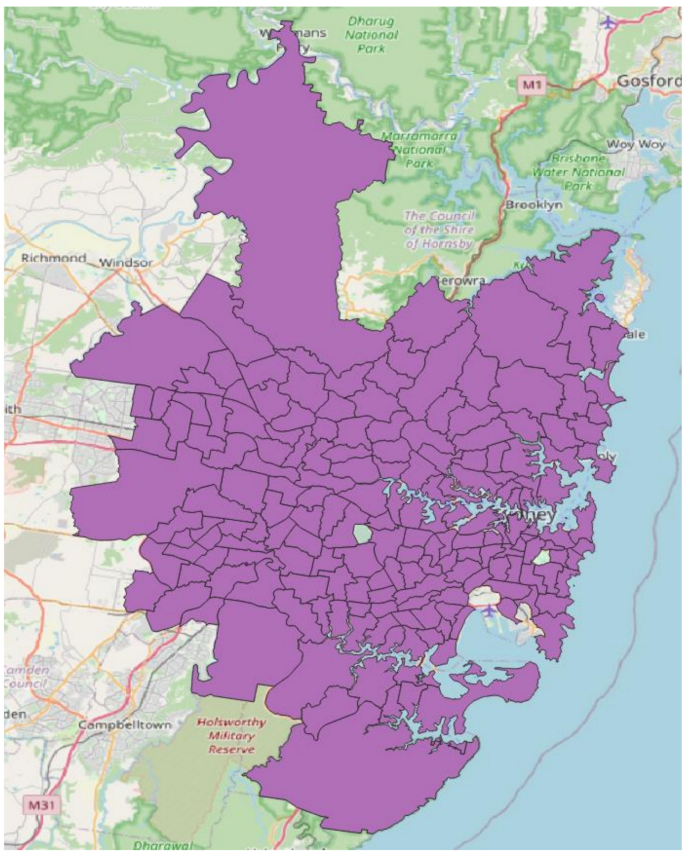



**Figure 10.** The SA2 zoning structure used for validation (total number of zones = 180).

However, several default properties of Rapidex are altered in the "observed" case to avoid any bias and to mimic the real-world conditions as much as possible. For example, in the "observed" case, the BPR parameters are set as random values ($\alpha$ in the range between 0.10 and 0.20 and $\beta$ in the range between 3.5 and 4.5), whereas for the "observed" case, the default parameters were used. Furthermore, the location and the number of centroids for each SA2 are altered. Finally, the capacity of each link in the "observed" case is multiplied by a random number between 0.75 and 1.25 to the default capacities. These changes are made because, in reality, one may not know the "true" link performance function, link capacities, and the centroid locations.

### *5.1. Validation Using the GEH Statistic*

The GEH statistic, outlined in the equation below, is typically used to to evaluate the performance of traffic models. Generally, a *GEH* value of less than 10 is considered acceptable in large-scale models [5,72].

$$GEH = \sqrt{\frac{2(M-C)^2}{M+C}}$$

In the current study, $M$ and $C$ can be treated as the modelled/estimated and observed traffic demand between an OD pair.

We calculated the GEH statistic for every OD pair, i.e., 32,200 observations. Figure 11 shows the frequency distribution of the GEH statistic. It can be seen that approximately 80% of the OD pairs have a GEH value less than 5, and over 93% have a GEH value less than 10.

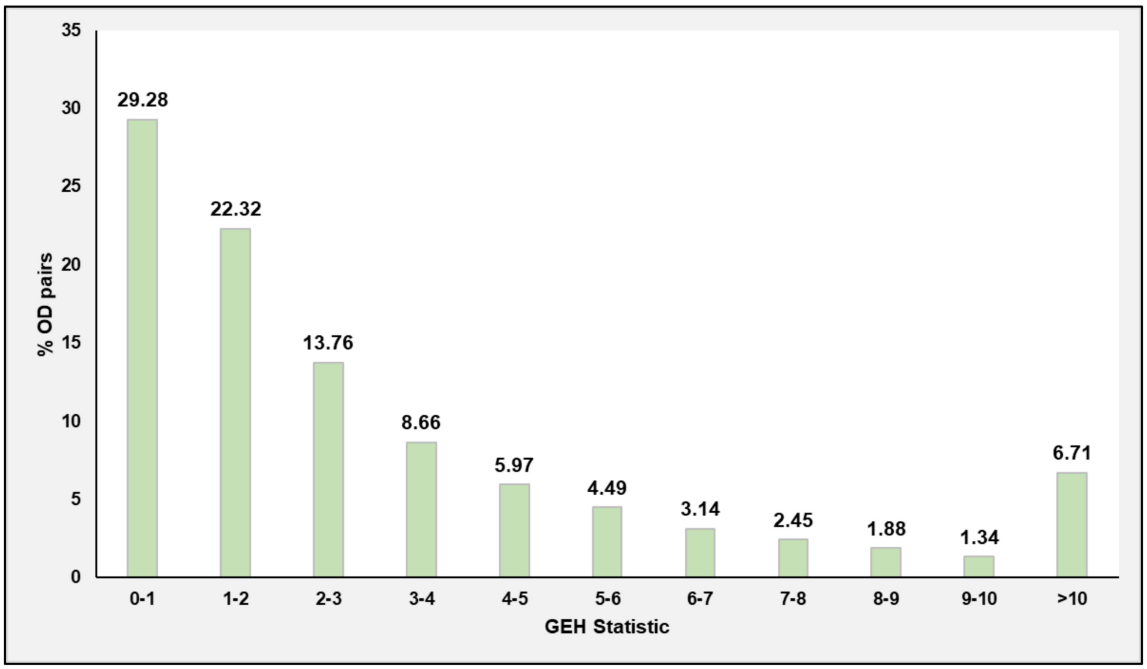


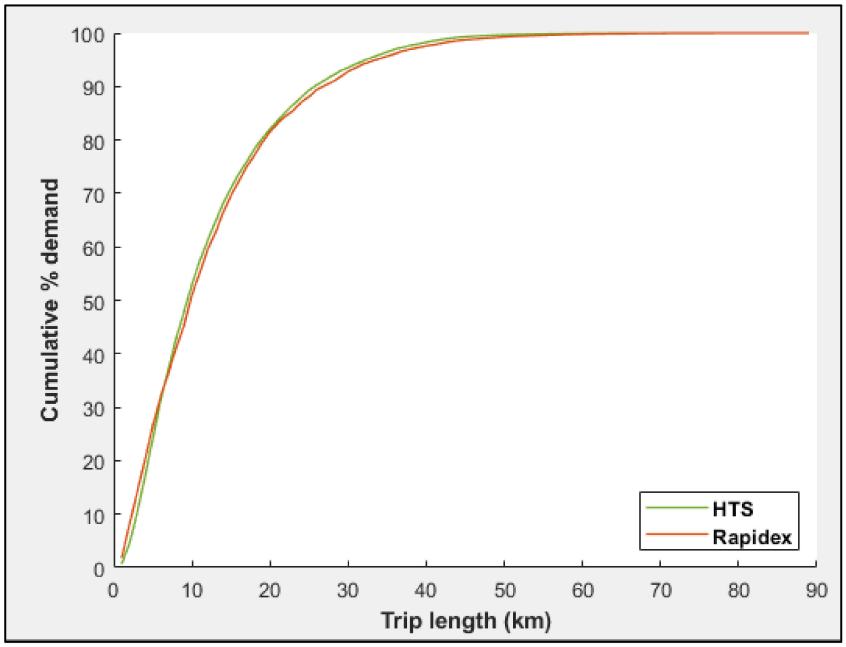

**Figure 11.** Frequency distribution of GEH statistic for every OD pair.

A further granular analysis of the GEH statistic in terms of zonal trip productions and attractions was conducted. All the 180 zones have their trip productions and attractions estimated within a GEH value of 10. Only two zones have the trip production GEH value above 5, and 16 (i.e., 9%) zones have the trip attraction GEH above 5.

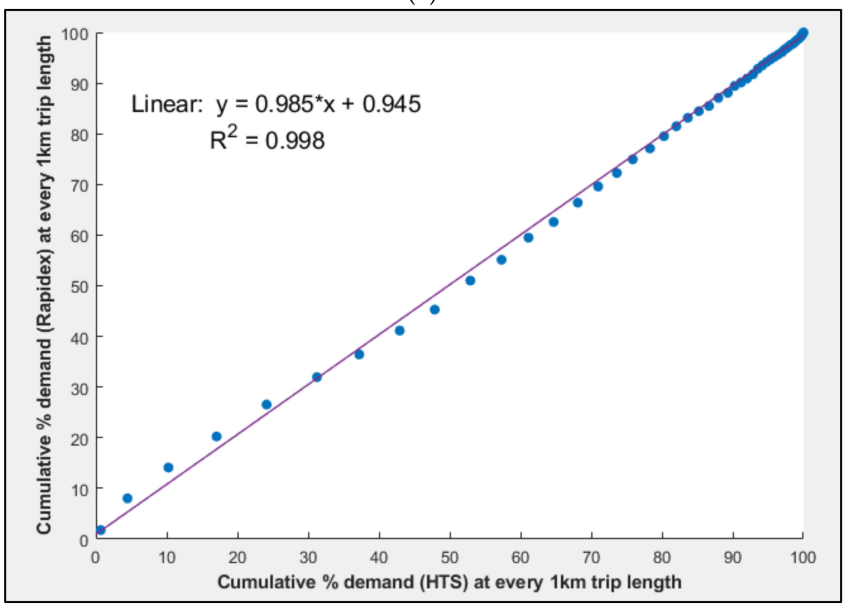
### *5.2. Validation Using Aggregated Metrics*

Aggregated metrics, namely trip length (and its distribution), congestion levels, and travel times, are used to evaluate the performance of the Rapidex approach. These ag-

gregated metrics, particularly the trip length distribution, are essential in calibrating and validating large-scale networks [73,74]. The "observed" values for total system travel time (TSTT), average trip length, average trip time, and congestion were 174,792 h, 11.6 km, 20 min, and 1.64, respectively. The corresponding values for the "estimated" case are 171,899 h, 11.4 km, 22 min, and 1.64, respectively. Furthermore, Figure 12a–d show the comparison of trip length and travel time distributions of the observed and estimated cases. It can be seen that the "estimated" trip length distribution matches perfectly with that of the "observed" case. Even the "estimated" travel time distribution aligns well with that of the "observed" case until the 70th percentile, after which Rapidex slightly underestimates the travel time. It could be due to the stopping criteria of the error function, i.e., MAPE-ODTT, which was set as 10%.

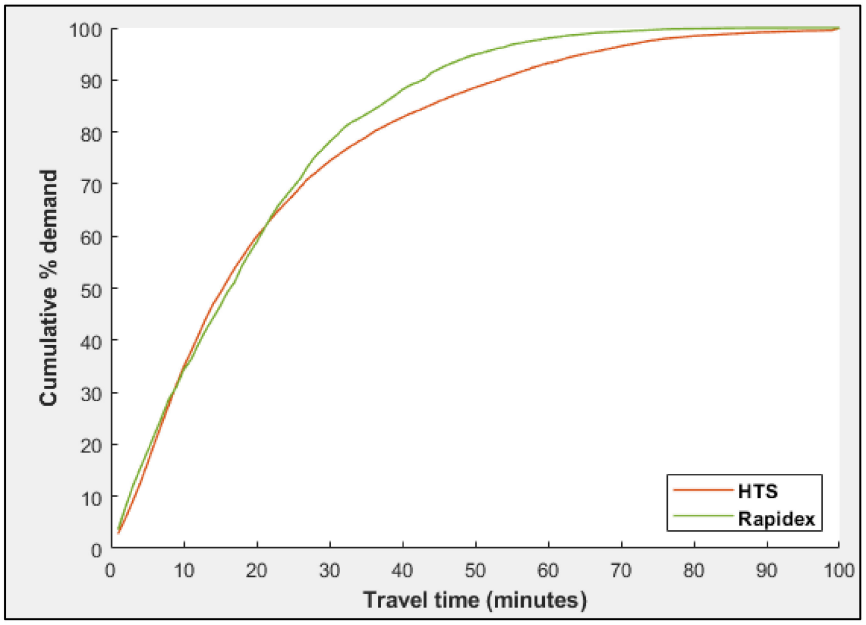



(**a**)

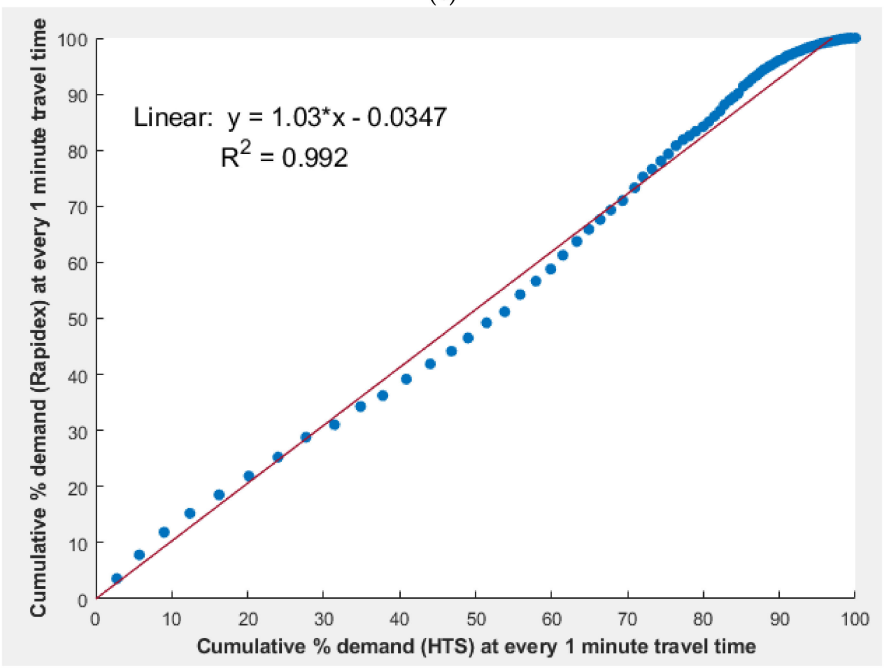



(**b**)

**Figure 12.** *Cont.*




(**c**)




(**d**)

**Figure 12.** Comparison of trip length and travel time distributions (HTS vs. Rapide). (**a**) Trip length distribution comparison (HTS vs. Rapidex); (**b**) trip length fit comparison (HTS vs. Rapidex); (**c**) travel time distribution comparison (HTS vs. Rapidex); and (**d**) travel time fit comparison (HTS vs. Rapidex).

### *5.3. Validation Using the Project Selection Method*

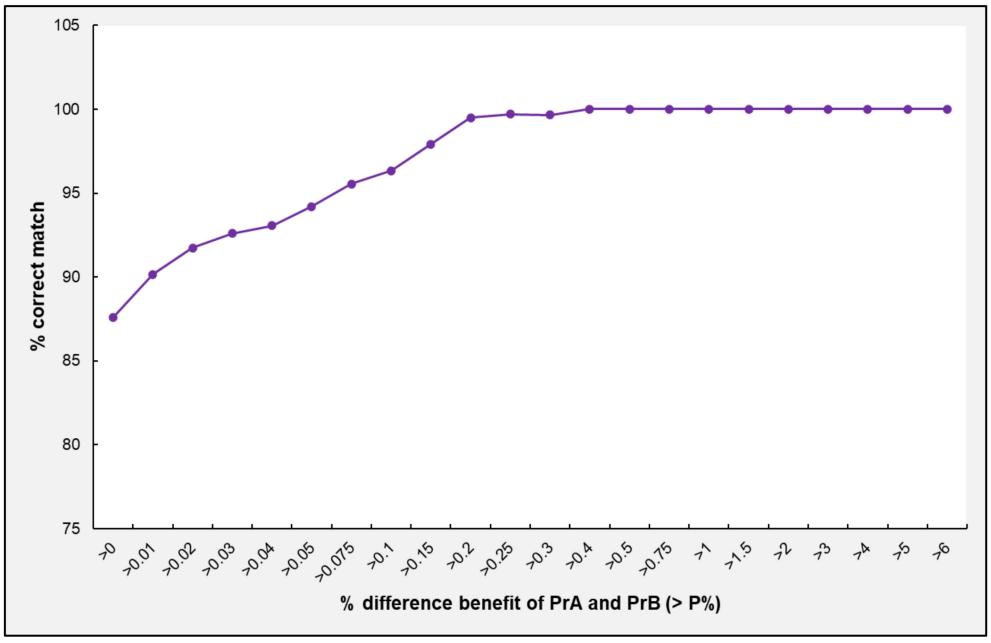
Major transportation projects are costly investments, which are aimed at improving the existing conditions. When multiple projects are in consideration, one must carefully prioritise the appropriate project, keeping in mind the budget constraints and the long-term project impacts [75]. The ultimate purpose of estimating OD matrices (and transportation planning models in general) for any network is to evaluate different transportation scenarios or/and infrastructure design projects [3,76].

In this second validation approach, we want to see how accurately the "estimated" matrix can perform in terms of project selection when compared with the "observed" OD matrix. We consider two projects, A and B, which are outlined below.

- **Project A—Increasing capacity**: Select a random corridor *i* (out of 69 major corridors, including key motorways and arterials in Sydney) and increase the capacity of all links along this corridor by a random number (set in the range of 100–1500 veh/h).
- **Project A—Cost**: $= C_{YA} * L_Y$,
  - where $C_{YA}$ is the capacity increment of route Y for Project A and $L_Y$ is the length of route Y.
- **Project B—Capacity increment:** Select a different random corridor X. The capacity increment for Project B is calculated such that both the projects cost an equal amount as $C_{XB} = \frac{C_{YA} * L_Y}{L_X}$,
  - where $C_{XB}$ is the capacity increment of route X for Project B and $L_X$ is the length of route X.
  - Here we assume that the project's cost is purely in terms of the capacity increment.

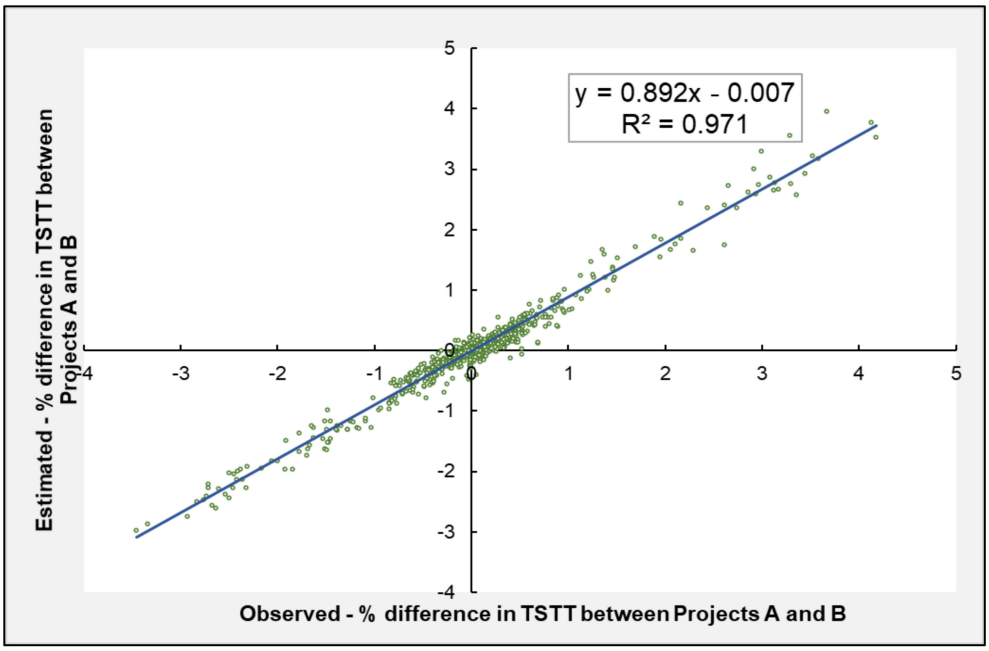
The process to select a better project between the two is to solve user equilibrium and evaluate the benefits in terms of networkwide metrics, e.g., TSTT. If the "estimated" and the "observed" OD matrices identify the same project as the better one, then we can say that the "estimated" OD matrix is useful for practical purposes. In this paper, we repeat the project selection 1000 times, i.e., each time we solve UE once for Project A and once for Project B, for both the observed and estimated cases. Thus, we run 4000 UE solutions. We observed that 88% of the time, the "estimated" and "observed" OD matrices identified the same project as the better one.

Further analysis (see Figure 13a) revealed that the projects were only ranked incorrectly when the difference between the benefits (in terms of TSTT) offered by Project A and Project B is insignificant. For example, when the percentage difference in benefits offered by the projects was greater than 0.1%, the projects were ranked correctly in 96% of the cases. Likewise, when the percentage difference increased to 0.25%, all the projects were ranked correctly. Furthermore, Figure 13b shows that the observed and estimated percentage difference of benefits between Project A and Project B align quite well.




(**a**)

**Figure 13.** *Cont.*




(**b**)

**Figure 13.** Validation using the project selection approach. (**a**) Percentage correct match of better project; and (**b**) percentage benefits comparison of Projects A and B (for observed and estimated cases).

All the three validation approaches show that the proposed Rapidex approach is appropriately validated.

## 6. Conclusions and Future Directions

This paper presented Rapidex, a new tool to estimate the trip table for any given city in the world. With the tool, users can download the road network, create zones, extract travel time data from pervasive data aggregators, estimate the OD matrix, produce critical outputs, and visualise them spatially on a map. Furthermore, the tool allows changes to be made to demand and network data to evaluate different scenarios. In the case study of Sydney, we saw some of the output produced by Rapidex, such as the maps of the road network, trip productions, attractions, and congestion levels. Furthermore, the model was validated using the HTS data of Sydney using the aggregated metrics and a project selection method.

Rapidex enables traffic authorities and practitioners to quickly make decisions regarding strategic long-term planning. It can help compare, contrast, and benchmark various policies by evaluating the trip tables across various networks. The tool can help in understanding day-to-day and within-day demand changes in a city, analysing the impacts of network and demand changes on critical performance metrics. Importantly, the tool can be beneficial for developing countries where urbanisation and population growth are rapid and yet lack adequate resources in capturing demand data.

A key limitation of most traditional existing demand estimation methods is the reliance on the limited number of observations of traffic counts, which are significantly lower than the unknown values (demand), causing the problem of under-determinacy. However, in Rapidex, the default objective function considers the travel times between every OD pair. There is also a provision to test the typically-used objective functions in the literature.

Significant future research is identified. For instance, the Rapidex tool is being updated and refined continuously. Currently, the project team is working towards including public transport and evaluating the impacts of the changes in demand and road capacity values on sustainability indicators, such as congestion, vehicle kilometres travelled, mode share, equity, etc., for multiple cities around the world. In addition, the sensitivity of different parameters such as centroid selection, zoning, GA parameters, etc., is being

tested. Furthermore, different user equilibrium traffic assignment heuristics are also being tested to further speed up the demand estimation process. Critically, it is noted that the current purpose of the tool is for quick decision making in long-term planning projects. Future research also involves developing a more dynamic version of the software that takes into account the fluctuations in within-day travel times so that it can also be used for operational purposes.

As with traditional aggregate traffic assignment approaches, a limitation of the overall demand value estimated by Rapidex is the reliance on link capacities (through the link performance function). The demand may be over or under estimated if the link capacities are set too high or low. Better solutions could be obtained if either the total demand or the link capacities are known. A limitation of Rapidex lies in the link performance function. Finally, it is important to note that the level of detail of OSM data and its data quality could differ across different countries and continents. Although Rapidex has a provision to import a pre-existing network file (in the form of a shapefile or spreadsheet), enhanced OSM data, particularly in developing countries, would make it even more accurate.

**Author Contributions:** Conceptualization, S.T.W.; methodology, S.T.W., S.C., A.Z. and V.V.D.; software, S.C., A.Z., D.N., C.N. and J.W.; validation, S.C. and A.Z.; formal analysis, S.C. and A.Z.; investigation, S.T.W., S.C., A.Z., D.N., C.N., J.W., X.Z. and V.V.D.; resources, S.T.W. and V.V.D.; data curation, S.C., A.Z. and D.N.; writing—original draft preparation, S.C. and A.Z.; writing—review and editing, S.T.W., S.C., A.Z., D.N., X.Z. and V.V.D.; visualization, S.C. and A.Z.; supervision, S.T.W. and V.V.D.; project administration, S.T.W.; funding acquisition, S.T.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to licensing restrictions from pervasive data providers.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. INRIX Scorecard. Available online: https://inrix.com/scorecard/ (accessed on 22 July 2020).
2. Ortmann, A.; Dixit, V.; Chand, S.; Jian, S. *Nudging towards a More Efficient Transportation System: A Review of Non-Pricing (Behavioural) Interventions*; Infrastructure Victoria: Melbourne, VIC, Australia, 2017.
3. Duell, M.; Waller, S.T. Implications of Volatility in Day-to-Day Travel Flow and Road Capacity on Traffic Network Design Projects. *Transp. Res. Rec.* **2015**, *2498*, 56–63. [CrossRef]
4. Sheffi, Y. *Urban Transportation Networks*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1985; Volume 6.
5. Duell, M.; Saxena, N.; Chand, S.; Amini, N.; Grzybowska, H.; Waller, S.T. Deployment and Calibration Considerations for Large-Scale Regional Dynamic Traffic Assignment: Case Study for Sydney, Australia. *Transp. Res. Rec. J. Transp. Res. Board* **2016**, *2567*, 78–86. [CrossRef]
6. Duthie, J.C.; Nezamuddin, N.; Juri, N.R.; Rambha, T.; Melson, C.; Pool, C.M.; Boyles, S.; Waller, S.T.; Kumar, R. *Investigating Regional Dynamic Traffic Assignment Modeling for Improved Bottleneck Analysis: Final Report*; Center for Transportation Research at The University of Texas at Austin: Austin, TX, USA, 2013.
7. Jafari, E.; Gemar, M.D.; Juri, N.R.; Duthie, J. Investigation of Centroid Connector Placement for Advanced Traffic Assignment Models with Added Network Detail. *Transp. Res. Rec.* **2015**, *2498*, 19–26. [CrossRef]
8. Haklay, M.; Weber, P. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]
9. Rapelli, M.; Casetti, C.; Gagliardi, G. Vehicular Traffic Simulation in the City of Turin from Raw Data. *IEEE Trans. Mob. Comput.* **2021**. [CrossRef]
10. Xia, F.; Rahim, A.; Kong, X.; Wang, M.; Cai, Y.; Wang, J. Modeling and Analysis of Large-Scale Urban Mobility for Green Transportation. *IEEE Trans. Ind. Inform.* **2018**, *14*, 1469–1481. [CrossRef]
11. Yedavalli, P.; Kumar, K.; Waddell, P. Microsimulation Analysis for Network Traffic Assignment (MANTA) at Metropolitan-Scale for Agile Transportation Planning. *Transp. A Transp. Sci.* **2021**. [CrossRef]
12. Chand, S.; Li, Z.; Dixit, V.V.; Travis Waller, S. Examining the Macro-Level Factors Affecting Vehicle Breakdown Duration. *Int. J. Transp. Sci. Technol.* **2021**. [CrossRef]

13. Schiefelbein, J.; Rudnick, J.; Scholl, A.; Remmen, P.; Fuchs, M.; Müller, D. Automated Urban Energy System Modeling and Thermal Building Simulation Based on OpenStreetMap Data Sets. *Build. Environ.* **2019**, *149*, 630–639. [CrossRef]
14. Kunkler, J.; Braun, M.; Kellner, F. Speed Limit Induced CO2 Reduction on Motorways: Enhancing Discussion Transparency through Data Enrichment of Road Networks. *Sustainability* **2021**, *13*, 395. [CrossRef]
15. Alarabi, L.; Eldawy, A.; Alghamdi, R.; Mokbel, M.F. TAREEG: A MapReduce-Based Web Service for Extracting Spatial Data from OpenStreetMap. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, Snowbird, UT, USA, 18 June 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 897–900.
16. Boeing, G. OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks. *Comput. Environ. Urban Syst.* **2017**, *65*, 126–139. [CrossRef]
17. Huber, S.; Rust, C. Calculate Travel Time and Distance with Openstreetmap Data Using the Open Source Routing Machine (OSRM). *Stata J.* **2016**, *16*, 416–423. [CrossRef]
18. Raifer, M.; Troilo, R.; Kowatsch, F.; Auer, M.; Loos, L.; Marx, S.; Przybill, K.; Fendrich, S.; Mocnik, F.-B.; Zipf, A. OSHDB: A Framework for Spatio-Temporal Analysis of OpenStreetMap History Data. *Open Geospat. Data Softw. Stand.* **2019**, *4*, 3. [CrossRef]
19. Cohn TomTom Traffic Index: Measuring Urban Traffic Congestion | TomTom Blog. Available online: https://www.tomtom.com/blog/traffic-and-travel-information/urban-traffic-congestion/ (accessed on 10 June 2021).
20. Aboudina, A.; Abdelgawad, H.; Abdulhai, B.; Habib, K.N. Time-Dependent Congestion Pricing System for Large Networks: Integrating Departure Time Choice, Dynamic Traffic Assignment and Regional Travel Surveys in the Greater Toronto Area. *Transp. Res. Part A Policy Pract.* **2016**, *94*, 411–430. [CrossRef]
21. Zhang, L.; Yang, D.; Ghader, S.; Carrion, C.; Xiong, C.; Rossi, T.F.; Milkovits, M.; Mahapatra, S.; Barber, C. An Integrated, Validated, and Applied Activity-Based Dynamic Traffic Assignment Model for the Baltimore-Washington Region. *Transp. Res. Rec.* **2018**, *2672*, 45–55. [CrossRef]
22. Stopher, P.R.; Greaves, S.P. Household Travel Surveys: Where Are We Going? *Transp. Res. Part A Policy Pract.* **2007**, *41*, 367–381. [CrossRef]
23. Antoniou, C.; Barceló, J.; Breen, M.; Bullejos, M.; Casas, J.; Cipriani, E.; Ciuffo, B.; Djukic, T.; Hoogendoorn, S.; Marzano, V.; et al. Towards a Generic Benchmarking Platform for Origin–Destination Flows Estimation/Updating Algorithms: Design, Demonstration and Validation. *Transp. Res. Part C Emerg. Technol.* **2016**, *66*, 79–98. [CrossRef]
24. Nair, D.J.; Gilles, F.; Chand, S.; Saxena, N.; Dixit, V. Characterizing Multicity Urban Traffic Conditions Using Crowdsourced Data. *PLoS ONE* **2019**, *14*, e0212845. [CrossRef]
25. Respati, S.; Bhaskar, A.; Chung, E. Traffic Data Characterisation: Review and Challenges. *Transp. Res. Procedia* **2018**, *34*, 131–138. [CrossRef]
26. Dixit, V.; Nair, D.J.; Chand, S.; Levin, M.W. A Simple Crowdsourced Delay-Based Traffic Signal Control. *PLoS ONE* **2020**, *15*, e0230598. [CrossRef]
27. Lin, Y.; Li, R. Real-Time Traffic Accidents Post-Impact Prediction: Based on Crowdsourcing Data. *Accid. Anal. Prev.* **2020**, *145*, 105696. [CrossRef]
28. Osorio, C. Dynamic Origin-Destination Matrix Calibration for Large-Scale Network Simulators. *Transp. Res. Part C Emerg. Technol.* **2019**, *98*, 186–206. [CrossRef]
29. Bauer, D.; Richter, G.; Asamer, J.; Heilmann, B.; Lenz, G.; Kölbl, R. Quasi-Dynamic Estimation of OD Flows From Traffic Counts Without Prior OD Matrix. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 2025–2034. [CrossRef]
30. Abrahamsson, T. *Estimation of Origin-Destination Matrices Using Traffic Counts—A Literature Survey*; Working Papers; International Institute for Applied Systems Analysis: Laxenburg, Austria, 1998.
31. Bera, S.; Rao, K.V.K. Estimation of Origin-Destination Matrix from Traffic Counts: The State of the Art. *Eur. Transp. Trasp. Eur.* **2011**, *49*, 2–23.
32. Duan, Z.; Zhang, K.; Chen, Z.; Liu, Z.; Tang, L.; Yang, Y.; Ni, Y. Prediction of City-Scale Dynamic Taxi Origin-Destination Flows Using a Hybrid Deep Neural Network Combined With Travel Time. *IEEE Access* **2019**, *7*, 127816–127832. [CrossRef]
33. Hai, Y.; Akiyama, T.; Sasaki, T. Estimation of Time-Varying Origin-Destination Flows from Traffic Counts: A Neural Network Approach. *Math. Comput. Model.* **1998**, *27*, 323–334. [CrossRef]
34. Barcelö, J.; Montero, L.; Marqués, L.; Carmona, C. Travel Time Forecasting and Dynamic Origin-Destination Estimation for Freeways Based on Bluetooth Traffic Monitoring. *Transp. Res. Rec.* **2010**, *2175*, 19–27. [CrossRef]
35. Tesselkin, A.; Khabarov, V. Estimation of Origin-Destination Matrices Based on Markov Chains. *Procedia Eng.* **2017**, *178*, 107–116. [CrossRef]
36. Kim, H.; Baek, S.; Lim, Y. Origin-Destination Matrices Estimated with a Genetic Algorithm from Link Traffic Counts. *Transp. Res. Rec.* **2001**, *1771*, 156–163. [CrossRef]
37. Saadi, I.; Mustafa, A.; Teller, J.; Cools, M. A Bi-Level Random Forest Based Approach for Estimating O-D Matrices: Preliminary Results from the Belgium National Household Travel Survey. *Transp. Res. Procedia* **2017**, *25*, 2566–2573. [CrossRef]
38. Krishnakumari, P.; van Lint, H.; Djukic, T.; Cats, O. A Data Driven Method for OD Matrix Estimation. *Transp. Res. Part C Emerg. Technol.* **2020**, *113*, 38–56. [CrossRef]
39. Cantelmo, G.; Viti, F. A Big Data Demand Estimation Model for Urban Congested Networks. *Transp. Telecommun.* **2020**, *21*, 4.
40. Van Zuylen, H.J.; Willumsen, L.G. The Most Likely Trip Matrix Estimated from Traffic Counts. *Transp. Res. Part B Methodol.* **1980**, *14*, 281–293. [CrossRef]

41. Hazelton, M.L. Statistical Inference for Time Varying Origin–Destination Matrices. *Transp. Res. Part B Methodol.* **2008**, *42*, 542–552. [CrossRef]
42. Maher, M.J. Inferences on Trip Matrices from Observations on Link Volumes: A Bayesian Statistical Approach. *Transp. Res. Part B: Methodol.* **1983**, *17*, 435–447. [CrossRef]
43. Cascetta, E. Estimation of Trip Matrices from Traffic Counts and Survey Data: A Generalized Least Squares Estimator. *Transp. Res. Part B: Methodol.* **1984**, *18*, 289–299. [CrossRef]
44. Bell, M.G.H. The Estimation of Origin-Destination Matrices by Constrained Generalised Least Squares. *Transp. Res. Part B: Methodol.* **1991**, *25*, 13–22. [CrossRef]
45. Spiess, H. A Maximum Likelihood Model for Estimating Origin-Destination Matrices. *Transp. Res. Part B Methodol.* **1987**, *21*, 395–412. [CrossRef]
46. Dixon, M.P.; Rilett, L.R. Population Origin–Destination Estimation Using Automatic Vehicle Identification and Volume Data. *J. Transp. Eng.* **2005**, *131*, 75–82. [CrossRef]
47. Stathopoulos, A.; Tsekeris, T. Framework for Analysing Reliability and Information Degradation of Demand Matrices in Extended Transport Networks. *Transp. Rev.* **2003**, *23*, 89–103. [CrossRef]
48. Yang, F.; Jin, P.J.; Cheng, Y.; Zhang, J.; Ran, B. Origin-Destination Estimation for Non-Commuting Trips Using Location-Based Social Networking Data. *Int. J. Sustain. Transp.* **2015**, *9*, 551–564. [CrossRef]
49. Liu, H.X.; He, X.; Recker, W. Estimation of the Time-Dependency of Values of Travel Time and Its Reliability from Loop Detector Data. *Transp. Res. Part B Methodol.* **2007**, *41*, 448–461. [CrossRef]
50. Alexander, L.; Jiang, S.; Murga, M.; González, M.C. Origin–Destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data. *Transp. Res. Part C Emerg. Technol.* **2015**, *58*, 240–250. [CrossRef]
51. Calabrese, F.; Di Lorenzo, G.; Liu, L.; Ratti, C. Estimating Origin-Destination Flows Using Opportunistically Collected Mobile Phone Location Data from One Million Users in Boston Metropolitan Area. *Estim. Orig. Destin. Flows Using Mob. Phone Locat. Data* **2011**, *4*, 36–44.
52. Behara, K.N.S.; Bhaskar, A.; Chung, E. Single-Level Approach to Estimate Origin-Destination Matrix: Exploiting Turning Proportions and Partial OD Flows. *Transp. Lett.* **2021**, 1–12. [CrossRef]
53. Michau, G.E.; Nantes, A.; Chung, E.; Abry, P.; Borgnat, P. Retrieving Dynamic Origin-Destination Matrices from Bluetooth Data. In Proceedings of the Transportation Research Board (TRB) 93rd Annual Meeting Compendium of Papers, Washington, DC, USA, 12–16 January 2014; Petty, K., Ed.; Transportation Research Board (TRB): Washington, DC, USA, 2014; pp. 1–11.
54. Villiers, C.; Nguyen, L.D.; Zalewski, J. Evaluation of Traffic Management Strategies for Special Events Using Probe Data. *Transp. Res. Interdiscip. Perspect.* **2019**, *2*, 100052. [CrossRef]
55. Cheng, Z.; Jian, S.; Rashidi, T.H.; Maghrebi, M.; Waller, S.T. Integrating Household Travel Survey and Social Media Data to Improve the Quality of OD Matrix: A Comparative Case Study. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 2628–2636. [CrossRef]
56. Lee, J.H.; Davis, A.; McBride, E.; Goulias, K.G. Chapter 9—Statewide Comparison of Origin-Destination Matrices Between California Travel Model and Twitter. In *Mobility Patterns, Big Data and Transport Analytics*; Antoniou, C., Dimitriou, L., Pereira, F., Eds.; Elsevier: Amsterdam, The Netherlands, 2019; pp. 201–228. ISBN 978-0-12-812970-8.
57. Liao, Y.; Yeh, S.; Gil, J. Feasibility of Estimating Travel Demand Using Geolocations of Social Media Data. *Transportation* **2021**, 1–25. [CrossRef]
58. Twitter Tweet Geospatial Metadata. Available online: https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata (accessed on 18 June 2021).
59. Moreira-Matias, L.; Gama, J.; Ferreira, M.; Mendes-Moreira, J.; Damas, L. Time-Evolving O-D Matrix Estimation Using High-Speed GPS Data Streams. *Expert Syst. Appl.* **2016**, *44*, 275–288. [CrossRef]
60. Mungthanya, W.; Phithakkitnukoon, S.; Demissie, M.G.; Kattan, L.; Veloso, M.; Bento, C.; Ratti, C. Constructing Time-Dependent Origin-Destination Matrices With Adaptive Zoning Scheme and Measuring Their Similarities With Taxi Trajectory Data. *IEEE Access* **2019**, *7*, 77723–77737. [CrossRef]
61. Rao, W.; Wu, Y.-J.; Xia, J.; Ou, J.; Kluger, R. Origin-Destination Pattern Estimation Based on Trajectory Reconstruction Using Automatic License Plate Recognition Data. *Transp. Res. Part C Emerg. Technol.* **2018**, *95*, 29–46. [CrossRef]
62. Mo, B.; Li, R.; Dai, J. Estimating Dynamic Origin–Destination Demand: A Hybrid Framework Using License Plate Recognition Data. *Comput. Aided Civ. Infrastruct. Eng.* **2020**, *35*, 734–752. [CrossRef]
63. Dabbas, H.; Fourati, W.; Friedrich, B. Floating Car Data for Traffic Demand Estimation—Field and Simulation Studies. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020; pp. 1–8.
64. Meng, C.; Yi, X.; Su, L.; Gao, J.; Zheng, Y. City-Wide Traffic Volume Inference with Loop Detector Data and Taxi Trajectories. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 7–10 November 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 1–10.
65. Zhang, Z.; Li, M.; Lin, X.; Wang, Y. Network-Wide Traffic Flow Estimation with Insufficient Volume Detection and Crowdsourcing Data. *Transp. Res. Part C Emerg. Technol.* **2020**, *121*, 102870. [CrossRef]
66. Yang, H.; Sasaki, T.; Iida, Y.; Asakura, Y. Estimation of Origin-Destination Matrices from Link Traffic Counts on Congested Networks. *Transp. Res. Part B Methodol.* **1992**, *26*, 417–434. [CrossRef]

67. Yang, H. Heuristic Algorithms for the Bilevel Origin-Destination Matrix Estimation Problem. *Transp. Res. Part B Methodol.* **1995**, *29*, 231–242. [CrossRef]
68. Yin, Y. Genetic-Algorithms-Based Approach for Bilevel Programming Models. *J. Transp. Eng.* **2000**, *126*, 115–120. [CrossRef]
69. Ou, J.; Lu, J.; Xia, J.; An, C.; Lu, Z. Learn, Assign, and Search: Real-Time Estimation of Dynamic Origin-Destination Flows Using Machine Learning Algorithms. *IEEE Access* **2019**, *7*, 26967–26983. [CrossRef]
70. Goldberg, D.E.; Deb, K. A Comparative Analysis of Selection Schemes Used in Genetic Algorithms. In *Foundations of Genetic Algorithms*; Rawlins, G.J.E., Ed.; Elsevier: Amsterdam, The Netherlands, 1991; Volume 1, pp. 69–93.
71. Kaya, Y.; Uyar, M.; Tekin, R. A Novel Crossover Operator for Genetic Algorithms: Ring Crossover. *arXiv* **2011**, arXiv:1105.0355 [cs].
72. Duell, M.; Amini, N.; Chand, S.; Grzybowska, H.; Saxena, N.; Waller, S.T. Large-Scale Dynamic Traffic Assignment: Practical Lessons from an Application in Sydney, Australia. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Gran Canaria, Spain, 15–18 September 2015; pp. 1735–1740.
73. Batista, S.F.A.; Ingole, D.; Leclercq, L.; Menéndez, M. The Role of Trip Lengths Calibration in Model-Based Perimeter Control Strategies. *IEEE Trans. Intell. Transp. Syst.* **2021**, 1–11. [CrossRef]
74. Batista, S.F.A.; Leclercq, L.; Geroliminis, N. Estimation of Regional Trip Length Distributions for the Calibration of the Aggregated Network Traffic Models. *Transp. Res. Part B Methodol.* **2019**, *122*, 192–217. [CrossRef]
75. Duthie, J.; Voruganti, A.; Kockelman, K.; Waller, S.T. Highway Improvement Project Rankings Due to Uncertain Model Inputs: Application of Traditional Transportation and Land Use Models. *J. Urban Plan. Dev.* **2010**, *136*, 294–302. [CrossRef]
76. Kockelman, K.; Xie, C.; Fagnant, D.; Thompson, T.; McDonald-Buller, E.; Waller, T. *Comprehensive Evaluation of Transportation Projects: A Toolkit for Sketch Planning*; ROSA P: Austin, TX, USA, 2010.