Università degli Studi di Napoli Federico II

Dipartimento di Ingegneria Civile, Edile e Ambientale

Dottorato di Ricerca in Ingegneria dei Sistemi Civili

XXXIII ciclo

Ph.D. Candidate

Angela Romano

Estimation/updating of origin-destination flows: recent trends and opportunities from trajectory data

Coordinatore di dottorato:

Prof. Ing. Andrea Papola

Supervisor:

Prof. Ing. Vittorio Marzano

Co-Supervisor:

Prof. Ing. Francesco Viti

UNIONE EUROPEA
Fondo Sociale Europeo

Ministero dell'Università e della Ricerca

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and background

1.1.1 The o-d flows estimation problem

Understanding the spatial and temporal dynamics of mobility demand is essential for many applications over the entire transport domain, from planning and policy assessment to operation, control, and management. Typically, mobility demand is represented by origin-destination (o-d) flows, each representing the number of trips from one traffic zone to another, for a certain trip purpose and mode of transport, in a given time interval (Cascetta, 2009, Ortuzar and Willumsen, 2011). Without loss of generality, reference will be made in the following to trips of passenger vehicles, thus the words trips and vehicles will be interchangeable. In statics, the time interval reflects a modelling horizon wherein stationary conditions hold, whilst in dynamics, the modelling horizon is discretized into time intervals to model within-day o-d flows evolution, e.g. the temporal distribution of trip departure times.

O-d flows have been generally unobservable for decades, thus the problem of o-d matrix estimation is still one of the most challenging in transportation studies. Estimation/updating of o-d flows has been studied extensively first in static traffic networks, following four main approaches: minimum information/maximum entropy (Van Zuylen and Willumsen, 1980); maximum likelihood (Bell 1983; Cascetta and Nguyen 1988), Generalized Least Squares (Cascetta 1984), Bayesian theory (Maher 1983). Drawing upon these broad general approaches, several generalizations and extensions of the o-d updating problem have been proposed: examples include incorporating the treatment of congested networks through bi-level optimization (Florian and Chen 1995; Yang 1995; Cascetta and Postorino 2001), accounting for the stochastic nature of traffic counts (Lo et al. 1996; Vardi 1996), estimating simultaneously o-d flows and the route choice model parameters (Lo and Chan 2003), or dealing with the availability of traffic counts on multiple days (Hazelton 2003). Extension to

the within-day dynamic framework was provided by Cascetta et al. (1993) through the proposition of GLS-based simultaneous and sequential o-d estimators. The former estimates jointly all o-d matrices for all time slices using the whole set of traffic counts, assuming the dynamic assignment matrix known, the latter estimates at each step the o-d flows for a given time slice θ expressed as a function of the traffic counts within θ and (some) of the already estimated previous o-d matrices. Usually, a prior estimate is obtained by a model and/or by a survey, and then an update of the prior estimate is performed to fit a set of traffic measurements, primarily link counts: this approach is characterized by two main issues.

The former is that the prior estimate is always biased. Traditionally o-d flow data is collected via high-cost, labour-intensive and time-spending surveys, such as household surveys, roadside intercept surveys, and video license capture methods, known to provide valuable information. However, these methods are dated and less viable by now due to concerns such as traffic safety and delay, privacy, and respondent burden (Tolouei et al. 2016). Furthermore, the collected sample can only represent a 'static' picture of the vehicle trip population. An alternative to the survey consists of deriving a prior estimate of the o-d matrix using analytical models. However, as demand models introduce simplifying behavioural assumptions, they inevitably yield to inherent biases.

The latter relates to the updating of the prior estimates (from both models or surveys) using indirect measurements of demand flows including a set of traffic measurements (link flows, speeds, densities, travel times, …), such that the final estimate can represent the correct traffic regime and reach a higher level of accuracy. Nonetheless, this problem is severely under-determined: the maximal number of network sensors (corresponding to the number of available linearly independent equations) is much lower than the number of the variables to be estimated (o-d flows). For this reason, a perfect fit of traffic measurements will not necessarily imply an accurate estimation of the o-d matrix.

Thus, a noteworthy research task arises, that is looking for approaches able to ameliorate the updating of prior o-d flows, possibly yielding satisfactory results irrespective of the quality of the prior estimate. This is a timely and classical issue in transport engineering, as recalled by the literature review presented in Chapter 2 of the thesis. Considering that the key issue is the imbalance between unknowns and equations and that in statics this balance cannot be met, solutions to the problem should be looked at in dynamics. After the seminal works by Hazelton et al. (2003) and Duong & Hazelton (2005), who dealt with day-to-day dynamics, most of the researchers focussed their attention on within-day dynamics. In principle, moving from statics to within-day dynamics does not alter the balance between unknowns and equations: given the

number of time slices considered in the modelling horizon, the number of equations and the number of unknowns increases proportionally in the same way. As demonstrated by the laboratory experiments on real-size synthetic networks carried out by Marzano et al. (2009), a satisfactory updating can be obtained when the ratio between the number of equations and the number of variables is close to one. These findings suggested the development of so-called quasi-dynamic o-d flow estimation/updating (Cascetta et al. 2013), which is, introducing hypothesis on the within-day demand evolution (i.e., between different time slices) to reduce the number of unknowns, based on a theoretical consideration: while the generation profile of each zone could be considered time-varying among the different time slices, the distribution percentages among the different destination zones could be considered linked to territorial aspects that vary more slowly across the day. Thus, according to the quasi-dynamic assumption, o-d shares are constant across a longer reference period (e.g. 60 minutes), whilst total flows leaving each origin vary for each sub-period within the reference period (e.g. 15 minutes). In this way, the equation-to-unknown ratio can be pushed toward the desired "one" value and true generation profiles. Importantly, the quasi-dynamic assumption has been tested only in monitored motorway systems (see Ashok and Ben-Akiva 2000 and Cascetta et al. 2013) and there is no evidence of its validity in urban contexts. Furthermore, the quasi-dynamic estimator, both offline (Cascetta et al. 2013) and online (Marzano et al. 2018), have been developed and tested only in uncongested conditions and assuming knowledge of the true underlying matrix. These unexplored aspects motivate the first part of the research presented in this thesis.

In recent times, unprecedented tracing and tracking capabilities have become available. The pervasive penetration of sensing devices (smartphones, black boxes, smart cards, …) adopting a variety of tracing technologies/methods (GPS, Bluetooth, …) could make in many cases o-d flows now observable. Considerable limitations remain, mainly related to privacy issues, organization and management of the tracking process, data transmission and storage, extraction of trips' characteristics other than just instantaneous positions and kinematics. The increasing availability of trajectory data sources has provided new opportunities to enhance observability of human mobility and travel patterns between origins and destinations, recently explored by researchers and practitioners, bringing innovation and new research directions on origin-destination (o-d) matrix estimation: a comprehensive literature review on trending research directions regarding this topic is reported in Chapter 2 of the thesis.

1.1.2 **Opportunities from sensing data**

Nowadays, direct measurements of o-d data can be collected leveraging a wide range of technologies that can be roughly categorized into two groups: fixed sensors technologies (e.g. loop detectors, cameras, Wi-fi and/or Bluetooth antennas) and moving sensors technologies, (e.g. Cellular Network, Global Navigation Satellite Systems). Fixed sensor technologies detect the number of vehicles/devices passing through the catchment area of a sensor installed at an opportunely defined location, whilst moving sensor technologies collected data from onboard devices, therefore recording with a pre-specified sample frequency the location of the vehicle moving along the network. For the scope of the thesis, this description mainly focuses on the second group of technologies (also known as point-to-point technologies) by means of which a preliminary estimate of the o-d matrix can be obtained. Characteristics of o-d flow observations can strongly vary according to the different capabilities and limitations of the technology used for the data collection; therefore, it is important to define which source can best address the study's scope and objectives. Concerning the o-d estimation problem, the primary technologies being tested and applied include cellular, GPS and Bluetooth (FHWA 2016), which are described in detail in the following:

- The GPS has established itself as a major positioning technology for providing locational data for Intelligent transport systems (ITS) applications. GPS satellites broadcast radio signals providing their locations, status, and time from onboard atomic clocks. The signals travel through space and are received by GPS receivers with their exact arrival times. Once a GPS receiver can detect at least four satellites (Zandbergen 2009), based on the time difference, geometric techniques can be utilized to locate the receiver's position on Earth in three dimensions (GPS.gov 2019). Even with some errors due to inaccurate timekeeping by the receiver's clock, GPS data usually has the highest accuracy and precision levels compared to other types of signals such as tower triangulation (see 1.1.3). We can distinguish between two types of GPS trajectory data according to the adopted sampling: opportunistic data and purpose-oriented data. In the former case the sample comes from various sources such as location-based applications for smartphones (as experimented by Zhao & Zang (2017), Cui et al. 2018 and Hasan and Ukkusuri (2014)) or navigation devices in fleets (e.g. trucks, taxi and on-demand service providers), while in the latter, data is generated from a well-structured process, designed with a specific purpose for transport research and applications, such as GPS-based household surveys (Cottrill et al. 2013; Tolouei et al. 2016; Bricka et al. 2012; Lee et al. 2016; Erhardt and Rizzo 2018; Vij e Shankari 2015) or travel diaries collection, in which users are actively solicited and asked to provide their movement

data. Non-purpose-oriented data should be carefully analysed to identify biases that could compromise population representativeness (Markovitc et al. 2019), for example, taxi data cannot represent total demand patterns. Therefore, it is common to use such datasets as probe data to get information on network status (supply performances), e.g. point-to-point travel times (Sanaullah et al. 2016), driving trajectories (Tang et al. 2018), traffic volume (Zhan et al. 2017).

- Cellular data derives from the interactions of mobile phones (which can be smartphones or any kind of mobile phones) with the cellular network, for this reason, cellular data is also referred as to event-driven data. An event occurs when the mobile device connects to the nearest tower of the cellular infrastructure and it can consist of a (starting or ending) call, SMS, an activity requiring data exchange or a handover, happening when the device is moving and automatically connecting to the next nearest tower to keep the connection/communication seamless. When an event occurs, a record containing the location of the active cell tower and the timestamp is generated, thus this type of data is also referred as to Call Detail Records (CDR). Normally, carriers collect data when the mobile device is off-call as well, but with a much lower sampling/recording frequency generating another type of records also known as "sighting" records. The location of the cell phone is calculated on the basis of its distance from the surrounding towers for e.g. by means of a tower triangulation technique (Toole et al. 2015). Its accuracy strongly depends upon the type of area in which the device is operating: in urban areas, a higher density of cell towers allows to obtain a better precision with respect to rural areas, in which cell tower distribution is more sparse. Investigating the location accuracy, Leontiadis et al. (2014) developed a method able to estimate stationary locations with a median error of 270 meters and in many case studies found in literature this error is reported around 300 meters (Wang et. al 2010; Wang et al. 2013; FHWA 2016). To derive o-d data, raw data is analysed to identify individual activity points, created when a device remains in the same location for a relatively long duration (e.g. five minutes). A broadly held conclusion regarding the o-d estimation problem states that mobile phone data can represent a proxy for human mobility, being able to capture travel patterns as well as to provide reliable estimates of o-d matrices for operational and planning applications. These accomplishments come with some important limitations affecting statistical results in mobility analysis**,** such as spatial resolution **(**e.g. the coarse granularity of cell locations leads to unreliable short-distance trip data collection (Janzen et al. 2018**),** poor location accuracy, the market share of the mobile operator

from which the dataset is obtained, sample bias in mobile phone users (e.g. high percentage of teenager users), calling plans limiting the number of phone activities, number of devices per person. Key studies validating mobile phone data processing methods and discussing the potential and limitations of mobile phone data can be found in Chen et al. (2014), Zilske e Nagel Yang et al. (2017), Chen et al. (2016), Chen et al. (2017), and Wang and Chen (2018). The validation of such methods relies on the possibility of effectively comparing passive data to socio-demographic/census data (Calabrese et al. 2011; Calabrese et al. 2013; Dash et al. 2014; Chen et al. 2014; Alexander et al. 2015; Bonnel et al. 2015; Çolak et al. 2015). The massive spread and the ubiquitous use of cellular devices represent the great potential of mobile phone datasets as they usually come with high penetration rates and widespread geographic (network) coverage, promising a significant level of population representativeness; nevertheless, these factors are evidently depending upon the market share owned by the mobile operators (Calabrese et al. 2011).

- Bluetooth technology is frequently embedded in mobile phones, GPS, and in-vehicle navigation systems; this technology is typically used for exchanging data over a short distance. Bluetooth-equipped devices can be detected as they approach the catchment area of an opportunely pre-located Bluetooth sensor. To track each device passing, the sensor identifies its unique alphanumeric identifier known as a Media Access Control (MAC) address. O-d data can be derived from Bluetooth technology by matching MAC addresses between locations where Bluetooth sensor equipment has been deployed. Given a study area, o-d flows data deriving from GPS and cellular technologies are based on estimated trip ends, thus they better suit for internal-external/external-internal trip estimation, while Bluetooth data does not contain trip ends and deriving o-d flows are based on the device location detected by the sensor, thus they are mainly used to estimate External-External (E-E) trips (FHWA, 2016). Generally, Bluetooth data is used for smaller-scale study cases, since data collection requires the installation of roadside sensors, unlike GPS or cellular data. Furthermore, Bluetooth o-d data is mainly used to compare and benchmark cell and GPS developed o-d data and for travel time estimation (e.g. Barceló et al. 2008).

Table 1.1 (Source: FHWA 2016) briefly illustrates the pros and cons of each technology/method, looking at few aspects. A first consideration relates to the sensing devices that implement these technologies/methods. Basically, two types of sensing devices can be

considered depending upon the data source, which is smartphone-based or onboard device-based (OBD-based). Typically, smartphones embed multiple tracking/tracing technologies/methods: a single device embeds numerous types of sensors and consequently a variety of data types with different level of accuracy can be obtained (Cellular, GPS, Wi-fi, Bluetooth data etc.), while OBD usually are single-technology .

O-D Data Element	Technology/Method		
	Cellular	GPS Data Stream	Bluetooth (E-E Only)
Data unit	Cell sighting based on event: call, text, data use/exchange, or network handover	GPS ping; time-stamped coordinate	MAC address of device
Positional accuracy	300 meters (average)	1–10 meters	About a 100 meter range
Data saturation/ penetration	Good, but varies	Relatively low	Varies by external station. Ranges from about 3–10 percent.
Sample frequency	Varies widely, in minutes	In seconds or minutes	In seconds
Continuous data stream?	No, random events	Sometimes, but typically pieces of trips captured	Yes, but only in about 100 meters range of reader
How trips and trip ends are estimated and defined	Based on activity points and clusters	Trip based on GPS data stream	MAC address matches between readers. Trip ends cannot be determined.
Anonymization	Encrypted to anonymize individual and device IDs through WISE technology	IDs scrambled and time/distance offsets applied. Actual trip ends may not be provided.	MAC address anonymized at field readers by removal of some digits of address, data aggregated prior to O-D table creation

Source: FHWA 2016

Table 1.1 - Key features of sensing data collection technologies/methods.

The wide and massive spread of smartphones and in general of all portable devices such as tablets, laptops, smartwatches etc. has enabled a rich data collection of users' activity points, movements, frequent visited places and so on. Nowadays these devices play a central role in people's life: in e.g. the 2017 global consumer trends by Deloitte reports that more than 80% of the Italian population is smartphone user and more than 90% owns a mobile phone of any kind. The situation is similar for each developed country, indicating the high level of penetration reached by this technology. Although they allow to provide a high level of information with high penetration rates, the most limiting aspect of smartphone-based data is its availability: data is often owned by private entities such as mobile carriers, smartphone providers, app developers etc. thus it can be inaccessible in most of the cases. Recently these entities have started to sell batch of data by encrypting users' personal information to protect their privacy anonymizing the specific individual or mobile device information, dramatically diminishing the informative content. Unlike smartphone-based data, on board device-based data is more available and more accessible to researchers and practitioners, but in most of the cases the data sample has a poor

penetration level. This can compromise the sample representativeness and variety, therefore scaling techniques are required to obtain a first estimate of the o-d matrix (see Chapter 5). Generally, this type of data derives from insurance providers or other private companies, opportunely encrypted to protect users' privacy. Recently dedicated companies have developed their own techniques to pre-process raw trajectory data and generate o-d data to sell it directly as a product, but the reliability of the data processing techniques and of the results cannot be accurately verified. The sample accuracy in this case is higher in terms of granularity and sampling frequency, but the level of socio-economic information provided is low, and if present, not accessible, therefore inference-based procedures are required to extract trips' characteristics from raw data.

To summarize, three critical aspects challenging trajectory data exploitation in transport studies and particularly in the o-d estimation problem emerged from the analysis of the literature:

- *Data ownership:* data is mostly owned by private entities, therefore there is a dramatically restricted availability of movements data. Consequently, a straightforward problem concerns the maximum achievable level of penetration. Available data is often offered as a product by vendors, e.g. mobile phone carriers or dedicated companies such as INRIX, OCTOTELEMATICS, TomTom, HERE etc. selling batch of data from various sources (e.g. freight carriers/shippers, car insurance providers etc.).
- *Penetration level and sample representativeness:* collected data should have a sufficient scale to apply inference techniques, thus the penetration level of the sample is a key factor for accurate demand estimates.
- *Sample bias and privacy concerns:* since data is often acquired from vendors, sample variety can be compromised. Furthermore, ethical issues on privacy protection arise, thus some characteristics of the sample are inevitably modified to remove sensitive information (e.g. precise location of a GPS trajectory starting point, SIM card ID for cellular data).

As mentioned, a crucial aspect in adopting trajectories for o-d flows estimation/updating relates to their penetration rate. Very high penetration rates, typical of mobile phone carrier or smartphone providers, can yield to effective o-d flows estimation methods; in addition, individual profiling helps associating socio-economic and trip information. Unfortunately, this type of data is very costly and not accessible to many researchers and practitioners, who usually must deal with trajectories collected with on-board-devices characterized by low penetration rates: by way of example, the dataset used in the thesis and described in Chapter 4, has an estimated overall penetration rate of about 6% of the total trips registered by census data, which

is considered already a good value. In addition, the estimation of the penetration rate itself is not straightforward: since the true underlying o-d flows are unknown, only inference based on census data (e.g. population, workforce, employees by traffic analysis zone) can be performed. Consistent with the above framework, trajectory data can be exploited for o-d flows estimation in various manners, as presented in the following and extensively explained in Section 2.2:

1. to obtain a first estimate of o-d flows;
2. to infer the assignment map;
3. to derive route choices underlying the assignment map;
4. to infer splitting rates at intersection;
5. to use the precedent information to ameliorate existing o-d flows estimation methods.

Currently, o-d flows updating methods are enriched with inputs from trajectory data, indeed opportunities from exploiting trajectory data are mainly related to novel and alternative formulations of the o-d flow updating problem itself: for instance, as recalled in the literature review in Chapter 2 , upscaling rates in place of o-d flows can be defined as variables of the optimization problem.

In general, although a few studies are available, to the author's knowledge no systematic assessment of the potential of these approaches has been developed yet. Furthermore, amongst methods proposed in the literature for o-d flows estimation/updating based on trajectory data, no assessment of the potential of the quasi-dynamic framework has been proposed, as also explicitly mentioned by Yang et al. (2017).

Overall, the above motivates the formulation of the research questions underlying this work, stated in Section 1.2. The outline of the thesis is illustrated in Section 1.3.

1.2 Thesis contribution

Consistently with the above considerations, this thesis focuses on a threefold objective:

- To conduct a preliminary study investigating the unexplored properties of quasi-dynamic assumption and to provide insights on how to implement the quasi-dynamic o-d estimation framework when dealing with congested networks (Chapter 3);
- To develop an extended analysis mining the entire range of variability and variety of trajectory data samples aiming at evaluating how the issues of representativeness and sample bias affect o-d estimation performances. This contribution has been carried out by means of both empirical (Chapter 4) and laboratory analysis (Chapter 5, Chapter 6). Part of the experimental analysis in Chapter 4 is dedicated to the assessment of quasi-

dynamic evolution of the demand in urban context, supporting the application of quasi-dynamic framework to o-d updating methods in presence of trajectory data.

- To investigate the potential of existing and new formulations of GLS-based estimators (simultaneous and quasi-dynamic GLS) in presence of trajectory data, defining its role on the basis of the considerations and results derived from the analysis described above (see Chapter 7).

Specifically, this thesis aims to provide a systematic study developed by means of laboratory experiments investigating on the potential of the various types of trajectory data samples for o-d related analysis, in terms of their characteristics (e.g. coverage, penetration level and distribution…). Given a set of opportunely defined ground truth data, synthetic experiments are used as powerful tools to provide an appropriate validation of proposed o-d estimation methods (e.g. Antoniou et al. 2016). Therefore the analysis will cover a wide experimental plan to investigate on the implications of sample characteristics variation and their impact on travel demand accuracy i.e. from the analysis of real trajectory data, evidence shows that penetration rate strongly vary among o-d pairs (see Section 4.2), thus a study assessing its variability and the consequent implications on o-d estimation becomes essential. Other hypothesis on the level and the distribution of penetration rate are considered: homogeneous by origin, by destination, by o-d pair or vice versa. In addition, the study sheds the light on the combined capability of different sets of data investigating on the simultaneous use of both trajectory data and link traffic data. Considerations on the implications in the dynamic and the static context for uncongested networks are developed by means of synthetic experiments, which can be one of the best tools guaranteeing a proper validation for o-d related analysis. Several o-d estimation methods are tested on variable scale networks to verify the sensitivity of the solution with respect to the network dimension. Furthermore, as also suggested by Yang et al. 2017, the quasi-dynamic hypothesis can be the next step to solve trajectory expansion rate estimation dramatically reducing the problem dimension. To this end, a preliminary study on quasi-dynamic methods is developed, calibrating its parameters for both congested and uncongested cases (see Chapter 3 and Section 4.5). Note that the scope of this thesis is car traffic, and, in terms of the modelling literature, the discussion involves both dynamic and static (stationary) applications.

This work, being part of the national research program namely “Programma Operativo Nazionale Ricerca e Innovazione 2014-2020”, has been developed in collaboration with the “Université du Luxembourg” under the supervision of Prof. Ing. Francesco Viti and in partnership with the company “PTV – Sistema” under the supervision of Ing. Lorenzo

Meschini. The real trajectory data sample analysed in this thesis has been obtained under the "STAR: Sostegno Territoriale Attività di Ricerca" program. Further details of the trajectory dataset and its characteristics are reported in Chapter 4.

1.3 Outline of the thesis

Consistent with the statement of research, the structure of the thesis is the following:

- Chapter 2: Literature review
- Chapter 3: Quasi-dynamic assumption in congested network
- Chapter 4: Trajectory data in Napoli
- Chapter 5: Performance analysis of direct scaling
- Chapter 6: Laboratory experiments to assess the reliability of traffic assignment map
- Chapter 7: o-d flows updating methods in presence of trajectory data
- Chapter 8: Conclusions and future research questions

2 Literature review

This chapter reports the literature review related to the two main topics of the thesis, that is o-d flows estimation/updating (Section 2.1) and the o-d estimation/updating methods in presence of trajectory data (Section 2.2). Finally, Section 2.3 summarises the outcomes of the literature review and the research contribution developed in the thesis.

2.1 Estimation/updating of o-d flows

Estimation/updating of the o-d matrix is a traditional problem in transport engineering. The procedure consists of updating a prior estimate of the o-d matrix fitting the information derived from a set of traffic measurements, primarily link counts. Traditionally a prior estimate of the o-d matrix derives from high-cost, labour-intensive and time-spending surveys, such as household surveys, roadside intercept surveys and video license capture methods, known to provide valuable information. However, the collected sample represents a 'static' picture of the vehicle trip population. An alternative to survey consists of deriving a prior estimate of the o-d matrix by means of analytical models. However, as demand models introduce simplifying behavioural assumptions, they inevitably yield to inherent biases. Thus, the updating of the prior estimates (from both models or surveys) using indirect measurements of demand flows including a set of traffic measurements (link flows, speeds, densities, travel times, …), is performed such that the final estimate can represent the correct traffic regime and reach a higher level of accuracy. This problem is severely under-determined: the maximal number of network sensors (corresponding to the number of available linearly independent equations) is much lower than the number of the variables to be estimated (o-d flows). For this reason, a perfect fit of traffic measurements will not necessarily imply an accurate estimation of the o-d matrix.
The problem formulation depends upon the assumptions on the temporal o-d flows evolution (stationary or dynamic conditions), and concerning the dynamic context, a further distinction can be made between on-line and off-line applications according to the temporal characteristics of the input data and the data feeding process. Furthermore, each case can be analysed for congested or uncongested networks.

2.1.1 Static O-D matrix estimation/updating

The static framework allows to perform the estimation/updating of the o-d matrix from traffic counts assuming time-independent conditions within the analysis horizon, such that the time interval reflects a modelling horizon wherein stationary conditions hold. Two main theoretical approaches have been proposed in the literature: the former approach comprises the "classical" estimation methods such as the Maximum Likelihood (ML) estimator proposed by Maher (1983) and Bell (1983) and the Generalised Least Squares (GLS) estimator proposed by Cascetta (1984), whilst the latter refers to the Bayesian framework proposed by Maher (1983). Following Cascetta and Nyguen (1988) and Cascetta (2001), the Maximum Likelihood (ML) estimator provides a final demand estimate ($\mathbf{d}_{ML}$) maximising the probability of conjunctively observing the o-d sampling survey data and the link counts data, under the assumption that these two probabilities are independent:

$$\mathbf{d}_{ML} = \arg\max_{\mathbf{x}\geq 0} \left[lnL(\hat{\mathbf{d}}|\mathbf{x}) + lnL(\hat{\mathbf{f}}|\mathbf{x}) \right] \tag{2.1}$$

wherein:

- $\mathbf{x}$ is the vector of the optimization variables, one for each o-d pair;
- $\hat{\boldsymbol{d}}$ is the demand vector derived by survey data;
- $\hat{\boldsymbol{f}}$ is the vector of link counts, one for each measured link.

Log-likelihood functions in equation 2.1 are specified based on assuming specific hypotheses on the probability distribution of demand counts $\hat{\boldsymbol{d}}$ and traffic counts $\hat{\boldsymbol{f}}$ respectively, conditional on the demand vector $\mathbf{x}$. The probability distribution of the prior demand estimate depends upon the sampling strategy adopted in the survey, while traffic counts are typically assumed as independently distributed as Poisson random variables or as Multivariate Normal random variables.

The Generalized Lead Squared (GLS) estimation method consists of an optimization problem aiming at providing a final estimate solving a system of linear stochastic equation. The goal is to minimise the distance (error) between collected data (link counts and a priori demand) and estimated values, such that the final estimate ($\mathbf{d}_{GLS}$) best fits the collected traffic measurements and the o-d survey data. This formulation yields to the following optimization problem:

$$\mathbf{d}_{GLS} = \arg\min_{\mathbf{x}\geq 0} \left\{ \frac{1}{2} \left(\mathbf{x} - \hat{\mathbf{d}} \right)^T \mathbf{Z}^{-1} \left(\mathbf{x} - \hat{\mathbf{d}} \right) + \frac{1}{2} (\hat{\mathbf{f}} - \mathbf{M}_\mathbf{f} \mathbf{x})^T \mathbf{W}^{-1} (\hat{\mathbf{f}} - \mathbf{M}_\mathbf{f} \mathbf{x}) \right\} \tag{2.2}$$

wherein:

- $\mathbf{M_f}$ is the sub-assignment matrix related to the set of available traffic counts;

- **Z** and **W** are respectively the covariance matrices related to the sampling error underlying the demand estimation and the measurement/assignment errors.

The Bayesian approach combines experimental information derived from measurement traffic data (link traffic counts) and non-experimental information based on the a priori knowledge or expectation about the demand probability function (e.g. coming from previous estimates or from an analytical model). The final estimate provided by the Bayesian-based estimators maximises the logarithm of the *a posteriori* probability, that is the probability function attributed to the unknown vector given the *a priori* estimate **d*** ($g(\mathbf{x}|\mathbf{d}^*)$) and the probability of observing the vector of traffic counts conditional on the unknown demand vector **x** ($L(\hat{\boldsymbol{f}}|\mathrm{x})$):

$$\mathbf{d}_B = \arg\max_{\mathbf{x}\geq 0}\left[ln\, g(\mathbf{x}|\mathbf{d}^*) + lnL(\hat{\mathbf{f}}|\mathbf{x}) \right] \tag{2.3}$$

The specific formulation of a Bayesian estimator and its performance strongly depends upon the assumptions on the probability functions $g(\mathbf{x}|\mathbf{d}^*)$ and $L(\hat{\boldsymbol{f}}|\mathbf{x})$. The unknown demand vector can be assumed to follow a multinomial random variable (in this case $ln\, g(\mathbf{x}|\mathbf{d}^*)$ becomes the entropy function of the unknown vector **x**), a Poisson random variable (in this case $ln\, g(\mathbf{x}|\mathbf{d}^*)$ becomes the information function of the unknown vector **x**), or a Multivariate Normal random variable. Interestingly, if all the relevant distributions are assumed as multivariate normal, the methods discussed above would result with identical objective functions.

Drawing upon these approaches, a number of generalizations and extensions have been proposed. Bell (1991) explored further theoretical properties of the GLS method. Other examples include applications dealing with congested network by incorporating o-d estimation and traffic assignment feedbacks through a bi-level formulation of the optimization problem (Yang et al. 1991; Florian and Chen 1995; Yang 1995; Cascetta and Postorino 2001). Lo et al. (1996) and Vardi (1996) among others, investigated on the stochastic nature of traffic counts. A further generalization by Lo and Chan (2003) jointly estimates o-d flows and route choice model parameters dispersion in the case of congested networks, whilst Hazelton (2003) extended the problem to applications dealing with the availability of traffic counts on multiple days considering time-series link counts.

2.1.2 **Dynamic o-d matrix estimation/updating**

The methods described for the static framework have been generalised to model within-day o-d flows evolution performing a dynamic estimation/updating using time-varying traffic measurements. Specifically, the dynamic framework prompted further research directions extending and adapting the static formulations to off-line and on-line applications, which will be separately described in the following Sections (2.1.2.1 and 2.1.2.2). Basically, offline

applications generally suit long-term purpose (e.g. planning) because measurement data is first collected in batch and then used to estimate the o-d flows, while online applications suit real-time (or short-term) management applications since the traffic data feeding process is continuous such that data amount increases at each time-interval.

2.1.2.1 *Off-line*

The dynamic o-d estimation problem was first formulated by Cascetta et al. (1993) proposing the simultaneous and the sequential GLS-based estimators. The simultaneous estimator jointly estimates the time-dependent o-d matrices for the total number of time slices using the whole set of traffic counts, thus it is designed for offline applications, whilst the sequential estimator is suitable for online applications, thus it will be discussed in details in the next section. The simultaneous estimator formulation can be smoothly derived from the estimator 2.2:

$$\{\mathbf{d}^{*1},\dots,\mathbf{d}^{*\theta},\dots,\mathbf{d}^{*n_\theta}\} = \underset{\mathbf{x}^\theta \geq 0 \;\forall \theta \in T}{\arg\min} \left\{ \sum_{\theta=1}^{n_\theta} \sum_{od=1}^{n_{od}} \frac{\left(x_{od}^\theta - \hat{d}_{od}^\theta\right)^2}{\sigma_{od}^\theta} + \sum_{\theta=1}^{n_\theta} \sum_{l=1}^{n_l} \frac{\left(\sum_{\theta'=\theta_l}^{\theta} \sum_{od=1}^{n_{od}} m_{odl}^{\theta'\theta} x_{od}^{\theta'} - \hat{f}_l^\theta \right)^2}{\sigma_l^\theta} \right\} \tag{2.4}$$

Wherein:

- $\mathbf{x}^\theta = \{x_1^\theta,\dots,x_{n_{od}}^\theta\} \;\forall \theta \in T$ represents the unknown demand vectors;
- $\mathbf{d}^{*\boldsymbol{\theta}} = \{d_1^{*\theta},\dots,d_{n_{od}}^{*\theta}\} \forall \theta \in T$ *is* the corresponding optimal solutions
- $\hat{\boldsymbol{d}}^\theta$ the $(n_o \cdot n_d)$ matrix of the prior demand estimates $\hat{d}_{od}^\theta$ for the time slice θ;
- $\hat{\boldsymbol{f}}^\theta$ the $(n_{lc} \cdot 1)$ vector of the observed link counts $\hat{f}_l^\theta$ for the time slice θ.
- $m_{odl}^{\theta'\theta}$ is the generic term of the dynamic assignment map linking time-dependent o-d flows with time-dependent link flows (i.e. it represents the fraction of o-d flow generated at the time slice θ' being on link l at the time slice θ);
- σ^θ_{od} and σ^θ_l are related to the dispersion matrix of the demand and of the counted flows distribution respectively;
- θ_l is the farthest time slice whose generated demand contributes to the link flows on θ.

Notably, the approach requires a dynamic traffic assignment (DTA) model to derive the dynamic assignment map and even though it is a robust estimator for offline applications, computational issues arise for moderate size networks, as reported in Cascetta and Russo (1997), Toledo et al. (2003) and Bierlaire and Crittin (2004).

Further adaptations have been proposed to deal with congested networks: since in the simultaneous approach the dynamic assignment matrix is exogenously determined, it might be

inconsistent with the estimated assignment mapping (Toledo et al. 2015). To avoid such inconsistency and take into account the complex interaction between o-d flows, path flows and link flows, the problem was formulated as a bi-level optimization problem (Bracken & McGill 1973) by Tavana and Mahmassani (2000) and Tavana (2001), proposing an iterative solution framework to estimate dynamic o-d demand matrix. In the upper-level problem, o-d flows are updated to fit the traffic measurements, while in the lower-level a DTA model maps time-dependent o-d flows with time-dependent link flows such that the estimated link flows are consistent with the demand values calculated at the generic iteration of the first level (Balakrishna et al. 2007). However, this circular dependency between od flows and traffic variables increases problem complexity, since the highly non-linear relationship between these two entities makes the problem highly non-convex. Frederix et al. (2011) analysed the influence of non-linearity of link-route proportion matrix on o-d estimation to obtain a more accurate representation of congestion phenomena, calculating the sensitivity (Jacobian matrix) of the link flows to the o-d flows through finite differences with marginal computation simulations based on first-order kinematic theory. Alternatively, Toledo and Kolechkina (2013) developed iterative algorithms to solve the estimation problem underlying a linear assignment matrix approximation. Several authors identified alternative ways to overcome the non-linearity issue: i.e. by integrating additional information within the o-d estimation framework (Dixon & Rilett 2002; Antoniou et al. 2006; Balakrishna 2006; Zhou & Mahmassani 2007; Barcelò, et al. 2010; Zhang et al. 2011; Rao et al. 2018; Nigro 2017) or developing assignment matrix-free algorithm i.e. by reproducing directly the relationship between measurement profiles and o-d flow profiles without using any assignment matrix explaining the demand pattern (e.g. Cremer and Keller 1981; Cremer and Keller 1984; Cremer and Keller 1987; Nihan and Davis 1987; Nihan and Davis 1989; Balakrishna et al. 2008; Carrese et al. 2017) or using machine learning to learn this nonlinear relationships (Wu et al. 2018; Krishnakumari et al. 2019). Other relevant research directions within this field dealt with joint estimation of demand and supply parameters: starting from Liu and Fricker (1996) to the development of efficient estimation algorithms in the recent contributions by Antoniou et al. (2015) and Cipriani et al. (2011). Some of the models proposed for within-day dynamics have been extended to day-to-day applications aiming at capturing the process of traffic evolution over multiple days as proposed by Hazelton (2003). A review of the early contributions in this field is reported by Balakrishna et al. (2005).

2.1.2.2 On-line

The primary requirement to develop on-line applications is to recursively provide estimations of the system state at the current step and future system state in the short term. The estimation

should be as fast as possible to dispose accurate real-time prediction status and support transport systems management and control. As mentioned in the previous section, the sequential estimator proposed by Cascetta et al. (1993) is generally designed for on-line or large-scale applications. Indeed, the sequential approach separates a large optimization problem into n_θ smaller problems, one for each time-slice considered in the time horizon, dramatically diminishing computational burden. Each problem estimates only the demand vector referred to the current time-slice using current traffic counts and taking in input as target demand, the demand vector estimated referred to previous time-slices. Additionally, traffic counts of the processed time-slice are expressed as a linear function of the demand vector referred to both the current time-slice and to the previous intervals. The corresponding GLS formulation is reported in Cascetta (2001). To overcome the limitation of using only current traffic measurement in the estimation process Ashok and Ben-Akiva (1993), following the seminal work of Okutani and Stephanades (1984), modelled the within-day demand evolution across time-slices by means of an autoregressive process and proposed a forward Kalman filter method to predict o-d flows for the time slice $\theta+1$ based on link flow measurements at the time slice θ. Indeed, to ameliorate their approach to take account for the previous time-slices measurements, Ashok and Ben-Akiva (1993) introduced an augmented state-space model using as state-space variables the deviation of the demand vector from its historical values. Subsequently, Ashok and Ben-Akiva (2000) suggested an alternative approach, having redefined the state variables as the deviations of departure flows from each origin and the shares headed to each destination. Except for different forms of transition equations, the approach has a similar framework as those they proposed earlier. Notably, the Kalman filter can be used also for off-line applications, as proposed by Balakrishna et al. (2005) and particularly by Gelb (1974), who suggested a double-step off-line estimation approach based firstly on a forward Kalman filter application and then on a backward Kalman smoothing, to account for the knowledge of link counts for all time slices in off-line contexts, providing more robust and reliable results with respect to a simple forward Kalman filter (Ashok 1996). Another approach to the on-line application introduced by Zhou and Mahmassani (2007) assumes a polynomial approximation for structural deviations of the demand with the respect to the historical estimate and for some of its derivatives. Chang and Wu (1995) and Ashok and Ben-Akiva (2002) dealt with the randomness of the dynamic assignment matrix introducing supply parameters such as travel time and path choice fraction within the state space formulation. Barceló et al. (2010) proposed a linear Kalman filtering, later extended by (Barceló et al. 2013) to cope with congestion effects

introducing a new formulation for general network structures integrating travel times measurements provided by Bluetooth technology.

Given the high number of variables to estimate in the short term, minimising computational time and load becomes crucial for an effective and efficient prediction. The seminal work by Bierlaire and Crittin (2004) addressed computational issues proposing an efficient algorithm to deal with large scale networks. Later on, Antoniou et al. (2009) proposed a joint calibration of the off-line and on-line DTA systems through a common framework such that both historical as well as real-time information are efficiently utilised to better reflect prevailing conditions. Cipriani et al. (2014) proposed a quasi-dynamic traffic assignment model which approximates the dynamic traffic model by steady-state intervals and applies approximate performance functions, reducing the computational burden. Another efficient method to reduce computational costs was proposed by Djiukic et al. (2012) by applying principal component analysis (PCA). The method allows for a significant reduction of the o-d estimation problem dimensionality, selecting the o-d pairs which preserve structural patterns and guarantee a negligible loss of accuracy.

The introduction of gradient-free algorithms into o-d estimation/prediction and DTA frameworks such as simultaneous perturbation stochastic approximation (SPSA) has played an important role in addressing computational issues. SPSA is an iterative derivative-free optimization algorithm proposed by Spall (1992, 1998a, 1998b) and designed for stochastic problems which significantly saves computational time for large-scale problems with the respect to traditional gradient methods (e.g. finite-differences stochastic approximation, FDSA). Indeed, differently from gradient-based algorithm which calculates the gradient of the objective function to individuate the direction of maximum function variation, SPSA approximate the gradient of the objective function with two successive measurements of the objective function, independently of the number of parameters to estimate. However, the gain in computational time comes with consistent loss of accuracy and information: approximating the gradient for each parameter using the aggregate error in the whole network across the entire simulation period introduces noise from uncorrelated measurements proportional to the size of network and the number of intervals. To overcome these limitations and to ameliorate its performance, extended and improved versions of the SPSA have been successfully proposed, i.e. Balakrishna and Koutsopoulos (2008) integrate o-d flows transition equations into the objective function, or Cipriani et al. (2011) proposing the asymmetric estimation and introducing a polynomial interpolation to compute the step size: with the SPSA-asymmetric design formulation, the number of necessary assignments to compute the gradient is reduced to

50% with respect to the basic SPSA with symmetric design (SD). To mitigate the noise from uncorrelated measurements and to enhance convergence and robustness Lu et al. (2015) proposed a Weighted SPSA (W-SPSA), embedding information of spatial and temporal correlation in a traffic network. The enhanced method outperformed SPSA when applied on a hundreds of thousands of o-d pair network (the entire Singapore expressway), demonstrating its efficacy on a very large scale network problem. Cantelmo et al. (2014a) proposed a second order SPSA introducing a scaling factor involving the inverse of the Hessian matrix estimate to mitigate the effect of different magnitude order variables in the objective function. If the assignment matrix is available a further SPSA specification, namely adaptive SPSA (Spall, 2000), allows to estimate the Hessian Matrix from o-d path proportion and speed up the algorithm convergence.

2.1.3 **Quasi-Dynamic o-d matrix estimation/updating**

A serious shortcoming still affecting the o-d demand estimation methods using traffic counts is the large imbalance between equations and unknowns: as it stands, the problem is severely under-determined meaning that the number of unknown is much larger than the number of available linearly-independent equations derived from traffic count measurements such that many combinations of demand patterns correspond to the same set of measurements. Additionally, the set of possible solutions increases with the size of the network and the alternatives available on the network. While in statics this condition cannot be ameliorate, in within-day dynamics an effective method to reduce problem dimensionality is making some assumptions on the within-day evolution of demand flows. As demonstrated by the laboratory experiments on real-size synthetic networks carried out by Marzano et al. (2009), a satisfactory updating can be obtained only when the ratio between the number of equations (i.e. independent observed link flows) and the number of unknowns (i.e. o-d flows) is close to one. In principle, moving from statics to within-day dynamics does not alter the balance between unknowns and equation: given the number of time slices considered in the modelling horizon, the number of equations and the number of unknowns increase proportionally in the same way. However, under reasonable hypotheses on o-d flow variation across time slices, the number of unknowns in within-day dynamic systems can be bound, thus achieving unknowns/equations ratios close to one. Along this research direction, the paper proposes a “quasi-dynamic” framework for estimation of o-d flows, hinted by Marzano et al. (2009), in which o-d shares are assumed constant across a reference period, whilst total flows leaving each origin are assumed varying for each sub-period within the reference period.

2.1.3.1 The quasi-dynamic assumption

Given a time horizon T of duration t_T divided into $n_\mathrm{q} = t_T/t_\mathrm{q}$ time slices θ of duration t_q, let d_{od}^{θ} be the generic o-d flow to be estimated for the time slice θ. d_{od}^{θ} can be expressed as the product between the generated demand g_o^{θ} by origin o during the time slice θ and the distribution probability $p_{d|o}^{\theta}$ of choosing destination d moving from o within the time slice θ. The quasi-dynamic (QD) assumption states that: while factors affecting generation profiles (g_o^{θ}) dynamic evolution are inherently within-day time varying (i.e. number of trips starting from o in θ), distribution shares $p_{d|o}^{\theta}$ dynamics can result more within-day stationary. This means that the distribution percentages among the different destination zones can be considered linked to territorial aspects that vary more slowly across a day. Therefore, distribution probabilities $p_{d|o}^{\theta}$ can be reasonably approximated to their average values across an opportunely pre-specified sub-period $\tau \subseteq T$ of duration $t_\tau \leq t_T$, encompassing a number of subsequent time slices given by $n_{\theta|\tau} = t_\tau/t_\theta$. In formulae, indicating the probability distribution average values over τ as $p_{d|o}^{\tau(\theta)}$, the "quasi-dynamic" o-d flow $d_{od}^{\theta,qd}$ can be obtained as follows:

$$d_{od}^{\theta} = g_o^{\theta} p_{d|o}^{\theta} \cong g_o^{\theta} p_{d|o}^{\tau(\theta)} = d_{od}^{\theta,qd} \tag{2.5}$$

wherein:

$$g_o^{\theta} = \sum_d d_{od}^{\theta} \tag{2.6}$$

$$p_{d|o}^{\tau(\theta)} = \frac{d_{od}^{\tau}}{g_o^{\tau}} = \frac{\sum_{\theta\in\tau} d_{od}^{\theta}}{\sum_{\theta\in\tau} g_o^{\theta}} = \frac{\sum_{\theta\in\tau} d_{od}^{\theta}}{\sum_{\theta\in\tau}\sum_d d_{od}^{\theta}} \;\forall\, \theta \in \tau \subseteq T \tag{2.7}$$

and $\tau(\mathrm{q})$ maps the time slices θ to sub-periods τ, i.e. $\tau(\mathrm{q})$ represents the specific subperiod τ which the time slice θ belongs to. The QD assumption allows reducing the number of unknowns from $n_\mathrm{q} \cdot n_{od}$ to $n_\theta \cdot n_o + n_\tau \cdot (n_{od} - n_o)$. Inevitably, the QD assumption introduces a bias which will be termed as "intrinsic error" ie_{od}^{θ}, given by the distance between prior od flow values d_{od}^{θ} and quasi-dynamic od flows $d_{od}^{\theta,qd}$:

$$ie_{od}^{\theta} = d_{od}^{\theta} - d_{od}^{\theta,qd} \tag{2.8}$$

This error comes from the fact that the ground-truth is not precisely quasi-dynamic, and increases the mathematical complexity of the quasi-dynamic estimator with respect to a standard GLS estimator leading to a bilinear form given by the product of the generation and distribution variables. To assess the magnitude of the intrinsic error Cascetta et al. (2013) developed a statistical analysis of the observed o-d flows to support the assumption of quasi-

dynamic o-d flows pattern in uncongested networks, using chi-squared and likelihood ratio tests, with acceptable goodness-of-fit measures even under the hypothesis of constant distribution shares for the whole day.

2.1.3.2 The offline GLS-based quasi-dynamic o-d flows estimator

Cascetta et al. (2013) proposed the theoretical formulation of the quasi-dynamic o-d flow updating framework, i.e. estimators based on the assumption of constant distribution shares across larger time horizons with respect to the within-day variation of the generation profiles, leading to an estimator which improves dramatically the unknowns/equations ratio. The quasi-dynamic GLS estimator, which in the following will be termed as QD-GLS, can be interpreted as a particularization of the simultaneous GLS estimator (Cascetta et al. 1993). Indeed, its formulation can be obtained adopting the new set of variables resulting from the quasi-dynamic assumption, yielding to an optimization problem formally expressed, under the assumption of diagonal dispersion matrices, as:

$$\{\boldsymbol{g}^{*^1}, \dots, \boldsymbol{g}^{*^\theta}, \dots, \boldsymbol{g}^{*^{n_\theta}}; \boldsymbol{p}^{*^1}, \dots, \boldsymbol{p}^{*^\tau}, \dots, \boldsymbol{p}^{*^{n_\tau}}\} = \arg \min_{\substack{g^1 \dots g^{n_\theta} \in S_g \\ p^1 \dots p^{n_\tau} \in S_p}} \left\{ \sum_{\theta=1}^{n_\theta} \sum_{od=1}^{n_{od}} \frac{(g_o^\theta \cdot p_{d|o}^{\tau(\theta)} - \hat{d}_{od}^\theta)^2}{\sigma_{od}^\theta} + \right.$$

$$\left. \sum_{\theta=1}^{n_\theta} \sum_{l=1}^{n_{lc}} \frac{(\sum_{\theta'=\theta_l}^{\theta} \sum_{od=1}^{n_{od}} m_{odl}^{\theta'\theta} g_o^{\theta'} \cdot p_{d|o}^{\tau(\theta')} - \hat{f}_l^\theta)^2}{\sigma_l^\theta} \right\} \tag{2.9}$$

s.t.

$$g^1 \dots g^{n_\theta} \in S_g: g_o^\theta \geq 0 \; \forall o, \forall \theta \in T$$

$$p^1 \dots p^{n_\tau} \in S_p: 0 \leq p_{d|o}^\tau \leq 1 \; \forall p_{d|o}^\tau \in \boldsymbol{p}_{d|o}^\tau \; \forall \tau \in T \; ; \sum_d p_{d|o}^\tau = 1 \; \forall o, \forall \tau \in T$$

Wherein:

- $\boldsymbol{g}^\theta$ is the $(n_o \cdot 1)$ vector of the generated demands g_o^θ for a given time slice θ;
- $\boldsymbol{p}^\tau$ is the $(n_o \cdot n_d)$ matrix of the distribution probabilities $p_{d|o}^\tau$ for a given sub-period τ;
- $\hat{\boldsymbol{d}}^\theta$ the $(n_o \cdot n_d)$ matrix of the prior demand estimates $\hat{d}_{od}^\theta$ for the time slice θ;
- $\hat{\boldsymbol{f}}^\theta$ the $(n_{lc} \cdot 1)$ vector of the observed link counts $\hat{f}_l^\theta$ for the time slice θ.

The unknowns are the demand generation profiles $\boldsymbol{g}^\theta$ for each time slice θ and the matrices of the distribution shares $\boldsymbol{p}^\tau$ for each sub-period τ respecting the feasibility sets described above. $m_{odl}^{\theta'\theta}$ is the generic term of the dynamic assignment map linking time-dependent o–d flows with time-dependent link flows (i.e. it represents the fraction of o–d flow generated at the time

slice θ' being on link l at the time slice θ), σ_{od}^{θ} and σ_{l}^{θ} are related to the dispersion matrix of the demand and of the counted flows distribution respectively, while θ_l is the farthest time slice whose generated demand contributes to the link flows on θ. The estimator 2.9 is a bilinear form with respect to the unknowns $\boldsymbol{g}^{\theta}$ and $\boldsymbol{p}^{\tau}$: the computational load required makes this formulation suitable only for off-line applications.

Cascetta et al. (2013) conducted the empirical analysis of the QD-GLS estimator on a real dataset for the case of uncongested network and fixed route choice, indeed the chosen test site consisted of a (closed) motorway system in which congestion phenomena can be considered negligible. Moreover, the linear mapping between link counts and o-d flows in this case can be expressed by means of a fixed and exogenous assignment matrix inferred from entry/exit times. Authors compared its performances with the simultaneous estimator proposed by Cascetta et al. (1993) (see Section 2.1) and the Kalman filter approach proposed by Ashok (1996): a recursive estimator typically used for on-line dynamic estimation. Interestingly, Ashok's Kalman Filter formulation introduced an alternative approach based on a property termed the "stability of shares", (a property according to which o-d shares remain stable over the course of a day relative to the departing trips), which can be considered as the seminal input for the formulation of quasi-dynamic hypothesis, but apart from redefining the state variables as the deviations of departure flows from each origin and the shares headed to each destination, the proposed method does not allow for a reduction of the problem dimensionality, thus the number of variables is not altered. The experimental analysis demonstrated that the QD-GLS estimator outperforms the simultaneous estimator in reproducing dynamic o-d flows estimates and the quality of the Kalman filter estimates is quite close to the quality of its seed o-d flows: as a consequence, the QD-GLS estimator is also very useful in supporting on-line applications, since using quasi-dynamic estimates as historical seeds allows the Kalman filter to provide good o-d flow estimates. Furthermore, aggregating QD-GLS estimates for successive time slices represents also the most effective way to reproduce o-d flows estimates for larger time horizons (e.g. hourly estimates) for static applications, outperforming in such way estimations coming by both using the classical static estimator proposed by Cascetta et al. (1984) and by aggregating the simultaneous dynamic o-d estimates of the corresponding time slices. Adopting the quasi-dynamic optimization variables framework (i.e. expressing o-d flows as the product of generations and distributions), Cantelmo et al. (2015) proposed a Two-Step approach separating the dynamic demand estimation problem into two sequential optimizations. Specifically, the first step uses a strict quasi-dynamic assumption to update the total generated demand volume for each traffic zone assuming fixed distribution shares over the entire time-horizon, while the

second step adjusts o-d flow values using a simultaneous GLS (Section 2.1.2.1). Subsequently, the two-step approach was applied on a large-scale network by Cantelmo et al. (2017) embedding Call Detail Records (CDR) data into the first step objective function to gain accuracy on generation demand volumes and extended to large-scale congested networks by Cantelmo et al. (2020) conducting a sensitivity analysis evaluating the introduction of different weights for speed and link counts measurements into the objective function. Further details regarding this approach can be found in Section 3.2, as it has been used as benchmark method to investigate the unexplored properties of the quasi-dynamic framework in the context of congested networks. Another recent contribution adopting the quasi-dynamic approach in the o-d flows estimation problem was provided by Bauer et al. (2017), proposing a method to eliminate the need for supplying the historical o-d matrix.

2.1.3.3 A Kalman filter for on-line quasi-dynamic o-d flow estimation/updating

To extend the quasi-dynamic framework to on-line applications Marzano et al. (2018) proposed an extended quasi-dynamic Kalman filter (QD-EKF), implemented for the case of uncongested networks. Its mathematical formulation is based on an augmented state-space variables composed of generation and distribution deviations from historical data following an autoregressive process.

In this respect, since Ashok and Ben-Akhiva (1993), Ashok (1996) and Ashok and Ben-Akhiva (2000), proposed already a Kalman filter based on an assumption similar to the quasi-dynamic termed “stability of shares” assuming the temporal evolution of the distribution shares to be less stochastic than the generation profiles, their mathematics is the natural starting point for the specification of the QD-EKF.

Denoting with θ the Kalman Filter time step, the state variables introduced were the deviations $g^{o}{}_{q}$ -$g^{oH}{}_{q}$ of the generation profiles and $\pi^{d|o}{}_{q}$ - $\pi^{d|oH}{}_{q}$ of the distribution shares from their respective historical estimates, denoted with superscript H. The state-space vector of the filter for the generation profiles is:

$$\Delta\boldsymbol{\gamma}_{\boldsymbol{\theta}} = \boldsymbol{\gamma}_{\boldsymbol{\theta}} - \boldsymbol{\gamma}_{\boldsymbol{\theta}}^{H} = \begin{bmatrix} \Delta\gamma_{\theta}^{o_1} \\ \dots \\ \Delta\gamma_{\theta}^{o_{n_o}} \end{bmatrix} = \begin{bmatrix} \gamma_{\theta}^{o_1} - \gamma_{\theta}^{o_1\,H} \\ \dots \\ \gamma_{\theta}^{o_{n_o}} - \gamma_{\theta}^{o_{n_o}\,H} \end{bmatrix} \tag{2.10}$$

with dimension $n_o \cdot 1$, being n_o the number of origins, whereas for the distribution shares is:

$$\Delta\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathbf{o}} = \boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathbf{o}} - \boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathbf{o}^H} = \begin{bmatrix} \pi_\theta^{d_1|o} - \pi_\theta^{d_1|o^H} \\ \dots \\ \pi_\theta^{d_{n_d|o}} - \pi_\theta^{d_{n_d|o}{}^H} \end{bmatrix} \tag{2.11}$$

with dimension $n_{d|o} \cdot 1$ of the distribution shares from origin o, in turn leading to the vector:

$$\Delta\boldsymbol{\pi}_{\boldsymbol{\theta}} = \boldsymbol{\pi}_{\boldsymbol{\theta}} - \boldsymbol{\pi}_{\boldsymbol{\theta}}^{H} = \begin{bmatrix} \Delta\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathbf{o_1}} \\ \dots \\ \Delta\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathbf{o_{n_o}}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathbf{o_1}} - \boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathbf{o_1}{}^H} \\ \dots \\ \boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathbf{o_{n_o}}} - \boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathbf{o_{n_o}}{}^H} \end{bmatrix} \tag{2.12}$$

of dimension $n_{od} \cdot 1$, i.e. containing all distribution shares for the time slice θ.

The corresponding transition equation for the generation profiles is expressed as an autoregressive process, given by:

$$\gamma_{\theta+1}^{o} - \gamma_{\theta+1}^{o}{}^{H} = \sum_{t=\theta-\theta_g}^{\theta} a_{\theta,t}^{o}\left(\gamma_t^o - \gamma_t^{o^H}\right) + \varepsilon^{\gamma^o}{}_\theta \tag{2.13}$$

being $a^o{}_{\theta,t}$ the coefficients of the autoregressive process, θ_g the order of the process encompassing n_g+1 time slices, and the error term. According to matrix terms, (2.13) becomes:

$$\Delta\boldsymbol{\gamma}_{\boldsymbol{\theta}+1} = \sum_{t=\theta-\theta_g}^{\theta} \mathbf{a}_{\boldsymbol{\theta},\mathbf{t}} \Delta\boldsymbol{\gamma}_t + \boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}} \tag{2.14}$$

being $\mathbf{a}_{\boldsymbol{\theta},\mathbf{t}}$ the diagonal square matrix of order n_o containing the coefficients of the autoregressive process related to time slices t and θ, given by:

$$\mathbf{a}_{\boldsymbol{\theta},\mathbf{t}} = \begin{bmatrix} a_{\theta,t}^{o_1} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & a_{\theta,t}^{o_{n_o}} \end{bmatrix} \tag{2.15}$$

and

$$\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}} = \begin{bmatrix} \varepsilon^{\gamma^{o_1}}{}_\theta & \dots & \varepsilon^{\gamma^{n_{o_n}}}{}_\theta \end{bmatrix} \tag{2.16}$$

the corresponding vector of error terms.

The transition of distribution shares is modelled through an autoregressive process as (2.13):

$$\pi_{\theta+1}^{d|o} - \pi_{\theta+1}^{d|o}{}^{H} = \sum_{t=\theta-\theta_p}^{\theta} b_{\theta,t}^{d|o}\left(\pi_t^{d|o} - \pi_t^{d|o^H}\right) + \varepsilon^{\pi^{d|o}}{}_\theta \tag{2.17}$$

wherein:

- $b_{\theta,t}^{d|o}$ are the coefficients of the autoregressive process;
- θ_p is the order of the process encompassing n_p+1 time slices and

- $\varepsilon^{\pi}{}_{\theta}^{d|o}$ is the error term.

The resulting $n_{od}\cdot 1$ vector of the error terms $\varepsilon^{\pi}{}_{\theta}^{d|o}$ is given by:

$$\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} = \left[\varepsilon_{\theta}^{\pi_{d_1|o_1}} \quad \dots \quad \varepsilon_{\theta}^{\pi_{d_{n_d}|o_1}} \quad \dots \quad \varepsilon_{\theta}^{\pi_{d_1|o_{no}}} \quad \dots \quad \varepsilon_{\theta}^{\pi_{d_{n_d}|o_{no}}} \right]' \tag{2.18}$$

The assumption of stability of shares introduced by Ashok (1996) and Ashok and Ben-Akiva (2000) is formally defined by setting Var[$\varepsilon^{\pi}{}_{\theta}^{d|o}$]<<Var[$\varepsilon^{\gamma}{}_{\theta}^{o}$], which expresses that the temporal evolution of the distribution shares is less stochastic than the generation profiles.

To efficiently introduce the quasi-dynamic assumption into the Kalman filter framework, Marzano et al. (2019) apply a state-space augmented filter specification spanning from the current time slice θ back to the time slice θ-θ_s, i.e. considering n_s+1 time slices.

A state-space augmented Kalman filter starts from (2.10), yielding the following augmented vector of generation profiles :

$$\Delta\boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}} = \boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}} - \boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}^H} = \begin{bmatrix} \Delta\boldsymbol{\gamma}_{\boldsymbol{\theta}} \\ \vdots \\ \Delta\boldsymbol{\gamma}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_{\boldsymbol{\theta}} - \boldsymbol{\gamma}_{\boldsymbol{\theta}}^{H} \\ \vdots \\ \boldsymbol{\gamma}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s} - \boldsymbol{\gamma}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s}^{H} \end{bmatrix} \tag{2.19}$$

and, from 2.12, yielding the following augmented vector of distribution shares :

$$\Delta\boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} = \boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} - \boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\pi}^H} = \begin{bmatrix} \Delta\boldsymbol{\pi}_{\boldsymbol{\theta}} \\ \vdots \\ \Delta\boldsymbol{\pi}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\pi}_{\boldsymbol{\theta}} - \boldsymbol{\pi}_{\boldsymbol{\theta}}^{H} \\ \vdots \\ \boldsymbol{\pi}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s} - \boldsymbol{\pi}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s}^{H} \end{bmatrix} \tag{2.20}$$

The resulting augmented state-space vector $\Delta\boldsymbol{\chi}_{\boldsymbol{\theta}}$ in terms of differences from historical values, related to time slice θ, is defined as:

$$\Delta\boldsymbol{\chi}_{\boldsymbol{\theta}} = \begin{bmatrix} \Delta\boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}} \\ \Delta\boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}} - \boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}^H} \\ \boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\pi}} - \boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\pi}^H} \end{bmatrix} = \begin{bmatrix} \Delta\boldsymbol{\gamma}_{\boldsymbol{\theta}} \\ \vdots \\ \Delta\boldsymbol{\gamma}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s} \\ \Delta\boldsymbol{\pi}_{\boldsymbol{\theta}} \\ \vdots \\ \Delta\boldsymbol{\pi}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_{\boldsymbol{\theta}} - \boldsymbol{\gamma}_{\boldsymbol{\theta}}^{H} \\ \vdots \\ \boldsymbol{\gamma}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s} - \boldsymbol{\gamma}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s}^{H} \\ \boldsymbol{\pi}_{\boldsymbol{\theta}} - \boldsymbol{\pi}_{\boldsymbol{\theta}}^{H} \\ \vdots \\ \boldsymbol{\pi}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s} - \boldsymbol{\pi}_{\boldsymbol{\theta}-\boldsymbol{\theta}_s}^{H} \end{bmatrix} \tag{2.21}$$

with dimension $[n_o(n_s+1)+n_{od}(n_s+1)]\cdot 1$. The corresponding transition between the augmented state-space vectors $\Delta\boldsymbol{\chi}_{\boldsymbol{\theta}+1}$ and $\Delta\boldsymbol{\chi}_{\boldsymbol{\theta}}$ is given by:

$$\Delta\boldsymbol{\chi}_{\boldsymbol{\theta}+1}^{\boldsymbol{\gamma}} = \boldsymbol{\Phi}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}} \Delta\boldsymbol{\chi}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}} + \mathbf{E}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}} \tag{2.22}$$

wherein the augmented state-space vector is given by (2.19) and the matrix of the coefficients of the autoregressive process is expressed by:

$$\boldsymbol{\Phi}_{\boldsymbol{\theta}}^{\boldsymbol{\gamma}} = \left[\begin{array}{ccc:c} \mathbf{a}_{\boldsymbol{\theta},\boldsymbol{\theta}} & \mathbf{a}_{\boldsymbol{\theta},\boldsymbol{\theta}-1} & \dots & \mathbf{a}_{\boldsymbol{\theta},\boldsymbol{\theta}-\boldsymbol{\theta}_s} \\ \hdashline \multicolumn{3}{c:}{\mathbf{I}} & \mathbf{0} \end{array}\right] \tag{2.23}$$

where the first n_o rows contain n_s+1 blocks $\mathbf{a_{\theta,t}}$ $t \in \theta...\theta\text{-}\theta_s$ of dimension $n_o \cdot n_o$ given by (2.15), and then a lower block with an identity matrix $\mathbf{I}$ of dimension $n_o n_s \cdot n_o n_s$ is appended with a matrix of zeros $\mathbf{0}$ of dimension $n_o n_s \cdot n_o$. Hence, matrix (2.23) is a square matrix of order $n_o(n_s+1)$. Furthermore, since usually the coefficients of the autoregressive process do not depend upon the specific time slice θ, but only on the relative lag between time slices, the coefficient matrix (2.23) is constant across time slices θ, i.e.

The error vector in equation (2.22) is given by:

$$\mathbf{E_\theta^\gamma} = \left[\boldsymbol{\varepsilon}_\theta^\gamma \;\; \boldsymbol{\varepsilon}_{\theta-1}^\gamma \;\; \dots \;\; \boldsymbol{\varepsilon}_{\theta-\theta_s}^\gamma\right]' \tag{2.24}$$

wherein the first n_o rows are given by the error vector (2.15).

The transition between distribution shares is analogous to equation (2.22), yielding to:

$$\Delta\boldsymbol{\pi}_{\tau+1} = \Delta\boldsymbol{\pi}_\tau + \mathbf{E}_\tau^\pi \tag{2.25}$$

wherein the matrix Φ_θ^π containing the coefficient of the auto-regressive process is given by:

$$\Phi_\theta^\pi = \left[\begin{array}{ccc:c} \mathbf{b}_{\theta,\theta} & \mathbf{b}_{\theta,\theta-1} & \dots & \mathbf{b}_{\theta,\theta-\theta_s} \\ \hdashline \multicolumn{3}{c:}{\mathbf{I}} & \mathbf{0} \end{array}\right] \tag{2.26}$$

where the first n_{od} rows contain n_s+1 blocks $\mathbf{b_{\theta,t}}$ $t \in \theta...\theta\text{-}\theta_s$ of dimension $n_{od} \cdot n_{od}$ as in 2.14, and then a lower block with an identity matrix $\mathbf{I}$ of dimension $n_{od} n_s \cdot n_{od} n_s$ is appended with a matrix of zeros $\mathbf{0}$ of dimension $n_{od} n_s \cdot n_{od}$. Hence, (2.26) is a square matrix of order $n_{od} \cdot (n_s+1)$. Finally, as in 2.24, the error term in equation 2.25 is given by:

$$\mathbf{E_\theta^\pi} = \left[\boldsymbol{\varepsilon}_\theta^\pi \;\; \boldsymbol{\varepsilon}_{\theta-1}^\pi \;\; \dots \;\; \boldsymbol{\varepsilon}_{\theta-\theta_s}^\pi\right]' \tag{2.27}$$

wherein all vectors have dimension $n_{od}(n_s+1) \cdot 1$ and $\boldsymbol{\varepsilon}_\theta^\pi$ is the error vector (2.18). The expression of the transition equation combining (2.22) and (2.25) in state-space augmentation is given by:

$$\Delta\boldsymbol{\chi}_{\theta+1} = \Phi_\theta \cdot \Delta\boldsymbol{\chi}_\theta + \mathbf{E}_\theta \tag{2.28}$$

wherein the augmented state-space vectors are defined by (2.21), the coefficients matrix of the process is a square matrix of order $(n_o+n_{od}) \cdot (n_s+1)$ given by:

$$\Phi_\theta = \left[\begin{array}{c:c} \boldsymbol{\Phi}_\theta^\gamma & \mathbf{0} \\ \hdashline \mathbf{0} & \boldsymbol{\Phi}_\theta^\pi \end{array}\right] \tag{2.29}$$

being $\boldsymbol{\Phi}_\theta^\gamma$ given by (2.23) and $\boldsymbol{\Phi}_\theta^\pi$ by (2.26), and the error vector is a $(n_o+n_{od}) \cdot (n_s+1)$ vector appending the corresponding error vectors (2.24) and (2.27):

$$\mathbf{E}_\theta = \left[\begin{array}{c} \mathbf{E}_\theta^\gamma \\ \hdashline \mathbf{E}_\theta^\pi \end{array}\right] \tag{2.30}$$

The corresponding variance matrix of the errors (2.30) will be denoted in the following as $\mathbf{Q_\theta}$ and is defined as:

$$\mathbf{Q_\theta} = \begin{bmatrix} \mathbf{Q_\theta^\gamma} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q_\theta^\pi} \end{bmatrix} \tag{2.31}$$

with the upper-left block of dimension $n_o(n_s+1)$ and the lower-right block of dimension $n_{od}(n_s+1)$, leading to an overall square matrix of order $(n_o+n_{od})\cdot(n_s+1)$. Normally, both $\mathbf{Q_\theta^\gamma}$ and $\mathbf{Q_\theta^\pi}$ are assumed diagonal, which can be calculated either based on the estimation of the autoregressive process or considered as design parameters of the filter set by the modeller.

To effectively implement the assumption of constant distribution shares within each quasi-dynamic interval, the formulation integrates a single state variable $\pi_\tau^{d|o}$ representing a constant distribution share over the duration of the state-space augmented rolling horizon spanning over n_s+1 time slices. All $\pi_\tau^{d|o}$ shares can be ordered in a vector $\mathbf{\Delta\pi_\tau}$ defined as:

$$\Delta\boldsymbol{\pi}_\tau = \boldsymbol{\pi}_\tau - \boldsymbol{\pi}_\tau^H = \begin{bmatrix} \Delta\boldsymbol{\pi}_\tau^{\mathbf{o_1}} \\ \dots \\ \Delta\boldsymbol{\pi}_\tau^{\mathbf{o_{n_o}}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\pi}_\tau^{\mathbf{o_1}} - \boldsymbol{\pi}_\tau^{\mathbf{o_1}\,H} \\ \dots \\ \boldsymbol{\pi}_\tau^{\mathbf{o_{n_o}}} - \boldsymbol{\pi}_\tau^{\mathbf{o_{n_o}}\,H} \end{bmatrix} \tag{2.32}$$

of dimension $n_{od}\cdot 1$ containing all distribution shares for the augmented state-space rolling horizon τ. Hence, the resulting augmented state-space vector $\Delta\boldsymbol{\chi}_\theta$, related to time slice θ and to a state-space augmentation horizon covering n_s+1 time slices, is defined as:

$$\Delta\boldsymbol{\chi}_\theta = \begin{bmatrix} \Delta\boldsymbol{\chi}_\theta^\gamma \\ \Delta\boldsymbol{\pi}_\tau \end{bmatrix} = \begin{bmatrix} \boldsymbol{\chi}_\theta^\gamma - \boldsymbol{\chi}_\theta^{\gamma\,H} \\ \boldsymbol{\pi}_\tau - \boldsymbol{\pi}_\tau^H \end{bmatrix} = \begin{bmatrix} \Delta\boldsymbol{\gamma}_\theta \\ \vdots \\ \Delta\boldsymbol{\gamma}_{\theta-\theta_s} \\ \Delta\boldsymbol{\pi}_\tau \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_\theta - \boldsymbol{\gamma}_\theta^H \\ \vdots \\ \boldsymbol{\gamma}_{\theta-\theta_s} - \boldsymbol{\gamma}_{\theta-\theta_s}^H \\ \boldsymbol{\pi}_\tau - \boldsymbol{\pi}_\tau^H \end{bmatrix} \tag{2.33}$$

Consequently, at each iteration the filter estimates n_s+1 (i.e. as many as the number of time slices embedded in the state-space augmentation) vectors of generation profiles $\boldsymbol{\gamma}$ and a single matrix of distribution shares $\boldsymbol{\pi}$, thus reducing the number of unknowns at each step of the filter. Consistently, the transition equation of the distribution shares is modified to account for the fact that a single vector of $n_{od}\cdot 1$ distribution shares for the entire duration of the rolling horizon is defined, thus yielding a consistent adjustment of the dimensions of the corresponding vectors/matrices. Therefore the transition equation is given by a special case of an autoregressive process of order -1:

$$\Delta\boldsymbol{\pi}_{\boldsymbol{\tau+1}} = \Delta\boldsymbol{\pi}_{\boldsymbol{\tau}} + \mathbf{E}_{\boldsymbol{\tau}}^{\boldsymbol{\pi}} \tag{2.34}$$

wherein the vectors $\Delta\boldsymbol{\pi}_{\tau+1}$ and $\Delta\boldsymbol{\pi}_{\tau}$ have dimension $n_{od}\cdot 1$.

The final transition equation is formally equal to 2.25 with the state-space vector defined in 2.34:

$$\Delta\boldsymbol{\pi}_{\tau+1} = \Phi^{\pi}_{\tau}\Delta\boldsymbol{\pi}_{\tau} + \mathbf{E}^{\boldsymbol{\pi}}_{\tau}$$

Consistently, the coefficients matrix of the process is a square matrix of order $n_o\cdot(n_s+1)+n_{od}$ given by:

$$\Phi_{\boldsymbol{\theta}} = \begin{bmatrix} \boldsymbol{\Phi}^{\boldsymbol{\gamma}}_{\boldsymbol{\theta}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{2.35}$$

and the error vector is a $n_o\cdot(n_s+1)+n_{od}$ vector given by appending the corresponding error vectors (2.24) and the one in (2.27):

$$\mathbf{E}_{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{E}^{\boldsymbol{\gamma}}_{\boldsymbol{\theta}} \\ \mathbf{E}^{\boldsymbol{\pi}}_{\tau} \end{bmatrix} \tag{2.36}$$

The corresponding variance matrix of the errors (2.36) can be denoted as $\mathbf{Q}_{\boldsymbol{\theta}}$ and defined as:

$$\mathbf{Q}_{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{Q}^{\boldsymbol{\gamma}}_{\boldsymbol{\theta}} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}^{\boldsymbol{\pi}}_{\tau} \end{bmatrix} \tag{2.37}$$

with the upper-left block of dimension $n_o(n_s+1)$ and the lower-right block of dimension n_{od}, leading to an overall square matrix of order $n_o\cdot(n_s+1)+n_{od}$. As above, both $\mathbf{Q}^{\boldsymbol{\gamma}}_{\boldsymbol{\theta}}$ and $\mathbf{Q}^{\boldsymbol{\pi}}_{\boldsymbol{\theta}}$ are assumed diagonal, which can be calculated either based on the estimation of the auto-regressive process or set by the modeller.

The measurement equation, under the assumptions of non-congested conditions and error-free assignment is expressed as follows, explicitly written in terms of deviations from historical estimates:

$$\begin{aligned} f^{l}_{\theta} &= \sum_{t=\theta-\theta_a}^{\theta}\sum_{o}\sum_{d} m^{od,l}_{t,\theta}\left(\gamma^{o\,H}_{t} + \Delta\gamma^{o}_{t}\right)\left(\pi^{d|o\;H}_{\tau(t)} + \Delta\pi^{d|o}_{\tau(t)}\right) = \\ &= \sum_{t=\theta-\theta_a}^{\theta}\sum_{o}\sum_{d} m^{od,l}_{t,\theta}\,\nu\left(\Delta\gamma^{o}_{t}, \Delta\pi^{d|o}_{\tau(t)}\right) \end{aligned} \tag{2.38}$$

wherein $\nu\left(\Delta\gamma^{o}_{t}, \Delta\pi^{d|o}_{\tau(t)}\right)$ is a bilinear form of the state variables.

In matrix notation, the generic terms $m^{od,l}_{t,\theta}$ of the assignment matrix in (2.38) for a pair of time slices t,θ can be aggregated into a matrix $\mathbf{M}_{\mathbf{t},\boldsymbol{\theta}}$ yielding to the following expression:

$$\mathbf{f}_{\boldsymbol{\theta}} = \mathbf{M}_{\boldsymbol{\theta}}\boldsymbol{\nu}\left(\Delta\boldsymbol{\chi}_{\boldsymbol{\theta}}\right) + \boldsymbol{\varepsilon}^{\mathbf{f}}_{\boldsymbol{\theta}} \tag{2.39}$$

Under the assumption of error-free link counts and assignment matrix, it occurs $\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{\mathbf{f}} = 0$ and the corresponding error covariance matrix $\mathbf{R}_{\boldsymbol{\theta}}$ should be assumed null as well.

The overall QD-EKF formulation allows to calculate a conditional prediction of the state variables and of their variances at the current time slice θ based on θ-1:

$$\Delta\boldsymbol{\chi}_{\boldsymbol{\theta}|\boldsymbol{\theta}-1} = \boldsymbol{\Phi}_{\boldsymbol{\theta}-1}\Delta\boldsymbol{\chi}_{\boldsymbol{\theta}-1|\boldsymbol{\theta}-1} \tag{2.40}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{\theta}-1} = \boldsymbol{\Phi}_{\boldsymbol{\theta}-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}-1|\boldsymbol{\theta}-1}\boldsymbol{\Phi}'_{\boldsymbol{\theta}-1} + \mathbf{Q}_{\boldsymbol{\theta}} \tag{2.41}$$

The prediction is subsequently updated at θ according to the first partial derivatives of the measurement equation with respect to the state variables ($\boldsymbol{\Omega}_{\boldsymbol{\theta}}$) and through the calculation of the gain matrix $\mathbf{G}_{\boldsymbol{\theta}}$ of dimension $[n_o(n_s+1)+n_{od}]\cdot n_{lc}$:

$$\boldsymbol{\Omega}_{\boldsymbol{\theta}} = \left.\frac{\partial \mathbf{M}_t \mathbf{v}\left(\Delta\boldsymbol{\chi}_{\mathbf{t}}\right)}{\partial \Delta\boldsymbol{\chi}_{\mathbf{t}}}\right|_{\Delta\boldsymbol{\chi}_{\mathbf{t}} = \Delta\boldsymbol{\chi}_{\boldsymbol{\theta}|\boldsymbol{\theta}-1}} \tag{2.42}$$

$$\mathbf{G}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{\theta}-1}\boldsymbol{\Omega}'_{\boldsymbol{\theta}}\left(\boldsymbol{\Omega}_{\boldsymbol{\theta}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{\theta}-1}\boldsymbol{\Omega}'_{\boldsymbol{\theta}} + \mathbf{R}_{\boldsymbol{\theta}}\right)^{-1} \tag{2.43}$$

$$\Delta\boldsymbol{\chi}_{\theta|\theta} = \Delta\boldsymbol{\chi}_{\theta|\theta-1} + \mathbf{G}_{\theta}\left[\mathbf{f}_{\theta} - \mathbf{M}_{\theta}\mathbf{v}\left(\Delta\boldsymbol{\chi}_{\theta|\theta-1}\right)\right] \tag{2.44}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{\theta}-1} - \mathbf{G}_{\boldsymbol{\theta}}\boldsymbol{\Omega}_{\boldsymbol{\theta}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}|\boldsymbol{\theta}-1} \tag{2.45}$$

The necessary input to initialize the QD-EKF at time slice θ=0, defining a prior state vector $\Delta\boldsymbol{\chi}_{\boldsymbol{0|0}}$, with the same dimensions and structure as (2.33) and defining a prior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{0|0}}$, with same dimensions and structure of (2.35). The performances of the QD-EKF have been tested in experiments on both toy and real-size networks, leading to results outperforming all estimators tested by Cascetta et al. (2013) and with satisfactory results also for prediction.

2.1.4 **Summary**

The development of the quasi-dynamic framework has left three main open challenges:

- The extension of the quasi-dynamic framework to congested networks;
- The exploration of its performances adopting a more effective algorithm for the solution of the optimization problem.
- The assessment of the quasi-dynamic hypothesis in urban context analysing the magnitude and distribution of the intrinsic error. Indeed, such contexts are normally characterized by more complex o-d flow patterns; however, a significant percentage of the overall urban mobility is represented by systematic trips, for which the quasi-dynamic assumption is expected to be sufficiently acceptable.

Part of this thesis attempts to bring research on these three aspects. Specifically, point 1 and 2 are discussed in Chapter 3, while the assessment of the quasi-dynamic hypothesis in urban context can be found in the trajectory data analysis presented in Chapter 4 (see Section 4.5);

2.2 o-d flows estimation/updating in presence of trajectory data

Given its accurate spatio-temporal resolution, trajectory data has been often exploited, besides o-d flows estimation/updating purpose, to retrieve information on relevant modelling variables contributing on the demand estimation model accuracy such as mode (e.g. Wang et al. 2010; Larijani et al. 2015; Huang et al. 2019), trip end and purpose, traffic status and traffic parameters (e.g. Gundlegård & Karlsson 2009; Demissie et al. 2013; Park et al. 2014) or for mobility survey development (e.g. Toluei et al. 2017). As introduced in Section 1.1, trajectory data can be used for o-d flows estimation in various manners:

- to derive direct o-d flows estimates
- to infer route choice set and route choice probabilities;
- to infer the assignment matrix;
- to derive splitting rates at intersection;
- to use the precedent information to ameliorate existing o-d estimation methods.

An in-depth analysis of the literature regarding each application is reported in the following sections.

2.2.1 Direct estimation of o-d flows

Point-to-point data allows to derive a first direct estimate of the o-d matrix: generally, given a opportunely defined traffic analysis zoning, o-d flows are extracted by aggregating the origin and destination zones of each vehicle trace.

Although nowadays sensing technologies allow to collect and store great amount of data potentially enabling to completely observe the o-d matrix, trajectory data accessibility remains still poor due to a circumscribed/limited data ownership; movements data is mainly owned by private entities, thus it is generally unavailable due to privacy concerns and to protecting their own business interests. Notwithstanding, the available data is often offered as a product by vendors, e.g. some mobile phone carriers or dedicated companies such as INRIX, OCTOTELEMATICS etc. selling batch of data acquired from various sources (e.g. freight carriers/shippers, car insurance providers etc.). Given the opportunistic nature of this data collection/aggregation, trajectory data samples show poor penetration level, insufficient scale to apply inference techniques and significant biases, basically lacking representativity and

variety. As mentioned in Section 1.1, a crucial aspect in adopting trajectories for o-d flows estimation/updating relates to their penetration rate. Very high penetration rates, typical of smartphone providers, yield effective o-d flows estimation. Unfortunately, this type of data is very costly and not accessible to many researchers and practitioners, who usually must deal with small samples (i.e. trajectories collected with on-board-devices), characterized by low penetration rates. In light of this phenomenon, trajectory dataset can be interpreted as a (distorted) sample of observations from a population (universe) of vehicle trips, thus a necessary and preliminary procedure to obtain a better estimate of the o-d matrix is to opportunely scale the o-d matrix derived from trajectory data. Intuitively, the penetration level of the sample and the sample distortion determines the complexity of the scaling technique to apply.

As pointed out in section 1.1, the estimation of the trajectory sample penetration level is not a straightforward operation itself, since the true underlying of the total amount of vehicle trips is unknown. However, considering the specific case of cellular data samples, the penetration level can be estimated considering the market share of the mobile operator and in addition, individual profiling could help associating further socio-economic and trip information useful to compare mobile phone data to census data, dramatically simplifying the scaling procedure. Indeed, trip volumes and o-d flows derived from cellular datasets are usually upscaled using a unique scaling factor and homogeneous per o-d pair, resulting from the ratio between the total number of vehicles from the census data and the total number of sampled vehicles.

$$\tilde{\varepsilon} = \frac{\sum_{od} d_{od}^{CENSUS}}{\sum_{od} d_{od}^{traj}} \tag{2.46}$$

The upscaling technique considering a unique expansion factor among all o-d pairs is referred here as to direct scaling, leading to a direct estimation of o-d flows.

Therefore, the rich opportunities deriving from cellular data described above have led to the proposition of numerous applications during the past decade performing a direct estimation of o-d flows (e.g. Calabrese et al. 2011; Ma et al. 2013; Bahoken & Raimond 2013; Wang et al. 2013; Larijani et al. 2015; Alexander et al. 2015; Toole et al. 2015; Wu et al. 2015; Gundlegård et al. 2016; Ge & Fukuda 2016; Bonnel et al. 2018; Hadachi et al. 2019; Bachir et al. 2019). Some of the most relevant works on direct o-d flows estimation from cellular data are broadly described below. Calabrese et al. (2011) exploited opportunistically collected mobile phone location data to estimate o-d flows, demonstrating that mobile phone data can represent a proxy for human mobility, being able to capture weekday and weekend patterns as well as seasonal variations when compared to census data. Ma et al. (2013) similarly derive o-d matrices and mobility patterns from a mobile phone dataset. Given the extensive coverage and the high

market share owned by the mobile operator they obtain a high accuracy level of the sample o-d matrix in fine spatial and temporal resolutions. The derived seed matrices are coupled with surveyed commute flow data and prevalent travel demand modelling techniques to provide the o-d matrices for operational planning applications (e.g. dynamic traffic assignment models). Bahoken et al. (2013) first obtain o-d matrices triangulating mobile phone data, socio-demographic data and link counts data, then they analyse the effects deriving from different spatio-temporal data filtering techniques, thus defining the quality of information that can be retrieved according to diverse spatial aggregation (zoning type) and temporal resolution. They demonstrate that better information can be retrieved with larger temporal windows, while concerning the spatial filtering technique, flows were aggregated at three different scales: Voronoi, urban area (UA) and node. Results showed that the loss of information is not relevant when flows are aggregated from Voronoi scale to UA scale and it is relevant (55%) when flows are aggregated at poles nodes. Wang et al. (2013), estimating travel demand by time-of-day and commuting traffic data along a traffic corridor based on a 6 week mobile phone data observation, concluded that, due to the low resolution of location using the network-based cell phone network, the use of cell phone network in collecting traffic data would be more feasible for long distance or inter-city trips. A long-time observation could increase the cell phone sample size and could be useful in obtaining stable cell phone traffic, as well as reduce the bias of the data. Alexander et al. (2015) estimate average daily origin–destination trips from triangulated mobile phone records of millions of anonymized users, inferring the type of location (home, work or other) and trip departure time from census data. Results are again, validated against national survey data. Toole et al. (2015) estimate multiple aspects of travel demand using call detail records (CDRs) from mobile phones in conjunction with open and crowdsourced geospatial data, census records, and surveys to generate representative origin–destination matrices and route trips through road networks. Iqbal et al. (2014) developed a methodology to extract trip patterns and o-d matrices from CDR mobile phone data and implemented an optimization-based estimation to identify time-varying grouped scaling factors. The scaling rates estimation proposed is based on link traffic count measurements and utilises a microscopic simulation model to reproduce the spatial and temporal propagation of o-d flows through the network. Although the study provides a relevant example of GLS-based scaling method, it overlooks the heterogeneity in call rates from different locations since it introduces only two scaling factors defined according to zone adjacency.

When link traffic counts are available, another direct scaling procedure can be applied. This method, proposed by Van Aerde et al. (1993) consists of scaling up the trajectory o-d matrix

($\boldsymbol{d}_{od}^{traj}$) by means of an unique upscaling factor defined for some period ϑ (γ^{ϑ}) obtained as the average ratio of the total number of vehicles observed from link traffic counts at time-interval ϑ to the total number of tracked vehicles traversing link l at time-interval ϑ:

$$\gamma^{\vartheta} = \frac{\sum_{l=1}^{nlc} f_l^{\vartheta}}{\sum_{l=1}^{nlc} f_l^{traj,\vartheta}} \quad \forall\, \vartheta \in T \tag{2.47}$$

$$\boldsymbol{d}_{od}^{\boldsymbol{EXP},\vartheta} = \gamma^{\vartheta} \cdot \boldsymbol{d}_{od}^{\boldsymbol{traj},\boldsymbol{\theta}} \tag{2.48}$$

Drawing upon this result, few studies in literature dealt with trajectory data sample scaling (Jing et al. 2011; Yang et al. 2017; Mitra et al. 2020). For the scope of the thesis , the models proposed by Yang et al (2017) and Mitra et al. (2020) are reported in the following. Yang et al. (2017) applied the method proposed by Van Aerde et al. (1993) to obtain a crude estimation of the o-d matrix. An equivalent result can be achieved implementing a GLS-based formulation aiming at finding the optimum constant scaling factor for the reference period *t*, homogeneous among all o-d pairs, able to best fit link traffic measurements, as proposed in Mitra et al. (2020) for stationary conditions. Following Yang et al. (2017), Mitra et al. (2020) demonstrated that direct scaling model expressed in Eq. 2.47 tends to produce biased o-d estimates and although the resulting o-d matrix constitutes a good starting point, it would require additional adjustment with more complex scaling techniques to reach a higher level of accuracy. Indeed, such scaled o-d matrix ($\boldsymbol{d}_{od}^{\boldsymbol{EXP},\vartheta}$) can be used as target o-d matrix in the o-d flows updating process, both for dynamic applications as proposed by Yang et al. (2017) involving the simultaneous GLS estimator described in section 2.1.2.1, and for stationary conditions as in Mitra et al. (2020) using a static formulation of the GLS estimator. For the scope of this thesis the static formulation applied by Mitra et al. (2020) is reported in the following:

$$\{\boldsymbol{d}_{od}^{*}\} = \arg\min_{x_{od}} \left\{ \sum_{od=1}^{n_{od}} \frac{(x_{od} - \boldsymbol{d}_{od}^{\boldsymbol{EXP}})^2}{\sigma_{od}} + \sum_{l=1}^{n_{lc}} \frac{(m_{l,od}^{traj} \cdot x_{od} - \hat{f}_l)^2}{\sigma_l} \right\} \tag{2.49}$$

wherein $\boldsymbol{d}_{od}^{\boldsymbol{EXP}}$ is calculated according to 2.48 considering the time interval θ in which stationary condition of demand flows holds. The stationary conditions imply a unique, scalar upscaling coefficient by means of which the direct scaling is performed.

In both studies, the o-d matrix resulting from direct scaling (2.48) is used as target matrix and the traffic assignment model is replaced by an observation of the assignment matrix from the trajectory data sample ($m_{l,od}^{traj}$) obtained as in 2.53. While Yang et al. (2017) performed a dynamic o-d flows updating by means of synthetic experiments (although with a quite limited experimental plan) on a small-scale network, Mitra et al. (2020) tested the GLS estimator for stationary conditions using real data from the road network of Turin, simulating a morning

reference time interval of one hour and using as validation procedure a dataset of hold-out traffic counts. As further explained in Chapter 5, validation techniques and especially synthetic experiments are crucial to assess o-d estimation methods (see Section 5.1). Additionally, Yang et al. (2017) proposed a new formulation of the simultaneous GLS integrating a new term into the objective function, as briefly described in section 2.2.3 while Mitra et al. (2020) presented two alternative formulations of the static GLS estimator. Both formulations assume the demand flows as a function of an attraction and a distribution upscaling coefficients (α_o and β_d), basically introducing the hypothesis according to which different traffic zones have different scaling rates, thus the o-d pair whose origin/destination is the same as the traffic zone will share the same scaling rate (respectively α_o and β_d). This hypothesis yields to the following expression of o-d flows:

$$d_{od} = \alpha_o \cdot \beta_d \cdot \boldsymbol{d_{od}^{EXP}} \qquad (2.50)$$

wherein $\boldsymbol{d_{od}^{EXP}}$ is obtained applying the direct scaling as in 2.48. The first GLS formulation proposed by Mitra et al. (2020) consists of a quadratic optimization in which the attraction and distribution upscaling factors are jointly adjusted :

$$\{\boldsymbol{\alpha_o^*}; \boldsymbol{\beta_d^*}\} = \arg\min_{\alpha_o,\beta_d} \left\{ \sum_{od=1}^{n_{od}} \frac{(\alpha_o \cdot \beta_d \cdot \boldsymbol{d_{od}^{EXP}} - \boldsymbol{d_{od}^{EXP}})^2}{\sigma_{od}} + \sum_{l=1}^{n_{lc}} \frac{(m_{l,od}^{traj} \cdot \alpha_o \cdot \beta_d \cdot \boldsymbol{d_{od}^{EXP}} - \hat{f}_l)^2}{\sigma_l} \right\} \qquad (2.51)$$

s.t.

$\alpha_o \geq 0$

$\beta_d \geq 0$

Conversely, the second model implements an iterative procedure alternately adjusting one expansion factor at time, maintaining the other constant: at iteration 1 (2.52) attraction factors are considered as optimization variables and distribution factors are maintained constant ($\bar{\beta}_d$), while in iteration 2 (2.53) distribution factors are adjusted while maintaining attraction factors constant ($\bar{\alpha}_o$):

$$\{\boldsymbol{\alpha_o^*}\} = \arg\min_{\alpha_o} \left\{ \sum_{od=1}^{n_{od}} \frac{(\alpha_o \cdot \bar{\beta}_d \cdot \boldsymbol{d_{od}^{EXP}} - \boldsymbol{d_{od}^{EXP}})^2}{\sigma_{od}} + \sum_{l=1}^{n_{lc}} \frac{(m_{l,od}^{traj} \cdot \alpha_o \cdot \bar{\beta}_d \cdot \boldsymbol{d_{od}^{EXP}} - \hat{f}_l)^2}{\sigma_l} \right\} \qquad (2.52)$$

s.t.

$\alpha_o \geq 0$

$\beta_d = \bar{\beta}_d = cost$

$$\{\boldsymbol{\beta_d^*}\} = \arg\min_{\beta_d} \left\{ \sum_{od=1}^{n_{od}} \frac{(\bar{\alpha}_o \cdot \beta_d \cdot \boldsymbol{d_{od}^{EXP}} - \boldsymbol{d_{od}^{EXP}})^2}{\sigma_{od}} + \sum_{l=1}^{n_{lc}} \frac{(m_{l,od}^{traj} \cdot \bar{\alpha}_o \cdot \beta_d \cdot \boldsymbol{d_{od}^{EXP}} - \hat{f}_l)^2}{\sigma_l} \right\} \quad (2.53)$$

s.t.

$\beta_d \geq 0$

$\alpha_o = \bar{\alpha}_o = cost$

Authors also tested the proposed methods conducting a sensitivity analysis of the parameter σ_{od}, expressing the variance of the estimation error, that is the level of reliability of the prior o-d flow estimates. Intuitively, the lower this value the higher is the level of reliability of the prior o-d matrix. Considering a range of variation of σ_{od} from 0 to 1 [0,1], the authors concluded that it would be convenient to search for an optimal solution which preserves the structure of the target matrix, specifically the best results were obtained running the formulation in (2.51) with $1/\sigma_{od}$ equal to 0.001.

The considerations deduced in few studies dealing with trajectory data-driven o-d estimation can be summarized as follow:

- To streamline the o-d flows updating, the main inputs for the implementation of a GLS formulation can be derived from trajectory data: the a priori matrix and the assignment map (see Section 2.2.3);
- The direct scaling of the observed o-d matrix as presented in 2.48 is an essential precondition to implement more complex scaling techniques;
- The GLS-based estimators proposed by Mitra et al. (2020) lead to a slight improvement of the o-d matrix obtained from direct scaling and tend to be susceptible to overfitting the link counts measurements.

More importantly, from the literature review emerged that, to evaluate and validate the performances of any proposed models, it is necessary to account for the estimated trajectory sample penetration rate, the level of sample distortion and the dimensionality of the link counts sample. Therefore, these results prompted the need of a systematic analysis by means of

laboratory experiments investigating the impact of trajectory and link counts sample characteristics on o-d estimation model performances. The development of this study is fully and extensively described in Chapter 5.

Additionally, concerning the estimation of o-d flows for the within-day dynamic context, it is important to notice that trajectory data allows to identify the departure time distribution during the entire period of observation (Mitra et al. 2020), although it strongly depends\ upon dataset composition and specifically upon sample variety. The trajectory data analysis and the experimental plan developed in this thesis shed the light on the impact of these two aspects (see Chapter 4 and Chapter 5).

2.2.2 **Route choice set and route choice probabilities**

One of the most challenging issue arising in route choice modelling relates to the dimension of a real-world network which consists of an high number of nodes, links and centroids. The number of the feasible paths connecting each o-d pair often makes unmanageable the possibility to explicitly take into account all of them. This is the reason why the choice set definition process is one of the most studied topic when dealing with route choice modelling. A common method to define users' choice-set follows a selective approach: intuitively, each decision maker actually considers only a few relevant paths for moving from an origin to a certain destination. Therefore, the choice-set is generated according to some algorithms: the most common one consists of iteratively implementing the shortest path algorithm. Once a choice set is exogenously defined, the second issue lies in evaluating to what extent the choice set alternatives are perceived as distinct, because of their degree of physical overlapping or their similarities, and the implications of such perception on the route choice probabilities. To this end, trajectory data sample can be a valuable source to derive an observed path choice set defining the most likely-to-use paths, limiting the behavioural assumptions that these models normally require. A recent example focussing on trajectory-based route choice set generation is provided by Yao & Bekhor (2020), in which different path generation algorithms are evaluated using a large GPS trajectory dataset. Experimental data shows that 60% percent of the total observations can be covered (assuming a threshold of 80% overlap) using a single path, which is significantly in contrast with previous literature findings. The research compares different choice set generation algorithms to test their final coverage demonstrating that an algorithm taking into account the preference for higher hierarchical roads provide the maximum level of coverage (97% with 80% overlap threshold). Other examples of inferring path choice dimension variables from trajectory dataset can be found in Mitra et al. (2020), Tang et al. (2019), Parry & Hazelton (2012). Concerning the o-d estimation problem, Nigro et al. (2018)

demonstrated by means of synthetic experiments that path choice probabilities deriving from Floating Car Data (FCD) can be a much more reliable and powerful information with respect to FCD origin–destination flows to inform an optimization-based o-d estimation model, since they represent the traffic conditions and behaviours that vehicles experiment along the path. Similarly, Carrese et al. (2017) showed how route choice information and point-to point travel times can significantly improve the spatial and temporal accuracy of the estimated demand for both off-line and on-line demand estimation problem.

2.2.3 **Assignment map**

Trajectory data can show temporal and spatial relationships among o-d/path flows and link arrival flows (Kim & Jayakrishnan 2010). Traffic counts have been the most popular data source for o-d estimation, therefore trajectory data has been used as supplementary data to derive the mapping between link counts and o-d flows. Substituting traffic assignment model with trajectory-based assignment map can dramatically simplify the formulation of the demand estimation model. The relation between o-d flows and link flows can be considered linear in the case of known and fixed route choice in uncongested networks, while it is non-linear when accounting for congestion phenomena and complex route choice (i.e. urban networks, numerous alternative routes connecting an o-d pair). The mapping between o-d flows and link flows can be described in within-day dynamics by the four-dimensional assignment matrix **M**[*t,j*], whose generic element m_{lod}^{tj} represents the fraction of the *od* flow generated at time interval *t* and being at link *l* at time interval *j*. The assignment fractions (m_{lod}^{tj}) depends upon two sets of information, whose estimates can be easily derived from trajectory data:

- *link-path fractions*: expressing the proportion of a path flow passing the link l and describing the spatial-temporal propagation of the route flows throughout the network
- *route choice probabilities*: expressing the proportion of an o-d flow interesting the path k and describing the spatial-temporal o-d flows distribution among different paths connecting an o-d pair (see section 2.3.2).

Inferring arc-path shares and route choice probabilities from trajectory data can provide an estimation of the assignment matrix, avoid complex dynamic traffic assignment models and streamline the o-d estimation process. Notably, both sets of information depends upon o-d flows evolution across time.

Given a set of sampled trajectories, a direct estimate of such assignment fractions ($\widehat{m}_{l,od}^{t,j}$) can be derived considering the ratio between the total number of sampled vehicles travelling from the origin *o* to the destination *d* departed at time-interval *t* and using link *l at* time-interval *j*

($n_{l,od}^{t,j}$), and the total number of sampled vehicles travelling between the considered o-d pair ($n_{od}^{t,j}$):

$$\hat{m}_{l,od}^{t,j} = \frac{n_{l,od}^{t,j}}{n_{od}^{t,j}} \tag{2.54}$$

Nevertheless, Simonelli et al. (2019), developing laboratory experiments on real-size network, demonstrated that to achieve a satisfactory level of accuracy of the assignment map, a high penetration trajectory sample is required (70%); being part of this work, the overall study is reported in Chapter 6. There are few examples in literature testing assignment-free demand estimation models (e.g. Yang et al. 2017; Krishnakumari et al. 2019; Mitra et al. 2020). Yang et al. (2017), performing numerical experiments by means of simulation datasets testing different formulations of an optimization-based dynamic o-d updating, explored a new way to construct assignment matrices directly from sampled probe trajectories to avoid sophisticated traffic assignment process. One of the proposed method, referred as to "Probe Ratio Assignment", explicitly considers the correlation between o-d probe vehicle penetration ratio (the proportion of probe vehicle in the total vehicle population within the same interval) and link probe vehicle penetration ratio (the ratio of observed link flow to corresponding link traffic flow during each interval), introducing the utilization of a new set of field observations. Furthermore, Mitra et al. (2020), substituted the traffic assignment computation by deriving an assignment map using path probabilities and arc-path shares observed from FCD. Furthermore, Krishnakumari et al. (2019) proposed an assignment-free data-driven o-d estimation method adopting only two main assumptions on human behaviour regarding path choice dimension: the magnitude of the number of chosen paths and the proportionality of path flows between these origins and destinations.

2.2.4 **Splitting rates at intersections**

A sample of trajectory data contains the turning fraction registered at every node interested by a trajectory, notably this is a valuable information regardless of the sample penetration level. An example of such application is proposed by Barceló et al. (2010), in which Bluetooth sensors are tested to detect mobile devices along a motorway corridor. The goal is to provide reliable estimates of average travel time and speed data, to be used as additional source of information to link traffic for the dynamic od estimation.

2.3 Summary

The classical o-d flows updating methods (described in Section 2.1) can be enriched with inputs from trajectory data. In light of this, in many of the studies mentioned in the previous sections, authors derive the information on relevant modelling variables and/or a prior o-d matrix by processing trajectory data; subsequently they adopt trajectory-based information to perform the classical o-d updating procedures or proposing novel and alternative formulations (Mitra et al. 2020; Carrese et al. 2017; Cantelmo et al. 2017; Ge and Fukuda 2016; Iqbal et al. 2014; Gudlegård et al. 2016; Kim et al. 2018; Krishnakumari et al. 2019; Michau et al. 2015; Montero et al. 2019; Nigro et al. 2018; Parry and Hazelton 2012; Wu et al. 2018; Yang et al. 2017). To summarize and provide a clear picture of the analysed literature, Table 2.1 reports a synopsis of the discussed works, wherein the first four columns indicate the trajectory-based applications of the previous Sections (2.2.1 – 2.2.4) and the last column indicates weather the authors perform the o-d updating procedure proposing novel and alternative formulations by integrating the information obtained by processing trajectory data.

Application / ***Study***	*Direct o-d Flows Estimation*	*Route Choice Probabilities*	*Assignment Map*	*Splitting Rates at Intersections*	*Trajectory data-based o-d Flows Updating*
Aerde et al. 1993	✓				
Alexander et al., 2015	✓				
Mitra et al. 2020	✓	✓	✓	✓	✓
Bachir et al., 2019	✓				
Bahoken and Raimond, 2013	✓				
Bonnel et al., 2015	✓				
Bonnel et al., 2018	✓				
Calabrese et al., 2011	✓				
Cantelmo et al. 2017					✓
Carrese et al., 2017	✓	✓			✓
Chen et al., 2017	✓				
Çolak et al., 2015	✓				
Ge and Fukuda, 2016	✓				✓
Gundlegård et al., 2016	✓				
Iqbal et al., 2014	✓				✓

Kim et al., 2018	✓				✓
Kim and Jayakrishnan, 2010			✓		
Krishnakumari et al., 2019	✓		✓		✓
Larijani et al., 2015	✓				
Ma et al., 2013	✓				
Markovic et al., 2019	✓				
Michau et al., 2015	✓				✓
Montero et al., 2019	✓				✓
Moreira-Matias et al., 2016	✓				
Mungthanya et al., 2019	✓				
Nigro et al., 2018		✓			✓
Parry & Hazelton, 2012		✓	✓		✓
Patrick, 2015	✓				
Toole et al., 2015	✓				
Wang et al., 2013	✓				
Yang et al., 2017	✓		✓		✓

Tabella 2.3 – Synopsis of trajectory data-based application for o-d estimation

2.4 Literature outcomes and research contributions

Literature studies exploring the use of trajectory data for o-d related analysis have demonstrated that a sample of trajectory data can provide a set of important information which effectively enriches existing and new model formulations of the o-d estimation problem. Few papers proposed enhanced o-d flows updating methods in presence of trajectory data, specifically dealing with scaling rate estimation, but none of them provided robust validation techniques to demonstrate proposed o-d estimation methods efficacy. These studies have raised several issues mostly related to the necessary and crucial investigation of trajectory data sample characteristics, significantly influencing models' outcomes at various levels (e.g. a priori o-d flows, assignment map entries, final o-d flows accuracy). Issues arising from various characteristics of trajectory data all converge to the fundamental challenge of identifying whether the dataset at hand represents the real travel pattern of a study area. Consequently, such analysis constitutes an essential step to be taken to generate a new level of understanding and

awareness pursuing this research path. Therefore, the main contribution of this research aims at investigating the impact of the various trajectory sample characteristics on o-d flows estimation process performances in order to provide structured guidelines for researchers and practitioners conducting o-d related analysis in presence of trajectory data. This contribution has been carried out by means of both empirical investigation, in which real trajectory data is analysed to identify the variability range of such sample characteristics (see Chapter 4), and of synthetic experiments, performing a sensitivity analysis to investigate the impact of different input parameters derived from trajectory data along the entire estimation process. Indeed the laboratory analysis is threefold, testing:

- Direct scaling methods providing a first estimate of the o-d matrix which can serve as the a priori matrix to initialise the updating process (see Chapter 5);
- Direct estimation of the assignment map entries (see Chapter 6);
- O-d flows updating procedures, implementing GLS-based estimators (the simultaneous and the quasi-dynamic GLS) on the basis of the outputs and the considerations derived from the analysis above (see Chapter 7).

Regarding the updating process, literature outcomes suggested that the use of trajectory data could be integrated with the quasi-dynamic approach to further reduce the unknown-to-observation ratio during the updating process. However, the development of the quasi-dynamic framework has left some open challenges: the extension of the quasi-dynamic framework to congested networks, the exploration of its performances adopting more effective algorithms for the solution of the optimization problem and the assessment of quasi-dynamic hypothesis in urban context. Therefore, this thesis attempts to bring light on these topics conducting a preliminary study investigating the unexplored properties of quasi-dynamic assumption, providing insights on how to implement the quasi-dynamic o-d estimation framework when dealing with congested networks (Chapter 3) and on the quasi-dynamic evolution of the demand in urban context (see Chapter 4) in order to support the application of quasi-dynamic framework to o-d updating methods in presence of trajectory data (Chapter 7).

using trajectory data to inform the QD-GLS method (see Chapter 7).

3 The quasi-dynamic assumption in congested networks

This Chapter provides experiments demonstrating quasi-dynamic estimators performances when applied to congested networks. This study allowed to test the QD-GLS performances in congested networks in terms of variance of the solution and reproduction congestion capabilities and to define the improvements that can be achieved using a derivative-free algorithm to solve the optimization problem.

3.1 Extension of the quasi-dynamic assumption to congested networks

To evaluate the performances of quasi-dynamic framework estimators when dealing with congested networks and to highlight the improvements achieved adopting derivative-free algorithms to solve the demand estimation problem, two methods of the quasi-dynamic framework have been considered: the QD-GLS estimator and the Two-Step (TS) approach developed by University of Luxembourg research group (Cantelmo et al. 2014, Cantelmo et al. 2015, Cantelmo 2018, Cantelmo and Viti 2020). Extending the quasi-dynamic o-d estimation framework to the case of congested networks implies the adoption of more complex methods and algorithms accounting for the bilinear form of the estimator, the non-convex objective function and the non-linear relation between link flows and o-d flows. Although the QD-GLS formulation remains the same as reported in equation 2.9, to adapt the QD-GLS method to the congested framework, the interaction between supply and demand must be developed by means of a dynamic traffic assignment (DTA) model which at each iteration of the estimation, computes the set of link flows consistently mapping the corresponding o-d flow.

Indeed, adopting the quasi-dynamic optimization variables framework (i.e. expressing o-d flows as the product of generations and distributions) and derivative-free algorithms to solve the estimation problem, Cantelmo et al. (2015) proposed a Two-Step approach separating the dynamic demand estimation problem into two sequential optimizations. Specifically, the first step uses a strict quasi-dynamic assumption to update the total generated demand volume for each traffic zone assuming fixed distribution shares over the entire time-horizon, while the second step adjusts o-d flow values using a simultaneous GLS (section 2.1.2.1).

The collaboration with the University of Luxembourg (Unilu) research group has allowed to conduct this study leveraging the research group solid set of expertise on the implications of congestion phenomena in dynamic demand estimation models and the crucial inputs for its development. Indeed, the two methods were tested on the heavy-congested test site of the inner ring of Antwerp (Belgium), a very-well validated test network by the extensive work of Unilu research group to assess the two-step approach and the derivative-free algorithms performances (Cantelmo et al. 2014, Cantelmo et al. 2015). Therefore, considering the two methods allowed not only to test the QD-GLS performances in congested networks in terms of variance of the solution and reproduction congestion capabilities, but also to define the improvements that can be achieved using a derivative-free algorithm to solve the optimization problem.

Although no considerations on demand values accuracy can be defined since no information on real demand values are available for the chosen test site network, the quasi-dynamic assumption provides robust results implementing different methods running and different algorithms, validating the estimation process reliability by proving that final demand estimates respect the real congestion pattern.

3.2 The tested methods

The Two-Step approach separates the demand estimation problem into two optimizations. Leveraging on a strict quasi-dynamic assumption, the first optimization procedure exploits the property of stability of shares adjusting prior estimates of total generated demand flows while maintaining constant the distribution probabilities. The methodology uses a strict quasi-dynamic simultaneous GLS estimator reformulating the objective function as:

$$\{\boldsymbol{g}^{*^1}, \dots, \boldsymbol{g}^{*^\theta}, \dots, \boldsymbol{g}^{*^{\boldsymbol{n_\theta}}}\} = \arg \min_{g^1 \dots g^{n_\theta} \in S_g} \left\{ \sum_{\theta=1}^{n_\theta} \sum_{l=1}^{n_{lc}} (f_l^\theta - \hat{f}_l^\theta)^2 \right\} \tag{3.1}$$

s.t.

$$g^1 \dots g^{n_\theta} \in S_g \colon d_{od}^\theta = g_o^\theta \cdot \hat{p}_{d|o}^\theta \ \forall o, \forall d, \forall \theta \tag{3.2}$$

wherein the unknowns are the demand generation profiles $\boldsymbol{g}^\theta$, $\hat{p}_{d|o}^\theta$ is the seed spatial/temporal distribution to move to destination *d* from origin *o* in time interval θ. The flow on the link *l* for the time slice θ (f_l^θ) is obtained directly by simulation, performing a DTA, underlying the dependence between supply and demand, while constraint (3.2) over-imposes, to the estimated matrix d_{od}^θ, the spatial/temporal structure of the historical demand.

Besides strongly reducing the number of decision variables, working with total generated trips can limit a demand overestimation during the demand estimation process, which is otherwise

likely to occur when dealing with congested networks. Furthermore, generation models are considered the most reliable models in transport engineering applications since total generated trips are more easily observable than o-d trips (Cascetta, 2009). This concept has also been recently analysed within a data driven framework (Krishnakumari et al. 2019). The main objective of this preliminary optimization phase is to obtain the right demand level, such that the updated demand matrix can provide a better initial point feeding the second step, consisting of a traditional optimization procedure using a simultaneous GLS estimator, thus improving temporal and spatial matrix distributions.

For the sake of clarity, the main differences between the two approaches of the quasi-dynamic framework are listed below:

- the Two-Step approach does not necessarily require to explicitly account for historical od flows within the objective function to reduce the number of possible solutions. To bind the research in the solutions space, constraint 3.2 imposes the distribution shares of the seed matrix and thus the information on the seed matrix structure.
- while the QD-GLS considers a probability function that captures the correlation between generation and distribution evolutions over a certain sub-period of time, the Two-Step approach assumes constant values of the distributions.

Concerning the optimization methods, different algorithms could be combined with both approaches. However, we propose here to analyse solution methods that have been adopted in the original papers in order to be able to compare our results with previous findings. Hence, the QD-GLS uses an interior-point algorithm (Karmarkar 1984), while the Two-Step uses the finite difference stochastic approximation (FDSA, Spall, 1992) for the first step and the simultaneous perturbation stochastic approximation - asymmetric design (SPSA-AD, Spall (1998a), Nigro et al. (2018)) for the second step, as briefly described below.

The interior point algorithm is also known as the “barrier method” because it introduces a barrier function into the objective function to bind the decision variables inside the feasible set. This enables the algorithm to avoid constraints violation at each iteration and solve a sequence of approximate unconstrained minimization problems exploiting estimated information about first and second order derivatives to search for the descent direction.

The FDSA follows approximately the gradient descent direction behaving like a gradient method. Specifically, at each iteration it calculates an approximation of the gradient perturbing every o-d pair independently, so the number of simulations required for computing the gradient at any iteration is equal to the number of o-d pairs plus the initial value of the objective function.

The SPSA is a path-search optimization method in which an approximation of the gradient is

computed simultaneously perturbing of all the variables. With respect to the FDSA, the gradient has a stochastic component, hence it requires a lower computational time to obtain the descent direction. With the SPSA-asymmetric design formulation, the number of necessary assignments to compute the gradient is reduced to 50% with respect to the basic SPSA with symmetric design (SD). It should be noted that the FDSA and the SPSA partially capture the nonlinear relationship between OD and link flows. However, this requires a higher accuracy and a longer computational time.

3.3 Test Site

The test case study relates to the inner ringway around Antwerp, Belgium (Figure 3.1). The network includes 56 links, 39 nodes, with 46 o-d pairs, all mainly connecting the different entry and exit points of this stretch of motorway and making rerouting options unlikely. The morning peak period considered occurs between 05:30 and 10:30. The field data - speeds and flows - were available every 5 minutes. The detectors are located at the on- and off-ramps and on some intermediate sections.

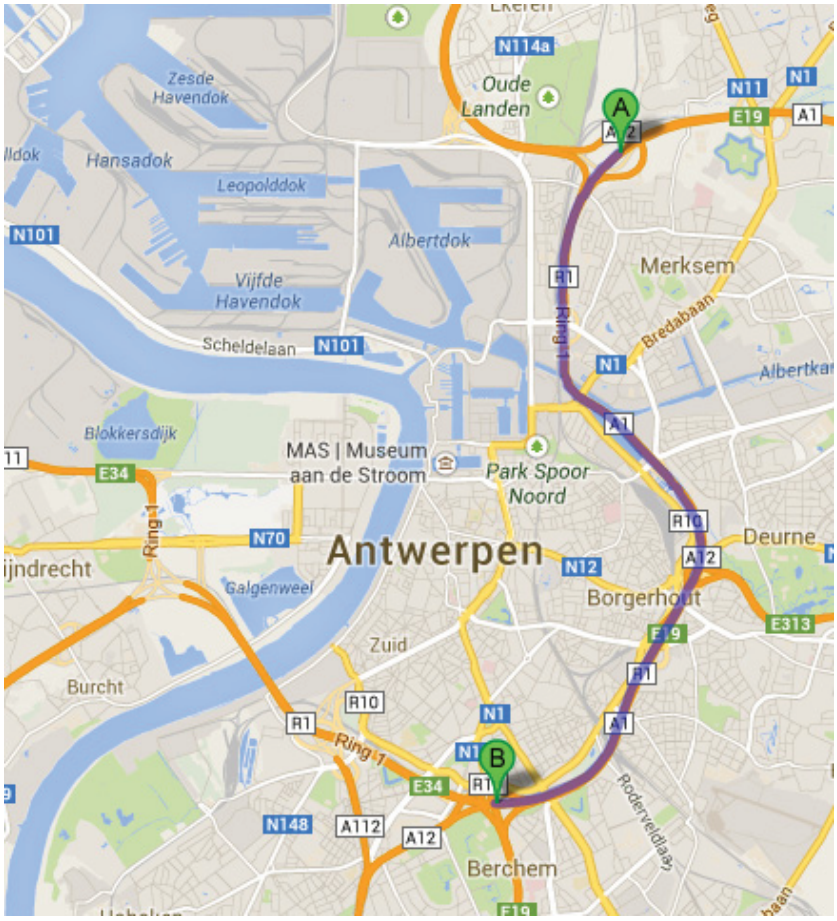


Figure 3.1 Test site: inner ringway of Antwerp, Belgium

The o-d flows were estimated for 15-minutes departure intervals, so the dynamic matrix contains 966 o-d pairs; the seed matrix, which amounts to 202,200 trips, is derived from an existing static o-d matrix by superimposing a time profile. Flows of a selection of o-d pairs were increased, so that the seed matrix has a congestion pattern similar to the actual one. In replicating the congestion pattern, the initial solution of the optimization also has a correct traffic pattern. shows the real congestion pattern, where each value on the y axis represents a (transversal) section of the inner ring, while on the x axes the time horizon is reported in minutes. As shown, the shaded blue colour area indicates low speeds, representing the time-space distribution of congestion phenomena affecting the inner ring.

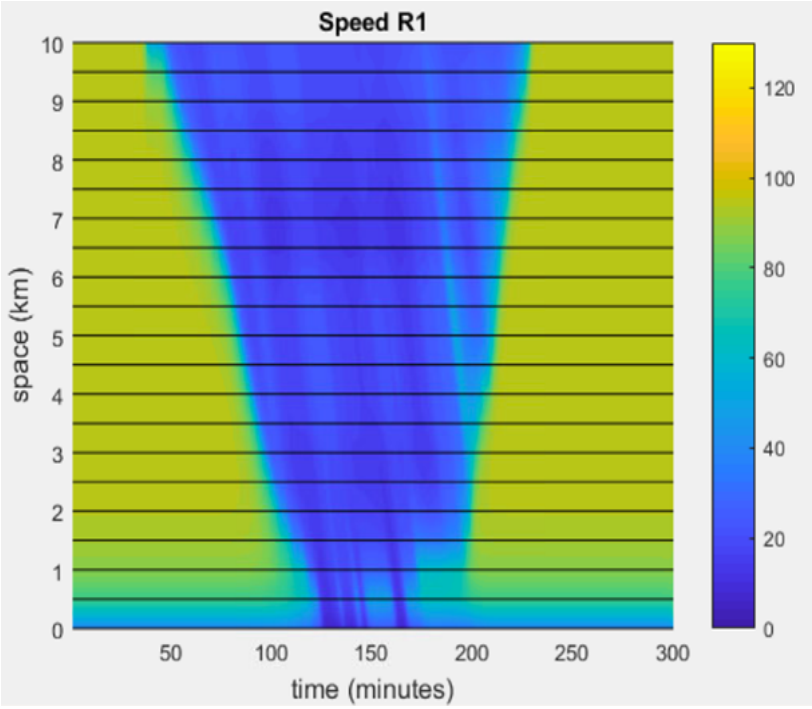


Figure 3.2 spatial and temporal plot of measured speeds on the network

3.4 Experimental Settings

Table 3.1 illustrates the experiments conducted adopting a wide experimental plan varying the settings for both methods and algorithms. Details of the chosen settings will be explained below.

		EXP	ALGORITHM	STEP SIZE	Obj. Function: GLS	CONVERGENCE CRITERIA (vehic/h)	n(θ|τ)	nvar	unkn/eq.
TS	STEP 1	I	FDSA	12	flows	Δ_{od}=10	-	126	0.33
		II	FDSA	15		Δ_{od}=5			
	STEP2		SPSA-AD	6	Flows & demand	Δ_{od}=1		966	2.56

QD	-	1	INTERIOR POINT	-	flows	STEP SIZE TOL. 10^{-10}	3	448	1.19
		2			Flows & demand				
		3			flows		21	172	0.46
		4			Flows & demand				

Table 3.1 Experimental Settings

Concerning the QD-GLS, two aspects were accounted to define the optimal settings for the experiments:

- the magnitude of the intrinsic error;
- the possible reduction of the number of the variables;

The presence of an intrinsic error has a key relevance as it represents a lower bound for the effectiveness of the QD-GLS estimator. Intuitively, the less reliable is the quasi-dynamic assumption within τ (i.e. the longer is the duration in which distribution shares are approximated to their average values to reduce the number of variables), the larger is the intrinsic error. Since both aspects depend on the duration of the reference period τ, the key parameters to set were the number of quasi-dynamic intervals ($n\tau$) and the corresponding number of time slices encompassing the reference period τ ($n_{\theta|\tau}$), as they define the trade-off between quasi-dynamic assumption and intrinsic error. In line with the findings provided by Marzano et al. (2009), we started from imposing a ratio between the number of variables and the number of equations equal to one leading to a value of $n_{\theta|\tau} = 3$ (EXP 1 – EXP 2). Choosing a number of time slices encompassing the reference period τ equal to 3 we assumed that distribution shares were stable in a reference period of 45 minutes duration. Afterwards, imposing this parameter equal to 21, we extended the assumption to the whole-time horizon duration (5 hours) and tested the minimum possible value for the ratio corresponding to the maximum reduction of the decision variables, although accepting a greater intrinsic error. Each quasi-dynamic setting was tested both including and not including the quadratic error on o-d flows (*flow, flow & demand*) into the objective function to evaluate the sensitivity of the algorithm in terms of distance from the starting matrix. The step-size and the descent direction for the QD-GLS experiments were iteratively computed by the algorithm choosing alternatively a projected conjugated gradient method or quasi-Newton method. As stopping criteria, a step size tolerance was adopted, thus the algorithm stopped when the relative change of the variables was less than a certain value reflecting the desired solution accuracy. In each experiment presented, no information related

to the dispersion matrix of the demand and of the counted flows distribution were introduced ($\sigma_{od}^{\theta}\ and\ \sigma_{l}^{\theta}$ set equal to 1).

For the Two-Step approach, the step-size (i.e. the advancement in the descent direction) is the most important parameter to be analysed. As the model combines the FDSA with a strict quasi-dynamic assumption, a small step leads to better estimations but implies greater computational times. After testing different values and different convergence criteria, the set of parameters reported in Table 1 provided the best compromise between efficiency and quality. For the first step, two experiments were conducted choosing respectively an initial step size equal to 12 (EXP I) and 15 (EXP II), while the convergence criteria were referred to a specific tolerance on the variables variation of two consecutive iterations, such that the algorithm stopped if the largest difference ($max\Delta_{od}$) between two consecutive solutions (o-d flows) was less than 10 vehicles per hour in the first experiment, while for the second experiment this parameter was set to 5. These values were related to the magnitude of starting matrix o-d flows, evaluated considering the average o-d flow initial values, resulting about 209.3 (vehicles/hour). For the second step a stricter convergence condition ($max\Delta_{od}=$ 1 vehicle/hour) was set to push the algorithm to deeply explore the solution space and thus guarantee no significant improvement of the solution was possible in a feasible computational time. On this purpose, both FDSA and SPSA-AD, implemented for the first and the second step respectively, were structured such that the algorithms allow to accept solutions even worse than the current one, to avoid getting stuck in the first local minimum found. To approximate the gradient, forward finite difference formulation is adopted in each experiment, since it has been demonstrated to be equivalent to the central finite difference formulation in terms of accuracy while more effective in terms of computational time (Cantelmo et al. 2014).

Table 3.1 also contains the ratio between unknowns and equations showing the level of balance reached: maintaining constant the distribution shares (STEP 1: EXP I – II), imposing no reduction technique (STEP 2) or for the QD-GLS experiments, choosing a different number of time slices encompassing the reference period (EXP 1-2, 3-4). To compute this ratio, a consistent number of equations deriving from collected link counts was considered, i.e. measurements referring to a time interval consistent with the estimation process time slice (15 min).

Results are presented in terms of some widely used indicators reported below:

$$Mean\ Square\ Error\ (\ MSE) \qquad \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n} \qquad (3.3)$$

$$Root\ Mean\ Square\ Error\ (RMSE) \qquad \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (3.4)$$

$$RMSE\ Coefficient\ of\ Variation\ (cvRMSE) \qquad \frac{RMSE}{\bar{y}} \qquad (3.5)$$

$$Coefficient\ of\ determination\ (r^2) \qquad \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}; \qquad (3.6)$$

Wherein:

- n is the number of observations
- $\boldsymbol{y}$ is a generic vector which can be specified either by starting demand vector either by measurement data vector
- $\bar{\boldsymbol{y}}$ is its average value
- $\hat{\boldsymbol{y}}$ is a generic vector which can be specified either by updated demand vector either by corresponding link flows/speeds vector deriving from the DTA simulation respectively.

3.5 Results

No considerations on o-d flows estimation accuracy can be made since no ground truth information related to real o-d flows are available. Therefore, results have been analysed in terms of distance from the historical matrix, link flow reproduction capabilities and reliability with respect to congestion pattern replicability of the real traffic regime.

Table 3.2 summarizes the results obtained. MSE and RMSE values are reported in vehicles/hour, while RMSE on speed data is reported in km/h. Final demand values were compared to historical ones (EST O-D MATRIX/SEED column), while link flows corresponding to the updated demand were compared against real measurements to evaluate data fitting performances (EST/ OBS LINK FLOWS column). Speed data were only used for validation purpose, comparing speed values corresponding to final demand against the real speed data (EST/OBS SPEED column).

The starting point of each experiment of the first step (two-step approach) presents an initial error on link flows corresponding to $cvRMSE_{link-flows} = 0.457, RMSE_{link-flows} = 1409\ veh/h$ providing a moderate fit with the traffic counts ($r^2 = 0.623$). For the quasi-dynamic method, the initial error on link flows is slightly different since the quasi-dynamic hypothesis introduces an intrinsic bias in the estimation process referred as to the aforementioned “intrinsic error”, depending on the duration of the reference period:

- n(θ|τ)=3: $cvRMSE_{link-flows} = 0.455$, $RMSE_{link-flows} = 1.400\ veh/h$;

- n(θ|τ)=21: $cvRMSE_{link-flows} = 0.459$, $RMSE_{link-flows} = 1.410\ veh/h$;

			EST O-D MATRIX/seed		est/ OBS Link Flows				est/OBS speed		
		EXP	cvRMSE	RMSE	cvRMSE	RMSE	r^2	%Reduction cvRMSE	RMSE	cvRMSE	C.T. [h]
TS	STEP 1	I	0.23	485	0.38	1170	0.659	16.8%	16.5	0.27	28.14
		II	0.50	104	0.34	1042	0.729	26.0%	17.8	0.29	110.88
	STEP 2	-	1.26	263	0.21	635	0.936	54.9%	20.6	0.33	228.00
QD	-	1	1.48	309	0.31	943	0.913	32.8%	20.7	0.33	24.10
		2	1.41	294	0.30	936	0.891	33.2%	18.4	0.29	65.44
		3	1.44	301	0.30	930	0.930	34.2%	15.2	0.25	10.44
		4	1.39	291	0.29	903	0.926	36.2%	16.2	0.26	5.56

Table 3.2 Results for the different settings reported in Table 3.1

Experiments on the QD-GLS estimator produce robust solutions since results are very close to each other although imposing different settings. Experiment number one and two provided good results slightly worse than experiment number three (EXP 3), which gave the best results in terms of link traffic counts and speed measurements reproduction, while experiment number four (EXP 4) provided the best improvement in terms of link flows root mean square error deviation reduction (36.2%) and the minimum distance from seed matrix. This show that imposing a stricter quasi-dynamic hypothesis satisfactory results can be obtained reaching convergence with low computational times (Figure 3.5). To evaluate the accuracy of the estimation in terms of o-d flows when dealing with congested networks, this aspect should be further tested in a controlled scenario in which true demand values are known.

As discussed in the previous works regarding the Two-Step approach, the first step is performed to preserve the starting o-d matrix structure maintaining constant the distribution probabilities and adjusting only total generated flows. This aspect can be observed analysing results on both experiments conducted for the first step (EXP I and II). Performance indicators comparing final demand and historical values (RMSE and RMSE coefficient of variance) indicate that the algorithm does not compromise the initial matrix structure, aspect further confirmed by a low dispersion from the root mean square value. Therefore, although in the first step results show

relatively poor data fitting performances, we optimize the total generated flows obtaining a better initial point for the second step.

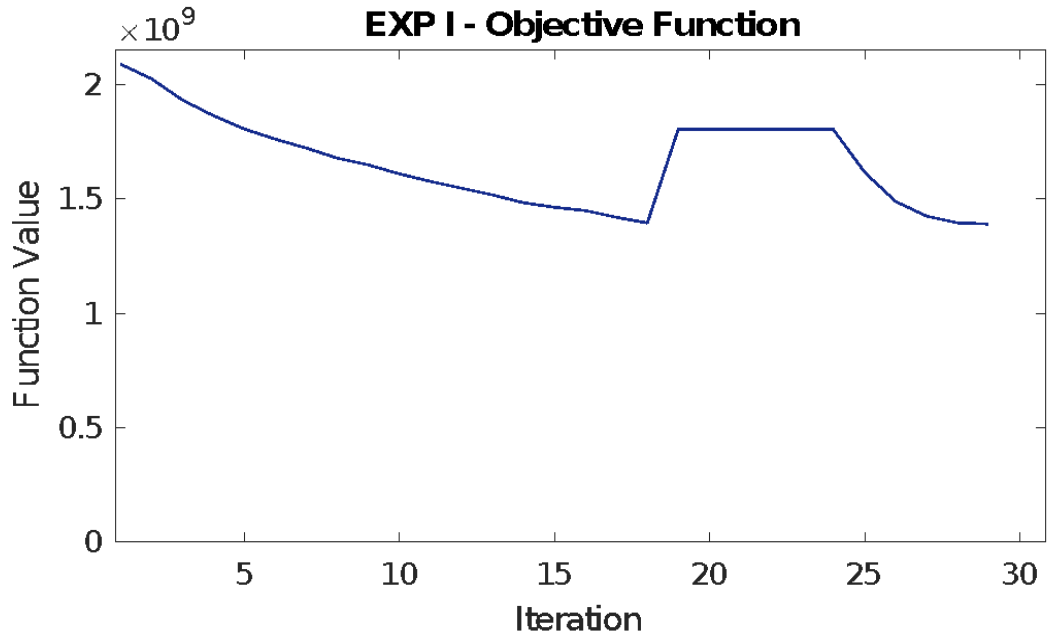


Figure 3.5 Evolution of the Objective Function of STEP 1 – EXP I;

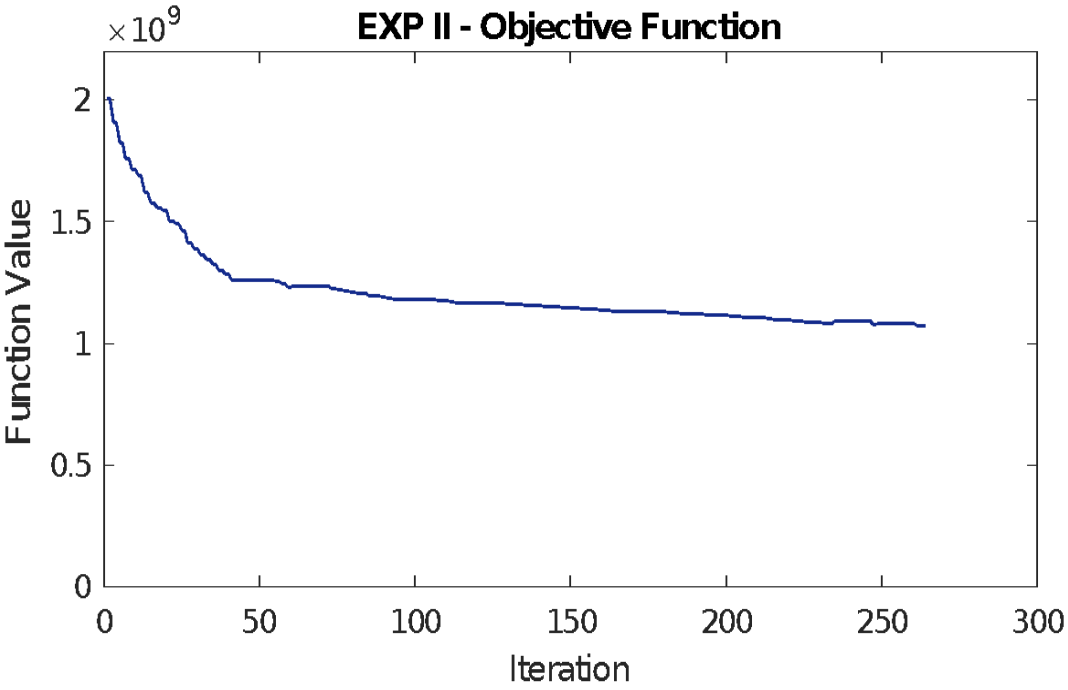


Figure 3.6 Evolution of the Objective Function of STEP 1 – EXP II

Comparing the performances of EXP I and EXP II varying the step size settings, it is known that running the algorithm with a smaller step can increase solution accuracy but also implies a longer computational time, since the algorithm requires a greater number of iterations to reach

a local minimum. In this case, setting a relaxed convergence condition (i.e. larger value of the maximum difference between consecutive solutions equal to 10 veh/h) the overall computational time remains moderate, reaching convergence after 29 iterations (Figure 2a). A better option would be to use a line search algorithm, which would drastically increase model performances. EXP II tests the opposite condition in which step size is increased, thus we experiment a rapid descent towards the minimum (i.e. with a greater slope) (Figure 2b). Nonetheless, imposing a stricter condition on the convergence, the overall computational time resulted longer than the first experiment (EXP I).

In the second step (STEP 2), the error on link flows further decreases, reaching the best level of fit ($r^2 = 0.936$) although with a significant computational time. Due to a reduced tolerance on the convergence criterion and the lack of a smart line search algorithm, the SPSA takes a significant amount of time in exploring the solution space, requiring a larger number of iterations to reach solution stability (*Figure* 3.7).

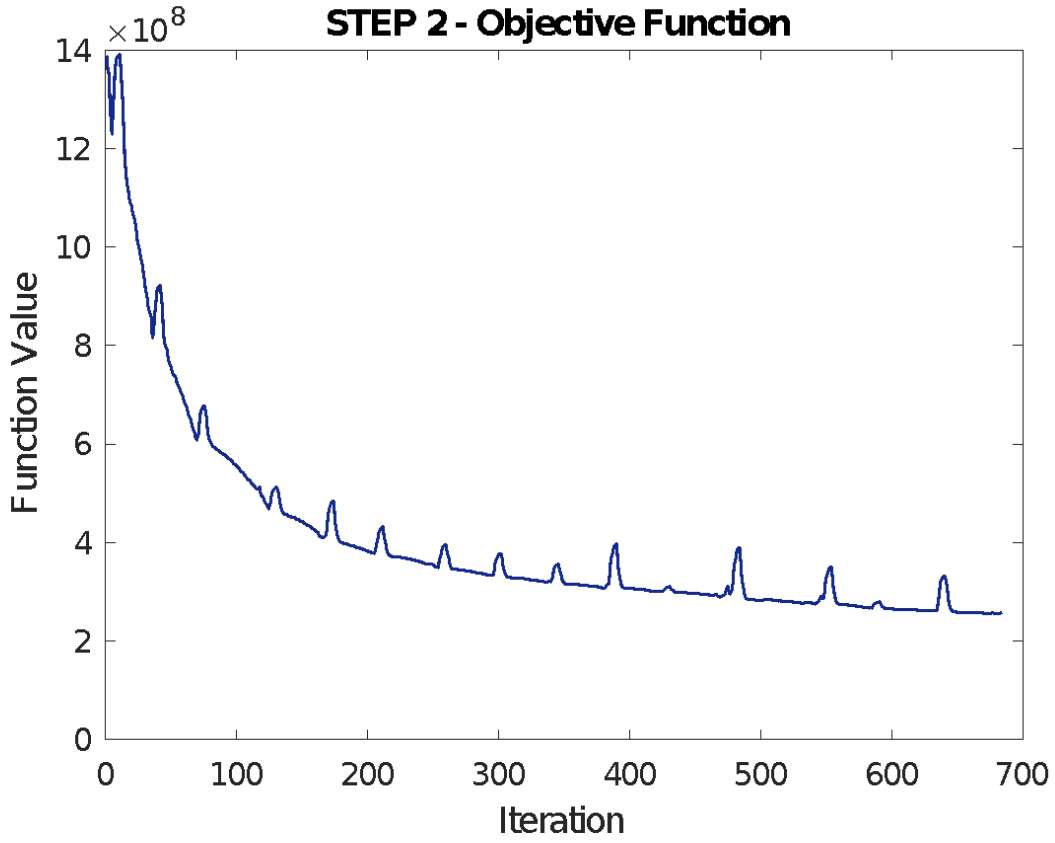


Figure 3.7 Evolution of the Objective Function of STEP 2 starting from the solution of STEP 1 EXP II

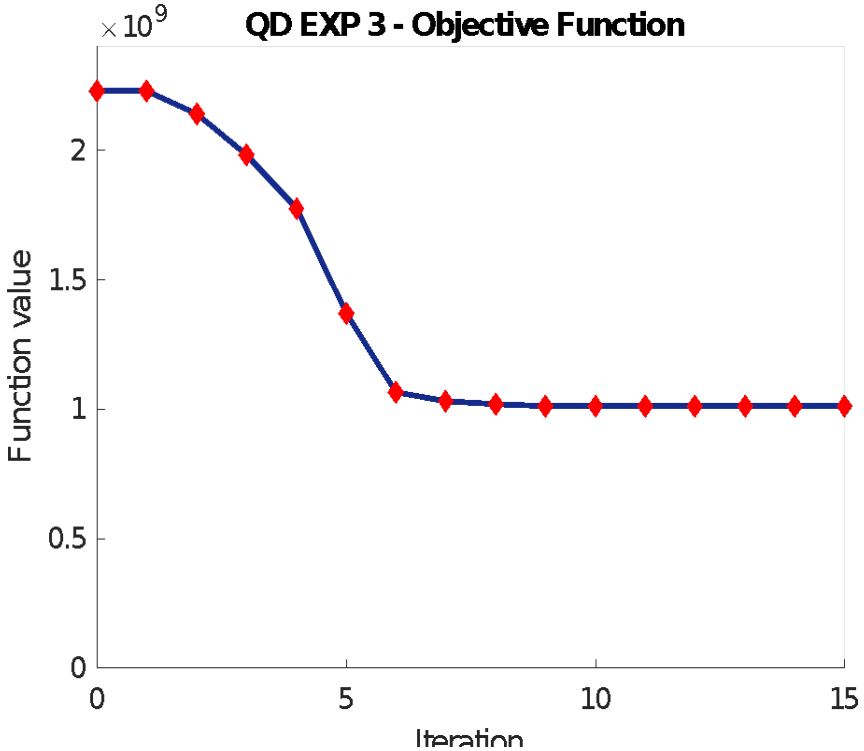


Figure 3.8 Evolution of the Objective Function of QD-GLS EXP 3

To evaluate solutions reliability, the estimated demand values were also validated in terms of speed and congestion patterns. A congestion pattern very close to the real one was obtained with all the tested scenarios as confirmed by observing speed RMSE and *cvRMSE* values reported in *Table* 3.2 (SPEED column). In Figure 3.9 and Figure 3.10 the space-time plots of the vector of the differences between simulated and measured speeds respectively resulting from the Two-Step approach and EXP 3 are presented, to validate best results obtained from both methods.

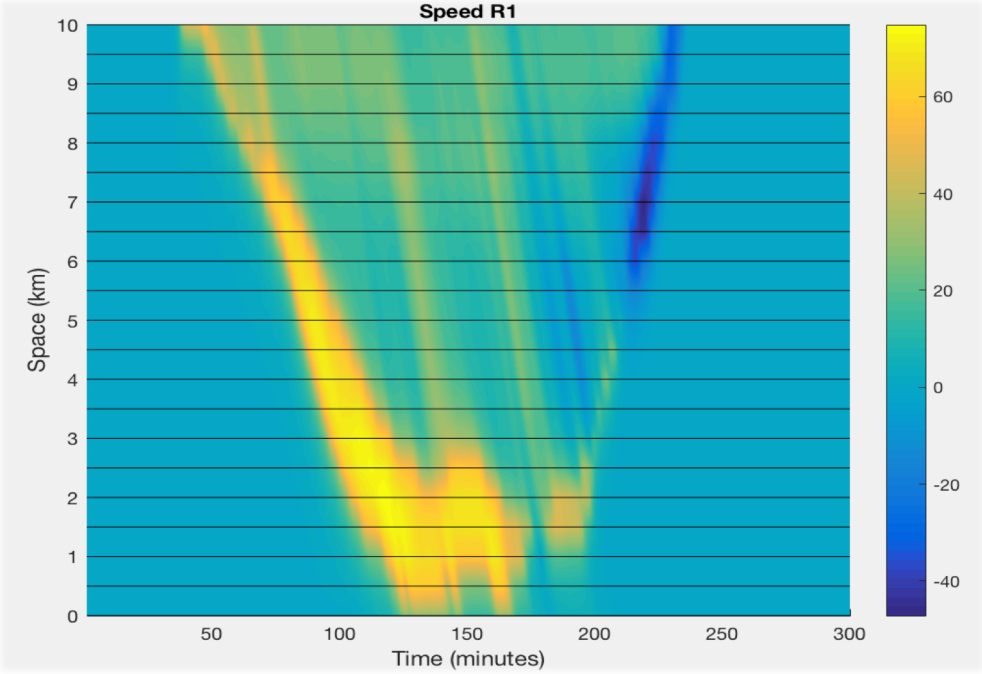


Figure 3.9 - Space-time Plot of the vector of the differences between simulated and measured speeds resulting from STEP 2

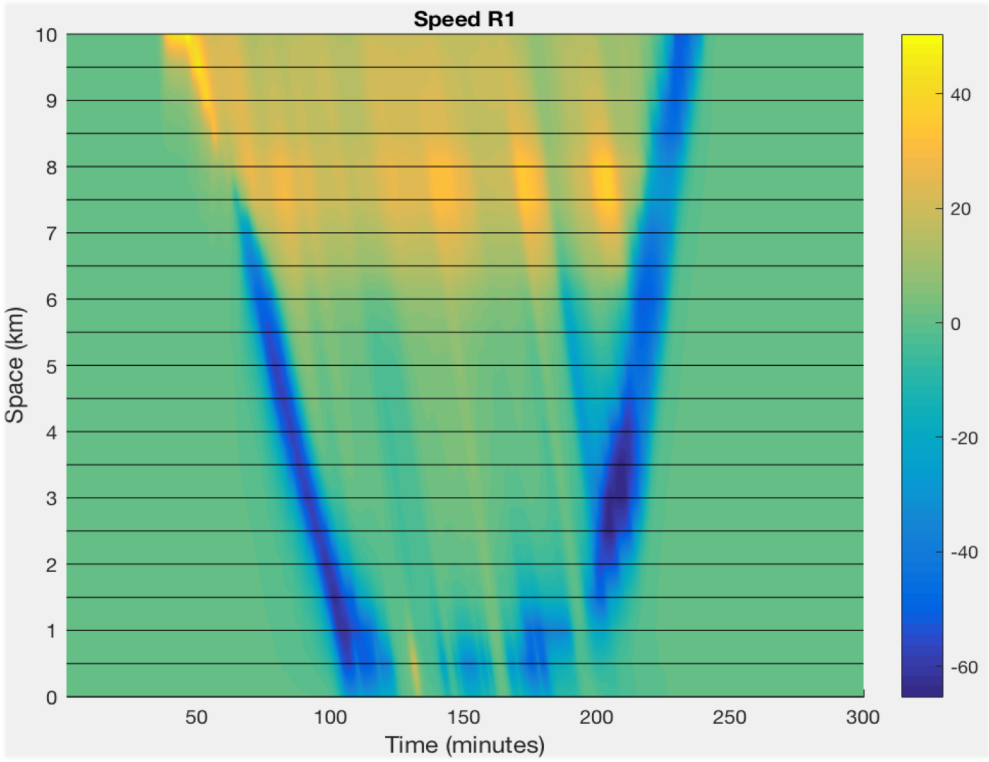


Figure 3.10 Space-time Plot of the vector of the differences between simulated and measured speeds resulting from EXP 3 – QD-GLS Experiments

While these differences show that the congestion pattern does not perfectly match, in both cases the model properly approximate both the beginning and the end of the congestion period, despite link speeds were not adopted within the objective function.

Overall, both models provide proper estimates with respect to all available traffic measures. This suggests that, on the one hand, the quasi-dynamic assumption is generally suited for o-d estimation problems and it is not dependent on one specific algorithm or implementation. On the other hand, we observed that the QD-GLS provide worse results but lower computational time, while the TS perform best for longer computational times. This suggests that different assumptions on the intrinsic error and its propagation have a huge influence on the performance of the model. Also, this phenomenon is not necessarily easy to predict, as for the QD-GLS a stricter assumption leads to shorter computational times, while for the TS, which also use a strict quasi-dynamic assumption, this is not the case. This suggests that the best compromise is to develop ad-hoc frameworks to leverage both these properties of the quasi-dynamic assumption, hence increasing model performance and fitting at the same time.

Best performances are achieved when the ratio between unknown variables and observations is lower than one. However, observability of the variables does not depend only on the number of sensors but also on their location (Yim et al. 1998, Bianco et al. 2001, Gan et al. 2005, Ehlert et al. 2006, Chen et al. 2007, Simonelli et al. 2012), as well as from the type of information available. To consider this, the TS approach combines a strict quasi-dynamic assumption (i.e. strict stationarity) with assignment matrix-free algorithms to optimally combine the quasi-dynamic assumption with multiple data sources. Then, in the second phase, the model relaxes the stationarity constraint to better fit the data. While this model is easier to implement, as it does not require to play with the intrinsic error generated by the quasi-dynamic assumption, the parameters of the optimization algorithm assume a central role to achieve good performance, as they establish when and how to relax the quasi-dynamic assumption.

Both models have been tested using real data on an highly congested network: the inner ring of Antwerp (Belgium).

3.6 Conclusions

Results from the two approaches can be summarized as follows:

- Both models provide good estimation results with respect to the error on the historical o-d flows and observed link flows, while acceptable for link speeds measurements;

- The QD-GLS performs best for a small ratio between unknown and equations, meaning assuming correlations between longer time periods. These settings ensure fast convergence and good results.

The TS, assuming constants distribution shares, provides the best results but also the highest computational time. These observations suggest that the quasi-dynamic assumption can have two corner solutions.

The first, is to introduce a relaxed quasi-dynamic assumption to quickly fit the data. This will result in a sub-optimal solution but computational times will be low. If the model is allowed to relax this assumption, a better solution will be found. However, computational time will increase at least linearly, as it increases with the number of variables (if SPSA/FDSA are used). This means that, under specific conditions, the quasi-dynamic model might find a good solution of the problem. Clearly, this depends on a series of considerations, including network topology, level of congestion and amount of information.

4 Trajectory data in Napoli

This Chapter contains an extended analysis of an opportunistic trajectory data sample composed of 50.933.281 GPS data points spanning over 31 days. The objective of this analysis is threefold:

- *Dataset statistical analysis;*
- *Trajectory o-d matrices and sampling rate estimation;*
- *Experimental analysis of the quasi-dynamic assumption in urban context;*

4.1 Dataset composition

The analysed trajectory sample was provided by a private company namely INRIX, one of the leading providers of mobility data. Its core business is to gather data from various sources such as road sensors and private and fleet vehicles to produce analytics and insights on human mobility to sell it as a product to mainly automotive and transport industries/agencies.
The dataset consists of 50.933.281 GPS data points spanning over 31 days of October 2017. The corresponding 2.328.471 trajectories were collected from 101.090 mobile devices, private and fleet vehicles crossing the geographic area highlighted in blue in Figure 4.1, approximately matching the entire city of Napoli and some of the surrounding suburb areas.

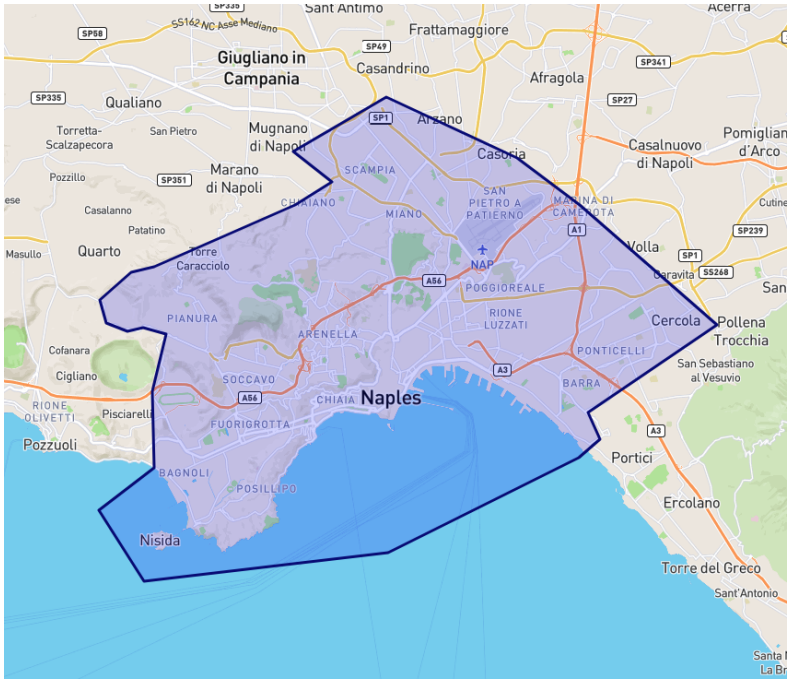


Figure 4.1 Naples urban area - Source: Inrix

In details, raw data included an array of .txt and .csv files, among which the most relevant were:

- *TripRecordsReportWaypoints.csv*: including the associated waypoints data for each trip (File I);
- *TripsRecordsReportProviderDetails.csv:* containing the trip providers details (File II);
- *TripRecordsReportReadMe.txt*: including the description of the fields contained in each .csv file (File III).

File I, II and III are further described in the following.

File I includes all of the details related to each of the 50.933.281 GPS data points collected during the entire period of observation, thus containing as many rows as the total recorded waypoints. Table 4.1 reports the most relevant data fields. The list of consecutive waypoints belonging to each trip can be either recognised from the waypoint sequence, starting with "1" and incrementing by one, either from the repetition of the same trip identifier along the first column. Each waypoint is spatially defined in terms of latitude and longitude and temporally by a timestamp (Capture Date) containing the capture date and time in UTC, ISO-8601 format.

Data field	Description
Trip ID	A trip's unique identifier corresponding to the waypoint recorded;
Waypoint Sequence	The order of the waypoint within the trip starting with "1" and incrementing by one;
Capture Date	The capture date and time of the waypoint in UTC, ISO-8601 format, example: "2014-04-01T08:33:35.000Z";
Latitude	The decimal degree latitude coordinates of the waypoint;
Longitude	The decimal degree longitude coordinates of the waypoint;

Table 4.1 waypoints details from raw data - field description

File II contains as many rows as the total number of recorded trips (2.328.471) and each row is associated with a number of fields containing specific characteristics of the trip (see Table 4.2). The trip's spatial and temporal information is specified by the origin location (Start point) and the destination location (End point) defined in terms of latitude and longitude, and by a timestamp containing the origins and destinations capture date and time in UTC, ISO-8601 format, namely "Start Date" and "End Date". Furthermore, this information is enriched with starting and ending day-type of the week (StartWDay and EndWDay fields) .

Data field	Description
Trip ID	A trip's unique identifier
Device ID	A device's unique identifier
Provider ID	A provider's unique identifier
Start Date	The trip's start date and time in UTC, ISO-8601 format, example: "2014-04-01T08:33:35.000Z"
Start WD	The weekday of the trip's start in UTC (1 = Monday, 2 = Tuesday, 3 = Wednesday, 4 = Thursday, 5 = Friday, 6 = Saturday, 7 = Sunday)
End Date	The trip's end date and time in UTC, ISO-8601 format, example: "2014-04-01T08:33:35.000Z"

End WDay	The weekday of the trip's end in UTC (1 = Monday, 2 = Tuesday, 3 = Wednesday, 4 = Thursday, 5 = Friday, 6 = Saturday, 7 = Sunday)
Start Location Lat	The latitude coordinates of the trip's start point in decimal degrees
Start Location Lon	The longitude coordinates of the trip's start point in decimal degrees
End Location Lat	The latitude coordinates of the trip's end point in decimal degrees
End Location Lon	The decimal degree longitude coordinates of the trip's end point in decimal degree
Geospatial Type	Type of trip, describes the trip's geospatial intersection with the requested zones (II - Internal to Internal; trips that starts and end within any requested zones; IE - Internal to External; trips that starts within any requested zone and end outside of any requested zone; EI - External to Internal; trip that start outside of any requested zone and ends within in any requested zone, and EE - External to External; trips that start and end but have one or more waypoints that intersect or completely traverse a requested zone)
Vehicle Weight Class	Numeral, representing the vehicle weight class

Table 4.2 trip details from raw data - fields description

Each trip is associated with a device unique identifier indicating the vehicle/mobile phone from which the trip was collected, as well as a provider unique identifier, which discloses the original source of the trajectory record. This allows to track the total number of trips collected from a specific vehicle/device.

A trip can be also distinguished according to its geospatial intersection with Napoli urban area, basically considering the location of its origin and destination; this information is contained in the field termed ‘Geospatial Type’. In light of this, four trip categories are defined:

- Internal to Internal (I-I): trips that starts and end within Napoli;
- Internal to External (I-E): trips that starts within Napoli and end outside of Napoli;

- External to Internal (E-I): trips that start outside of Napoli urban area and ends within the study area
- External to External (E-E): trips that start and end outside the study area but have one or more waypoints that intersect or completely traverse Napoli urban area.

Figure 4.2 presents a simplified schematization of the four possible trip geospatial types illustrating Napoli urban area divided in 10 zones corresponding to the 10 municipalities (administrative zoning of the city of Napoli):

- Chiaia, Posillipo, San Ferdinando (1);
- Avvocata,Montecalvario, Mercato,Pendino, Porto, S. Giuseppe (2);
- Stella, San Carlo all'Arena (3);
- San Lorenzo, Vicaria, Poggioreale, Zona Industriale (4);
- Arenella, Vomero (5);
- Ponticelli, Barra, S. Giovanni a Teduccio (6);
- Miano, Secondigliano, S. Pietro a Patierno (7)
- Piscinola, Marianella, Chiaiano, Scampia (8);
- Soccavo, Pianura (9);
- Bagnoli, Fuorigrotta (10);

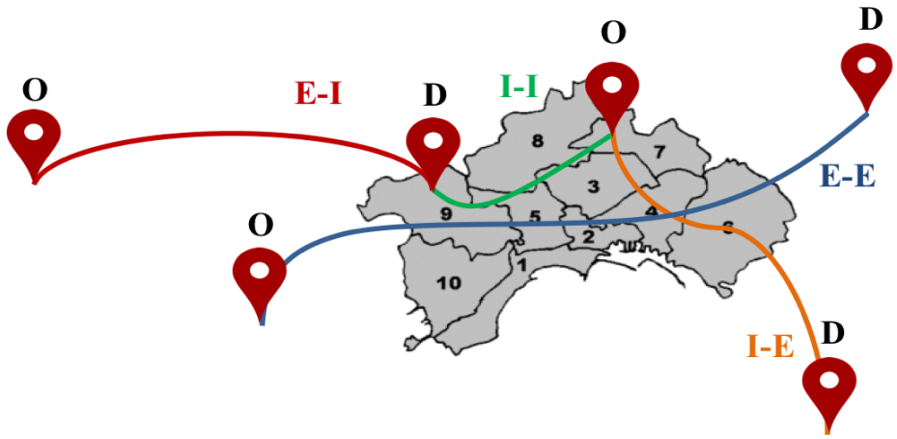


Figure 4.2 geospatial type categories: Internal to Internal (I-I); Internal to External (I-E); External to Internal (E-I); External to External (E-E); zoning:10 municipalities (administrative zoning of Napoli city)

The majority of origins and destinations from respectively E-I and I-E trips are distributed all over the Italian peninsula and only few origins/destinations fall outside Italian boarders, probably corresponding to freight distribution trips. An example showing this aspect can be observed in Figure 4.3 depicting the origins of all the sampled trips.

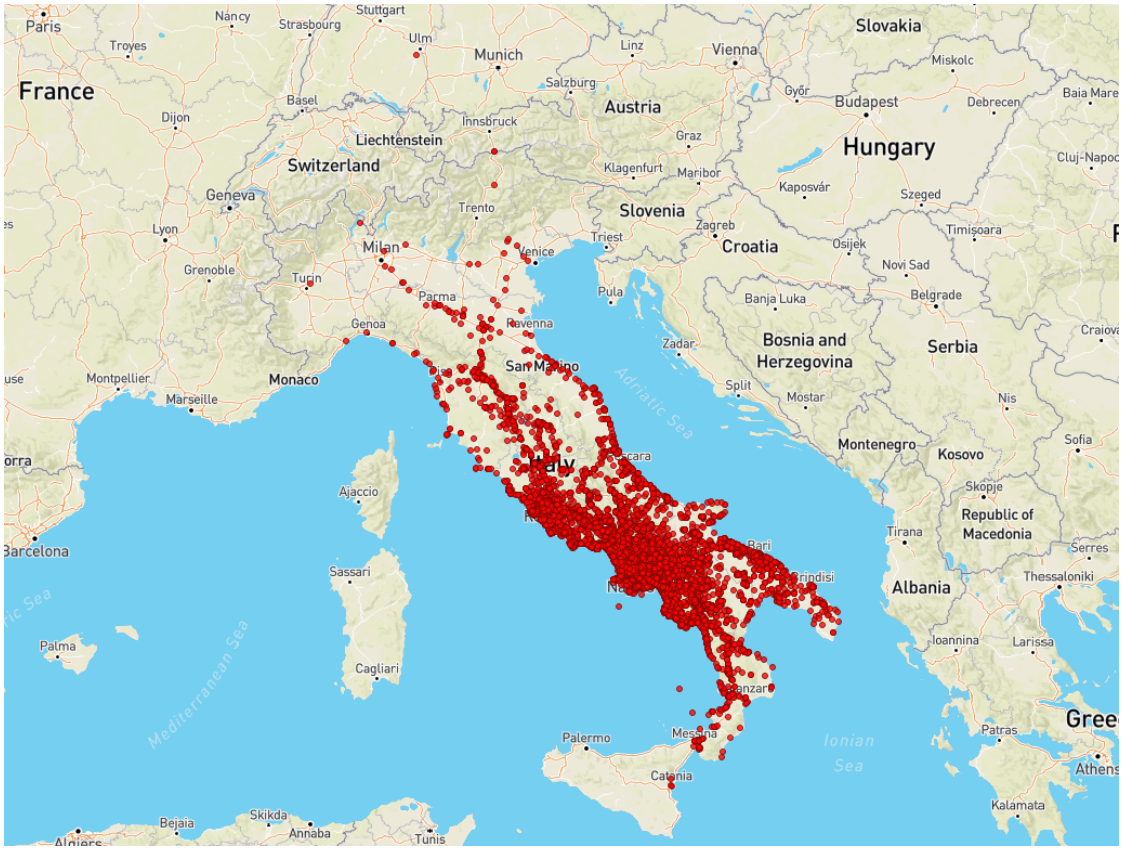


Figure 4.3 trip origins from raw trajectory data

Collected trips derive from miscellaneous sources, private and fleet vehicles, thus a field of file II contains the vehicle weight class. In light of this, tracked vehicles are categorised in three different classes according to their weight while the specifics on probe source type are included in the information reported in file III. The vehicle weight classes are defined as follow:

- Light Weight Vehicles (LWV): vehicles with weight up to 6.35 ton
- Medium Weight Vehicles (MWV): vehicles with weight from 6,35 to 11,8 ton
- Heavy Weight Vehicles (HWV): vehicles with weight over 11,8 ton

This distinction is essential to identify private vehicles (LWV such as personal cars, scooters, etc.) and vehicles dedicated to freight transport (MWV and HWV such as trucks, lorries etc.): this is a relevant information especially for o-d matrix derivation purpose and for the other scopes of the analysis as well.

File III reports the details of the multiple providers (i.e. consumers/users, fleet) from which the trajectory data was collected in the first place. Additionally, this information allows to identify the type of deployed sensor (*Probe Source Type*).

Data field	Description
Provider Id	A provider's unique identifier
Provider Type	Describes the provider type: 1 = Consumer, 2 = Fleet
Provider Driving Profile	Driving class, additional detail about type of provider: 1 = Consumer Vehicles, 2 = Taxi/shuttle/town car services, 3 = Field Service/Local Delivery Fleets, 4 = For hire/private trucking fleets
Vehicle Weight Class	Lists one of three weight classes provider: 1 = Light Duty Truck/Passenger Vehicle: Ranges from 0 to 14,000 lb; 2 = Medium Duty Trucks / Vans: ranges from 14001–26000 lb; 3 = Heavy Duty Trucks: > 26000 lb.
Probe Source Type	Type of sensor (i.e. mobile device or embedded GPS)

Table 4.3 provider details from raw data – fields description

4.2 Data cleaning and descriptive statistics

To develop descriptive statistics and to derive o-d matrices from the raw dataset, some preliminary data cleaning operations were required. To handle the great amount of data and speed up the data reading process, waypoints data from file I necessitated to be divided into smaller .csv files. Specifically, waypoints data was divided into 31 csv files, one per each day of observation. Secondly, data from the different input files (I-II-III, see section 4.1) had to be consistently joined into a unique database. To this end, the data cleaning process started with importing raw data into Matlab as a 'tabular text datastore', a Matlab object to manage large collections of text files containing column-oriented or tabular data. Identifying the trip ID as the key field, a unique table was created allowing for descriptive statistical analysis.

Since the vehicle weight class is particularly relevant for mobility analysis, specifically for o-d matrices estimation, a first investigation concerned the composition of the overall volume of trips according to this parameter. The analysis revealed that trips related to LWV were more the 71.1% of the total number of collected trajectories, while 28.1 % of the trips were from MWV and only 0.8% were from HWV. Therefore, the majority of the trips collected derives

from private vehicles, as shown in Figure 4.4. An analogous investigation has been developed considering the trip geospatial type (Figure 4.5), and subsequently focused on to trips detected from light weight vehicles, yielding to the following results: 46.7% of the trips were starting and ending in Naples urban area (I-I trips), around 21% were I-E and E-I and the remaining were E-E (see Figure 4.6) This suggests that the dataset, containing a high percentage of intra-urban trips (I-I) detected from light weight vehicles, results adequate to conduct the analysis of the applicability of the quasi-dynamic assumption in the case of urban context and further considerations on urban mobility.

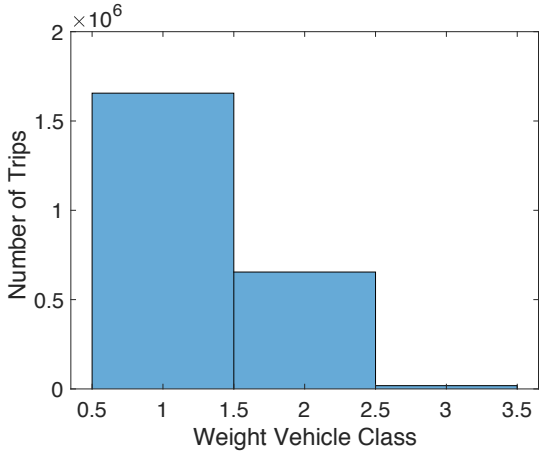


Figure 4.4 Trips per weight vehicle class: 78.1% LWV, 21.1% MWV, 0.8% HWV

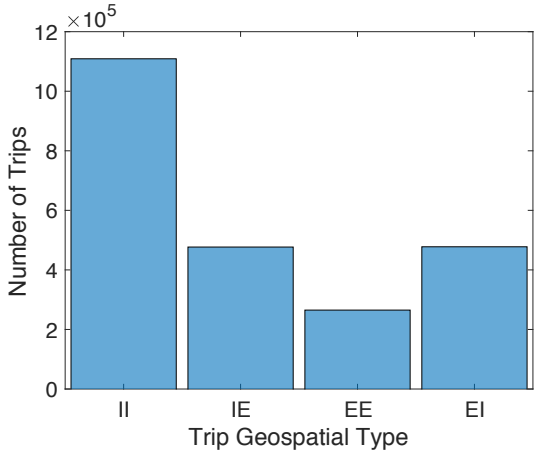


Figure 4.5 Trips per geo-spatial type : 46.7% I-I, 21% I-E & E-I, 32.3% E-E

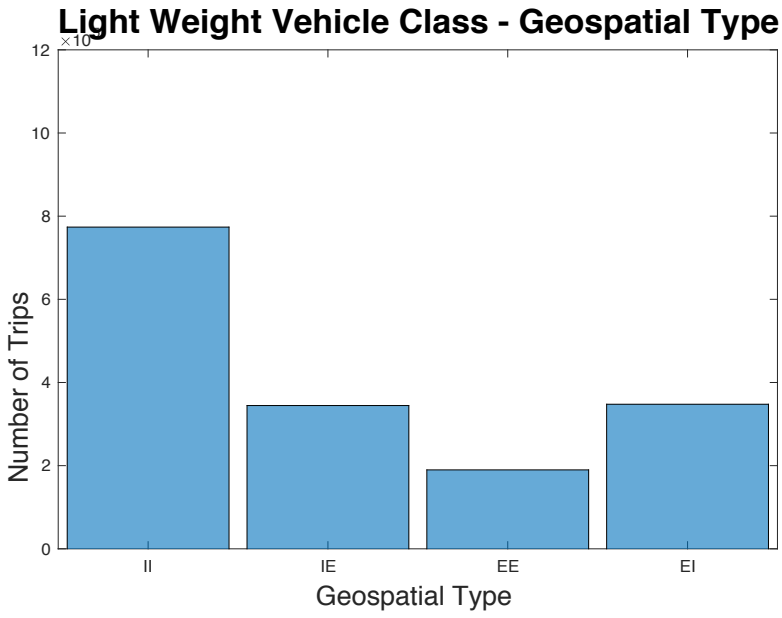


Figure 4.6 Light Weight Vehicle Trips per geo-spatial type

Basic waypoints statistics revealed insights on the spatial temporal distribution of recorded trips:

- *Average Number of Detected Trips per day*: 75.108 (Figure 4.7); this suggests an uniform distribution of the collected trajectories spanning the 31 days.
- *Average Number of Detected Points*: *1.643.009*. Figure 4.8 shows per each day reported on x-axes the average number of detected points among the total number of detected trips, then, averaging these quantities on the observation period, the overall average number of detected points has been obtained.
- *Average Polling Frequency*: 51,2 s (Figure 4.9), obtained with the same procedure as specified in the previous . Specifically, more than 70% of the trajectories has a polling frequency equal to 60 seconds. The higher is the polling frequency, the better can perform the map-matching process.
- *Average Distance between two GPS Points*: 355,9 m (Figure 4.10); this is particularly useful to get insights on the potential path choice inference, indeed the smaller is the distance between two GPS points, the higher is the possibility of reconstructing the real route and the higher is the potential route choice probability accuracy.

The average values reported above are all denoted in the following figures by the green dotted line.

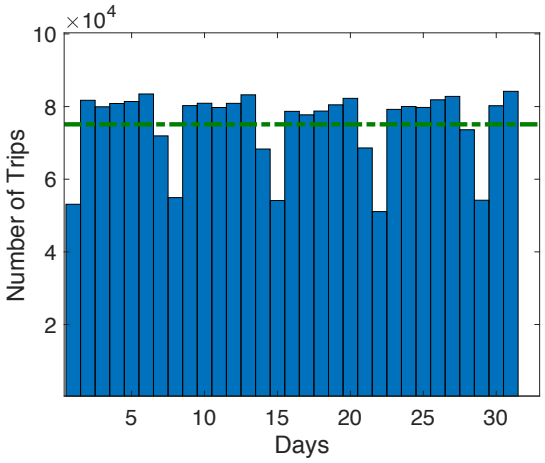


Figure 4.7 Daily Number of Detected Trips

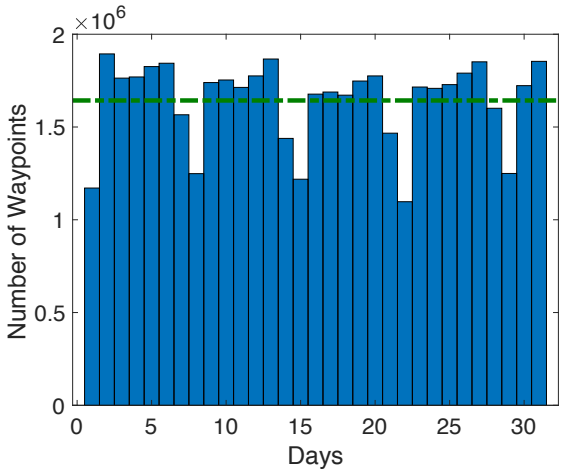


Figure 4.8 Daily Number of Detected Waypoints

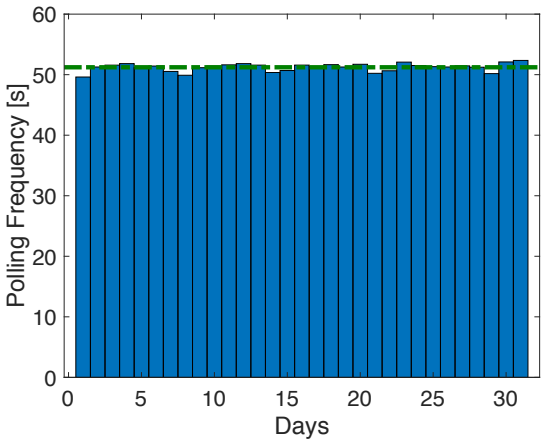


Figure 4.9 Mean GPS Polling Frequency on detected trips per day

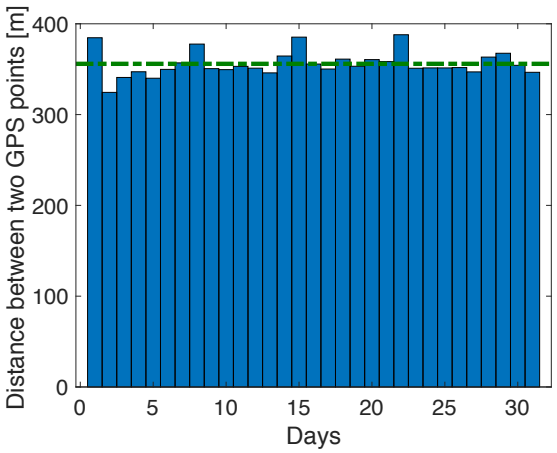


Figure 4.10 Mean distance between two GPS points on detected trips per day

Raw data was pre-processed by INRIX company anonymising and encrypting trip data to protect users' identity and sensitive information. During this operation, the company reassigned a unique ID to each tagged vehicle (see field "Device ID" in Table 4.2), allowing to consistently keep the observation of multiple trips deriving from the same vehicle over multiple days and reconstruct the total number of the trips collected from each vehicle for the entire period of observation. In light of this, it has been possible to define the distribution of the number of trips per each tagged vehicle. Around 35% of the vehicles registered only one trip, 12% registered two trips, less than 5% registered three and four trips, thus 57% registered less than five trips and the remaining 43% registered more than four trips (Figure 4.11) indicating an adequate heterogeneity of the dataset.

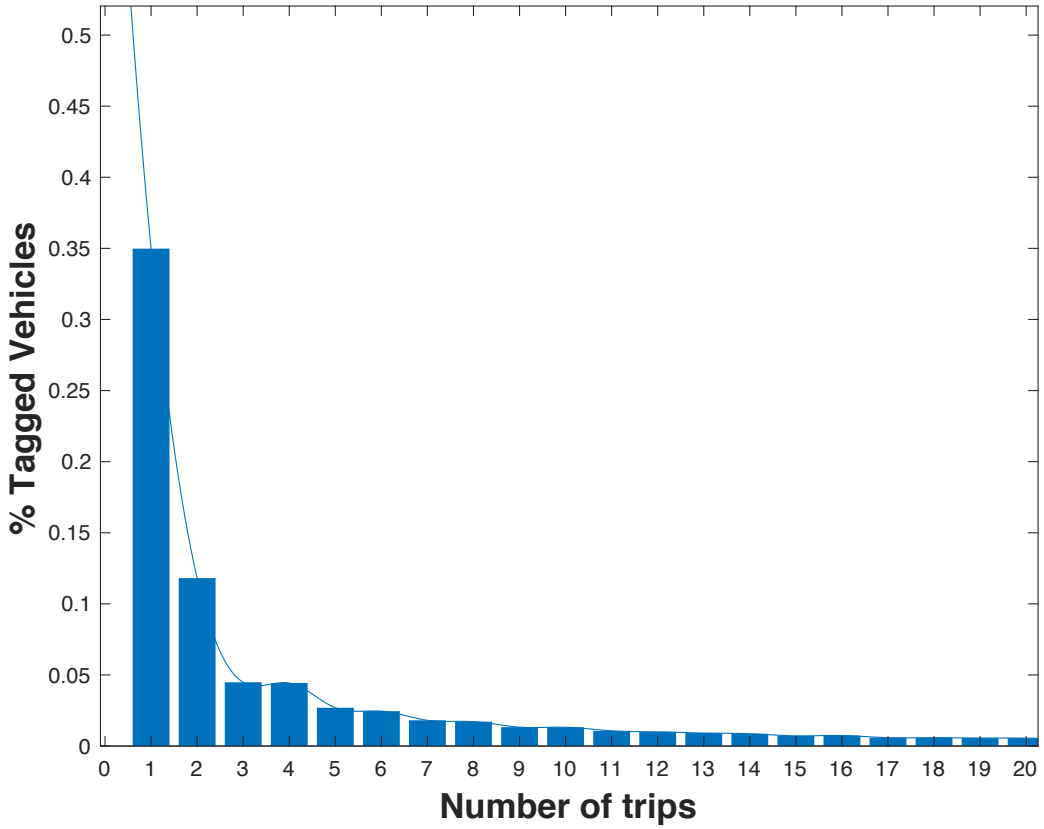


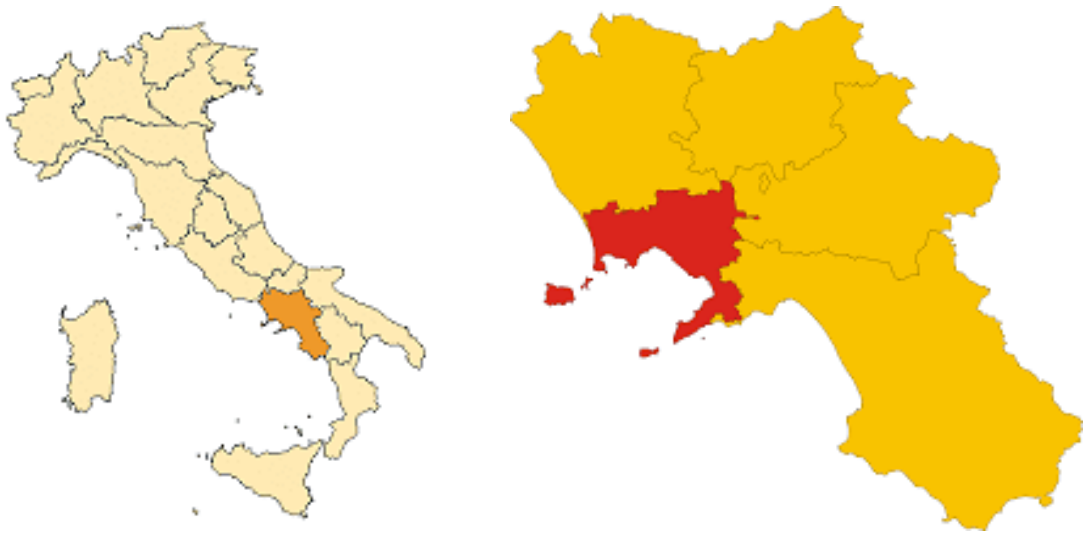
Figure 4.11 Trips per vehicle

4.3 Estimation of o-d matrices from trajectory data

The analysed trajectory dataset allows to derive a first crude estimation of time-dependent o-d matrices representing the collected trips taking place between two specific locations (traffic analysis zones) at a certain time of a day. This process requires defining the study area, a spatial discretization (zoning) of the study area and a temporal discretization (time-slice duration) of the analysed time-horizon (i.e. trajectory data observation period). Since the trajectory data analysis conducted in this study attempts to address diverse objectives, the o-d matrices have been referred to different geographic levels and multiple temporal discretization levels. To implement this process, a Matlab tool has been developed, able to select the trip list and derive the o-d matrices given one combination of these inputs, which can take the following values as shown by the Figure 4.11 (lines 44-48):

- *Study Area*: the considered study areas are referred to the territory of Campania region or alternatively to Napoli city, which is the capital of Campania region. The region of Campania is located in the southern part of the Italian peninsula, whose area is depicted in orange in Figure 4.12 (a). Since the area of interest covered by the trajectory dataset

is Napoli city, for the analysis concerning the demand evolution in the urban context, the observed o-d matrices have been also referred to its urban area whose boarders are illustrated in Figure 4.12 (b). In light of this, since the trajectory dataset contained waypoints spreading inside and outside the Italian borders, a preliminary operation to implement the tool consisted in filtering out of the total number of recorded waypoints a subset of waypoints falling inside the Campania region: this selection was performed using TransCAD software and the resulting data was subsequently imported into Matlab.



(a) *(b)*

Figure 4.12 Study Areas: (a) Campania region; (b) Napoli city urban area.

- *Zoning*: Cities (applicable for Campania Region only, referring to the 550 cities of Campania region registered in 2017), Municipalities (applicable for Napoli City only, referring to the 10 municipalities of Napoli city and macro-tract census zones for a higher level of disaggregation (30 zones used only for Napoli city but applicable for both territories); The origin and the destination zones of each trajectory have been identified selecting the areas where first and the last waypoints of the recorded sequence referring to the same trip ID.

Campania Region

(a)

Napoli City

(b)

(c)

Figure 4.13 Reference Zoning systems: (a) Cities of Campania region (550); (b) ASC of Napoli city (30); (c) Municipalities of Napoli city (10)

- *Time-slice duration*: the o-d matrices can refer to any duration ranging from 15 minutes to 1440 minutes (o-d matrix corresponding to the entire day): the recommended step in this range is 15 minutes;

Additionally, the tool architecture allows to build o-d matrices referred to the whole trajectory data observation period (31 days), and classified according to a given day type (working day, before holiday, holiday) and eventually to a specific weight vehicle class (Low, Medium, Heavy), as shown in Figure 4.14.

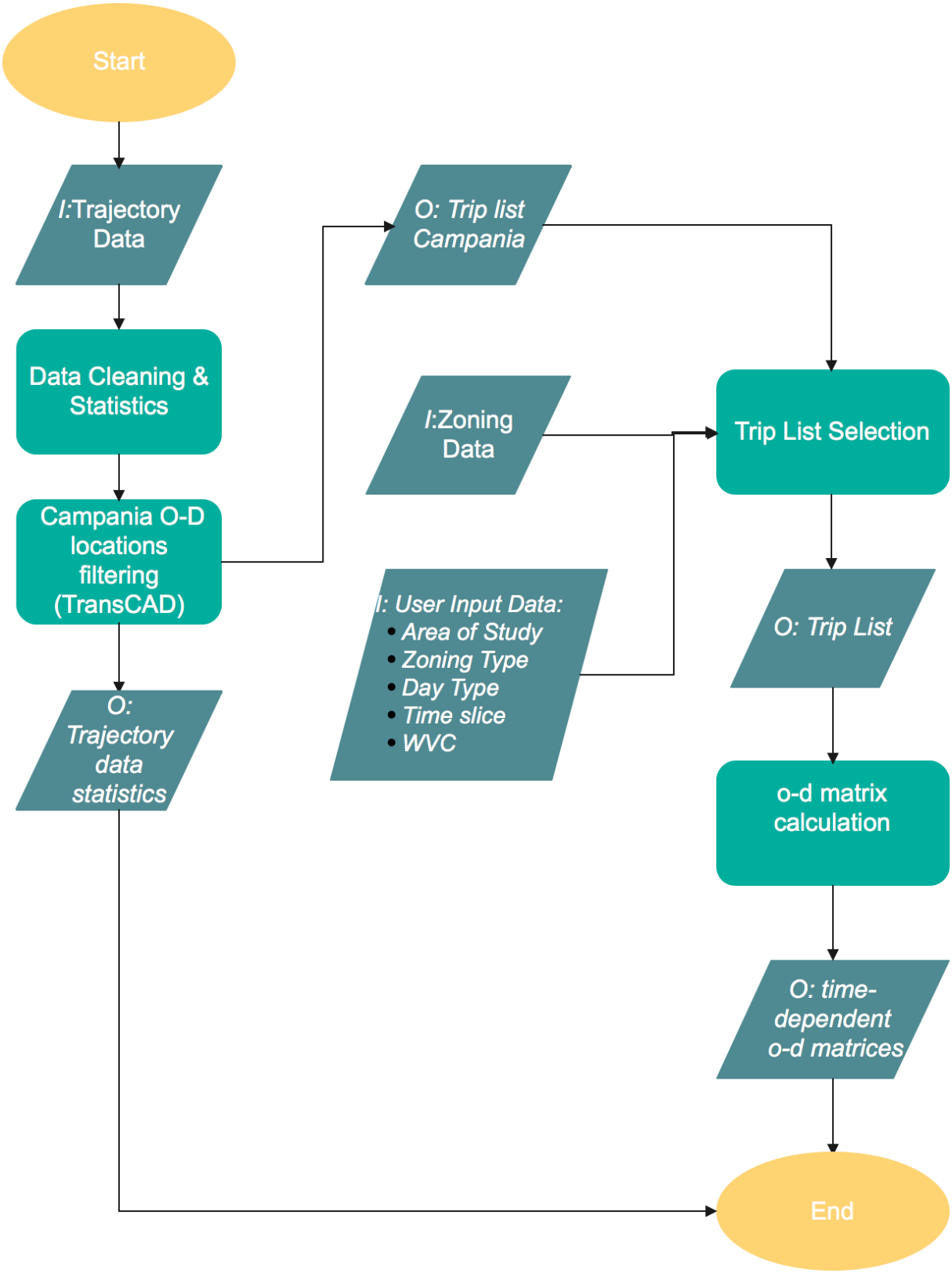


Figure 4.14 MatLab tool flowchart to obtain trajectory data statistics and classified o-d matrices by trajectory data

Once the input parameters have been set, the tool filters out from the list of all the trips collected in the trajectory sample the trips selected by the input criteria (Figure 4.14) and maps their origins and destinations to the chosen geographic zoning, generating the requested o-d matrices. The magnitude of the resulting o-d flows depends upon the granularity of the chosen spatial and temporal discretization: by way of example, Figure 4.15 reports the hourly-based demand profile of one o-d pair in Napoli urban area over a typical working day, using municipalities as reference zoning system, as depicted in Figure. In this case the maximum registered value among o-d flows is around 30 vehicles in the time interval from 4 pm to 5 pm. Additionally, the example shows that municipality zoning and 1 hour time-slice constitutes an adequate spatial and temporal discretization to avoid matrix sparsity, a well-known problem of o-d matrices derived from trajectory data samples.

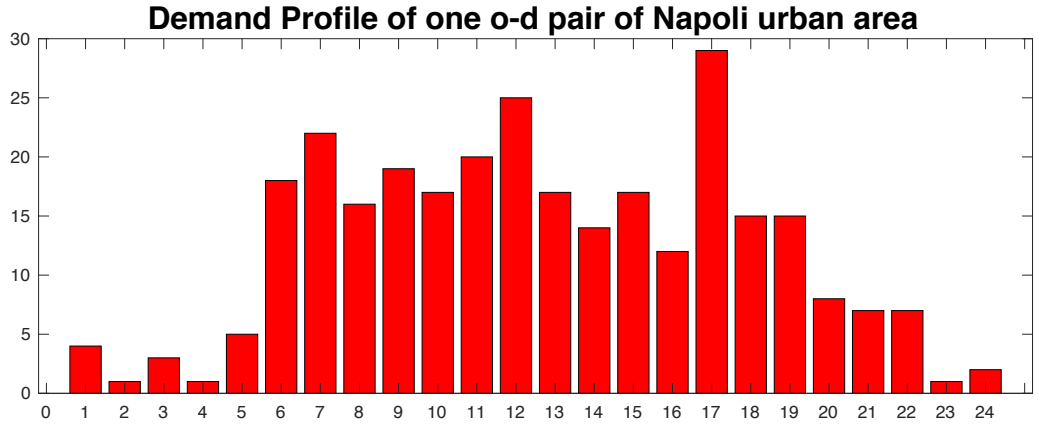


Figure 4.15 hourly-based demand profile of one o-d pair in Napoli urban area spatially discretised in Municipalities zones as in Figure 4.13 (c);

To gain more insights on o-d flows daily patterns, daily o-d flows of Napoli urban area have been categorised and analysed to quantify the variability among o-d flows resulting in the same day-type (working days, before holidays and holidays). To this end, only light vehicle trips were filtered and mapped with an acceptable granularity to the 10 municipalities of Naples city (Figure 4.13 (c)). Additionally, outbound and inbound demand flows deriving from the exchange with the remaining territory of Campania region and crossing the city area have been considered into the analysis, defining an 11x11 o-d matrix for each day, by referring the eleventh o-d pair to the exchange flows with external areas.

Subsequently, the o-d matrices were grouped according to the day-type (Figure 4.14). To perform the variability assessment for each day-type group, each daily o-d matrix has been compared to an o-d matrix containing group mean values and evaluating the coefficient of variation of the root mean square error (cvRMSE), a very well-known indicator already introduced in Section 3.1.3 (Equation 3.5) Figure 4.16 and Figure 4.17 show that the variability

of daily o-d flows is significant for both working days and before holydays, while for holidays results quite stable (Figure 4.18).

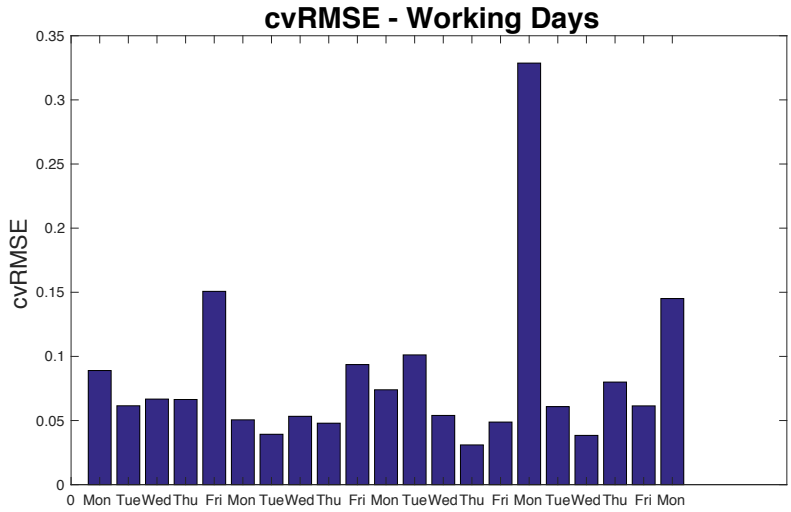


Figure 4.16 working days o-d matrices variation with the respect to group mean values

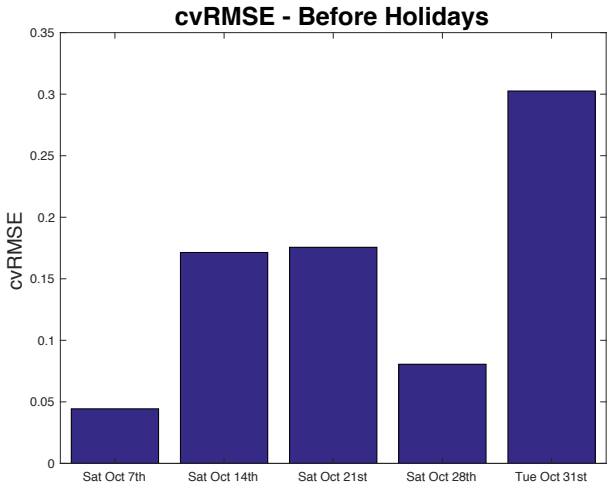


Figure 4.17 before holidays o-d matrices variation with the respect to group mean values

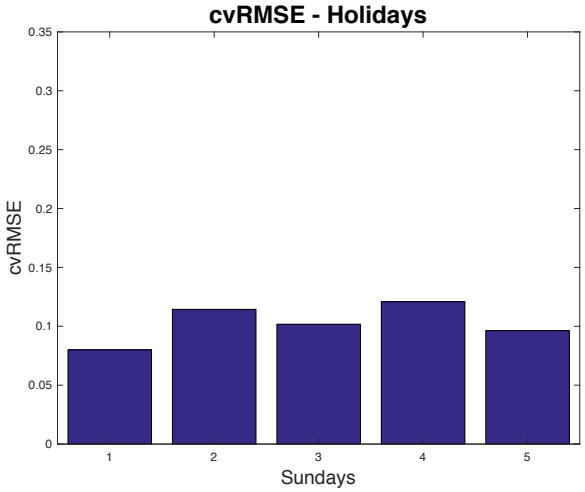


Figure 4.18 holidays o-d matrices variation with the respect to group mean value

4.4 Penetration rate estimation

4.4.1 Methodology

As previously mentioned, trajectory data represents only a subset of vehicles on the road, thus it is important to define its representativeness of the true underlying phenomenon. To this end, it is necessary to estimate the penetration rate of the collected sample. Furthermore, understanding the distribution of the penetration rate among origins, destinations and o-d pairs becomes essential to develop ad-hoc scaling techniques and properly expand the trajectory sample. Since the true o-d flows are unknown, the estimation of the penetration level and its distribution is not straightforward, thus only an inference based on census data (e.g. population, workforce, employees by traffic analysis zone, commuter trips) can be performed. While, when census data is unavailable, the analysis can be performed using as benchmark data a reliable historical o-d matrix, usually provided from previous studies. When commuter trips data and/or population data are available, a first crude estimation of the overall penetration rate can be derived comparing the total number of trajectories detected in a sample ($\sum_{od} d_{od}^{traj}$) and the total number of trips reported in census data ($\sum_{od} d_{od}^{CENSUS}$):

$$\tilde{\rho} = \frac{\sum_{od} d_{od}^{traj}}{\sum_{od} d_{od}^{CENSUS}} \tag{4.1}$$

This operation is essential to develop a direct scaling technique as presented in section 2.2.1, especially when link counts data are not available. Indeed, the expansion rate reported in equation 2.46 to perform a direct scaling is basically given by the inverse of the overall penetration rate calculated in equation 4.1:

$$\tilde{\varepsilon} = \frac{1}{\tilde{\rho}} \tag{4.2}$$

Note that, to opportunely define the overall penetration rate it is necessary to guarantee consistency between o-d flows derived from trajectory data and census data o-d flows in terms of: set of covered o-d pairs, scope of the trip, transport mode and reference period.

To complete the analysis and evaluate the penetration rate distribution among origins, destinations and o-d pairs, it has been defined:

- A penetration rate depending on the origin zone, assuming that the o-d pair whose origin is the same will share the same penetration rate:

$$\omega_o = \frac{\sum_d d_{od}^{traj}}{\sum_d d_{od}^{CENSUS}} \cdot 100 \ \forall \ o \ \in O \tag{4.3}$$

- A penetration rate depending on the destination zone, assuming that the o-d pair whose destination is the same will share the same penetration rate:

$$\varphi_d = \frac{\sum_o d_{od}^{traj}}{\sum_o d_{od}^{CENSUS}} \cdot 100 \ \forall \ d \ \in D \tag{4.4}$$

- A penetration rate depending on the o-d pair, basically assuming a different penetration rate per each o-d pair:

$$\sigma_{od} = \frac{d_{od}^{traj}}{d_{od}^{CENSUS}} \ \forall \ od \ \in I \tag{4.5}$$

Wherein $\sum_d d_{od}^{traj}$ is the total demand flow originated from origin *o* and $\sum_o d_{od}^{traj,wd}$ is the total demand flow reaching the destination *d;*

Note that, since trajectory data total coverage is not guaranteed per each origin, destination or o-d pair, the non-zero values can be calculated only for the origins/destinations/o-d pairs resulting from the intersection of the sampled o-d pairs in the trajectory dataset and the covered o-d pairs from commuting survey.

4.4.2 **Results**

The census data used in this part of the thesis refers to the last available commuting census referring to the 15th population census carried out in October 2011 by the Italian Institute for statistics whose acronym is ISTAT, standing for "Istituto Nazionale di Statistica" (literally: National Institute for Statistics). Specifically, this survey is periodically conducted every 10 years and its aim is to register the internal flows of any Italian city and the inbound and the outbound flows from any Italian city to any other city in the country. The recorded o-d flows refer to 28.871.447 individuals moving to go to the work/study place during the morning peak hours (approximately 7.15 – 9.15 AM) of a typical working day (usually Wednesday). Besides

the o-d pair, the commuting trips are classified by the scope of the trip, the type of residency (private or co-living households) and the travellers' gender. 80% out of the total number of records contained in the survey are furtherly classified by the mode of transport, the time range of departure and the approximated travel time. Therefore, only the latter set of records specifying the mode of transport can be used for the scope of the analysis. A preliminary operation to obtain a first estimate of the overall penetration rate ($\tilde{\rho}$) is to guarantee a consistent comparison between the two terms in equation 4.1 according to the set of covered o-d pairs, the scope of the trip, the transport mode and the reference period.

To obtain the terms in equation 4.1, a preliminary step consisted of deriving the o-d matrices relative to the morning peak hours (7.15 AM – 9.15 AM) of the working days collected in the trajectory dataset and the commuting o-d matrix derived from the trips reported by ISTAT by means of the Matlab tool structured as in Figure 4.14 and described in Section 4.3, using cities as reference zoning system (*Figure* 4.13). Subsequently, since the trajectory dataset refers to Napoli urban area (see Section 4.1), a consistent set of covered o-d pairs to compare trajectory data to census data was obtained selecting out from ISTAT records the internal trips of Napoli city, the inbound and the outbound flows from/to each city of the provinces of Campania region to/from Napoli city. In other words, a consistent set of o-d pairs for both datasets refers to trips inside the Campania region and having as origin and/or destination Napoli city area.

To demonstrate that the comparison between the two sets of data referring to two different years (2017 for trajectories and 2011 for census data) can hold despite the significant temporal gap, an evaluation of the population variation registered from 2011 to 2017 for all the cities of Campania region has been carried out. As shown in Figure 4.19, negligible changes occurred to population census data between the two years, indeed, only in 35 out of 550 cities the variation (delta) corresponds to a value presenting an order of magnitude equal to 10^4. Therefore considerations on the results deriving from the comparison of trajectory data to 2011 census data can be considered reliable.

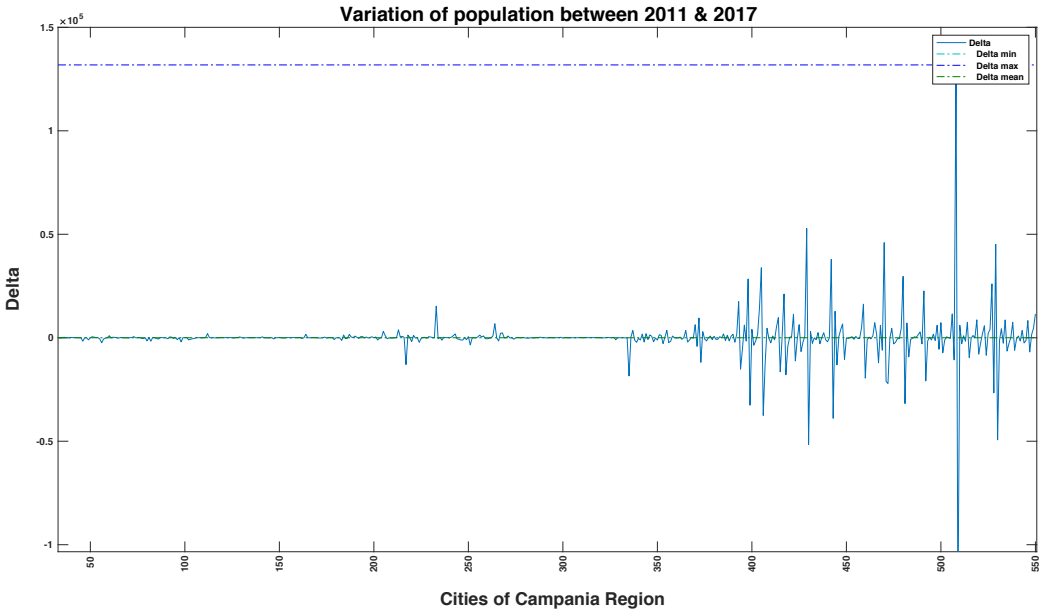


Figure 4.19 Variation of population between 2011 and 2017 of Campania region cities

A first observation concerns the number of non-zero o-d flows and the related set of reported o-d pairs by both datasets: interestingly, while the census dataset referring to the morning peak of a typical working day reports 700 o-d pairs (e.g. referred to non-zero o-d flows) out of the total 1099 o-d pairs related to inbound, outbound and internal trips of Napoli city, the trajectory dataset reports more than 1000 non-zero o-d flows referred to the same type of the day and time interval. This implies that a considerable percentage of the non-zero o-d flows derived by trajectory data refers to o-d pairs not reported in the census dataset. By way of example, the number of o-d pairs referring to non-zero o-d flows by trajectory and not collected in the census dataset has been reported in Table 4.4 for the four Wednesdays of the trajectory sample (respectively Oct 4th 2017, Oct 11th 2017, Oct 18th 2017 and Oct 25th 2017):

	Total number of non-zero flow o-d pairs by trajectory	***non-zero flow o-d pairs by trajectory not reported by census***	***% of o-d pairs collected in both trajectory and census***
Oct 4th 2017	1033	757	39%
Oct 11th 2017	1048	781	38%
Oct 18th 2017	1033	762	39%
Oct 25th 2017	1023	764	37%

Table 4.4 comparison of the number of o-d pairs referring to non-zero o-d flows by trajectory and o-d pairs not collected in the census dataset

Therefore, around more than 70% of the o-d pairs referring to non-zero o-d flows collected in the trajectory sample is not reported by census commuting data; similar values have been derived for all working days in the sample. The calculations on the trajectory sample penetration rate defined per origin, destination and o-d pair as respectively presented in Equations 4.3, 4.4 and 4.5 have been developed considering the set of o-d pairs covered by both datasets, thus the deriving considerations on the penetration rate estimates must account for this relevant aspect. Furthermore, it is worth of notice that although the two sets of data can be considered consistent, a further investigation should be conducted to estimate the scope of the trips detected from trajectories: as introduced in Section 1.2, one of the main drawbacks of trajectory data is its enormous lack of socio-economic information, which are hidden on purpose to protect users' privacy. Indeed, in this analysis it is reasonably assumed that the total number of trips happening between the morning peak hour are systematic, however this could only partially correspond to reality.

To visualise the area of interest used for the trajectory sample penetration rate estimation, a simplified map of Campania region in *Figure* 4.12 (b) highlights Napoli province area in red, while all other provinces of Campania region are depicted in yellow, which are arranged in clockwise order as: Caserta, Benevento, Avellino and Salerno. Therefore, a slightly smaller part of the red area corresponds to the sampling area of INRIX trajectories, which is Napoli urban area as depicted in Figure 4.1. Furthermore, for the scope of the analysis only the trips by private vehicles (e.g. car or motorbike), originating from/reaching any city of Campania region (*Figure* 4.13 (a)) and to/from Napoli city (as in Figure 4.1) were selected from ISTAT census data, while from the INRIX trajectories only LWV class trips collected during the morning peak of the 21 working days in Napoli province were selected.

Denoting with $d_{od}^{traj,wd}$ the total number of trips by trajectory detected during the morning peak of a typical working day (formally *wd*) and with $\sum_{od} d_{od}^{ISTAT}$ the total number of trips contained in the ISTAT commuting census corresponding to the same time interval, the penetration rate $\tilde{\rho}$ can be estimated applying equation 4.1 for each working day reported in the trajectory sample. Since (almost) all trips collected during the morning peak are systematic, it is possible to perform a direct comparison with ISTAT commuting census data. The values of the penetration rate obtained applying equation 4.1 referred to the morning peak of the twenty-one working days by trajectory are reported in the following table:

Working day	*Morning Peak penetration rate*
1	0.07
2	0.06
3	0.06
4	0.07
5	0.06
6	0.06
7	0.06
8	0.06
9	0.06
10	0.06
11	0.06
12	0.06

13	0.06
14	0.06
15	0.06
16	0.03
17	0.06
18	0.06
19	0.06
20	0.06
21	0.06

Table 4.5 values of the penetration rate obtained applying equation 4.1 referred to the morning peak of the twenty-one working days collected in the trajectory sample

An overall value of the estimated penetration rate can be derived as the mean value over the 21 working days of the sample:

$$E[\tilde{\rho}] = E\left[\frac{\sum_{od} d_{od}^{traj,wd}}{\sum_{od} d_{od}^{ISTAT}}\right] 100 = 6.1\% \quad (4.6)$$

Another calculation of the overall penetration rate can be performed comparing the total number of daily trips collected in the trajectory sample and an estimation of the total number of daily trips obtained from the total reported by ISTAT commuting census data. This calculation lead to a similar value of the overall penetration rate referred to the morning peak total demand reported in equation 4.6, assessed around 6%.

This is a satisfactory value compared to the range of variability of GPS trajectory sample penetration rate, indeed as reported by FHWA in 2016, the sampling rate generally ranges from small percentages up to 10% (see Chapter 2 for literature references). This result suggests that applying a scaling technique is a crucial operation to obtain a reliable estimate of the o-d matrix. In light of this, the direct scaling techniques presented in section 2.2.1 are primarily considered as an essential step to scale the volume of the captured trajectories such that a first reliable estimate of generated demand could be derived.

An ideal sampling technique guaranteeing an effective representativeness of the overall demand would imply an uniform distribution of the sampling rate among o-d pairs, e.g. using basic sampling techniques such as simple random sampling. This distribution can be obtained when building a purpose-oriented trajectory dataset in which the sampling process is well structured and its characteristics are defined a priori, conversely these conditions do not hold when dealing with non-purpose oriented trajectory data, which remarks the importance of this analysis. For this reason, to better understand the trajectory sample representativeness, a further analysis

pertains the penetration rate distribution among origins, destinations, and o-d pairs applying equations 4.2, 4.3, 4.4 respectively.

The analysis concerns the o-d matrices relative to the morning peak hours (7-9 AM) of the four Wednesdays ($d_{od}^{traj,wd}$), among all working days included in the trajectory dataset, in light of the fact that ISTAT commuting data are approximately referred to this time range and usually to Wednesdays. These matrices have been individually compared to commuting o-d flows reported by the ISTAT commuting survey (d_{od}^{ISTAT}). Since trajectory data total coverage is not guaranteed per each origin, destination or o-d pair, the sampling rates have been calculated only for the origins/destinations/o-d pairs resulting from the intersection of the available o-d pair in the trajectory dataset and the available o-d pair from ISTAT commuting survey, thus it is expected a different number of origins/destinations/o-d pairs covered per each analysed o-d matrix referring to a specific working day. To further evaluate the representativeness of generated demand observed from trajectory data the sampling rate per origin has been additionally calculated applying Equation 4.2 using as benchmark the population data collected during the census survey conducted in 2017; the obtained values are presented in Figure 4.20.

The population data has been filtered considering only people whose age is eligible for obtaining the driving license. By way of example Figure 4.20, Figure 4.21, Figure 4.22 and Figure 4.23 *illustrate the results referred to one of the working days of October 2017, while*

Table 4.6 reports relevant statistics on the measured values. Similar values have been obtained for all working days;

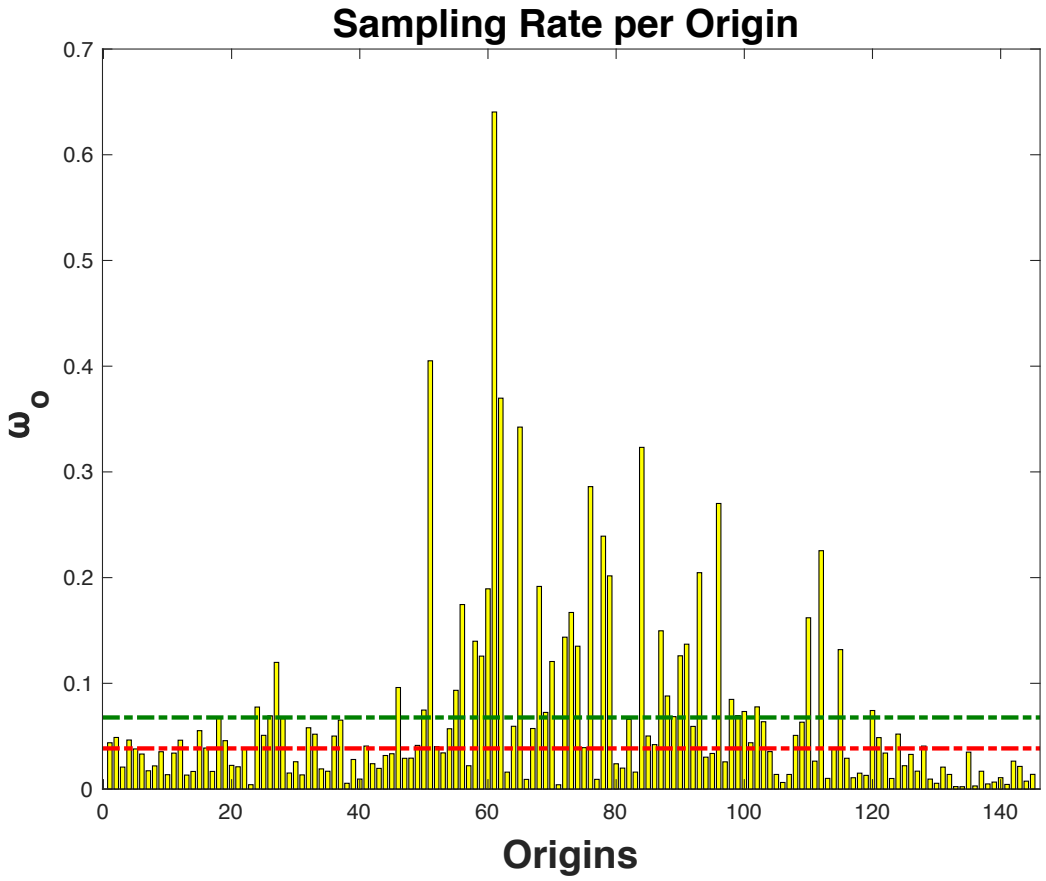


Figure 4.20 Sampling rate per Origin - values calculated comparing generated demand observed from trajectory data and population census data (collected in 2017) using as zoning system the cities of Campania region depicted in Figure 4.13 (a)

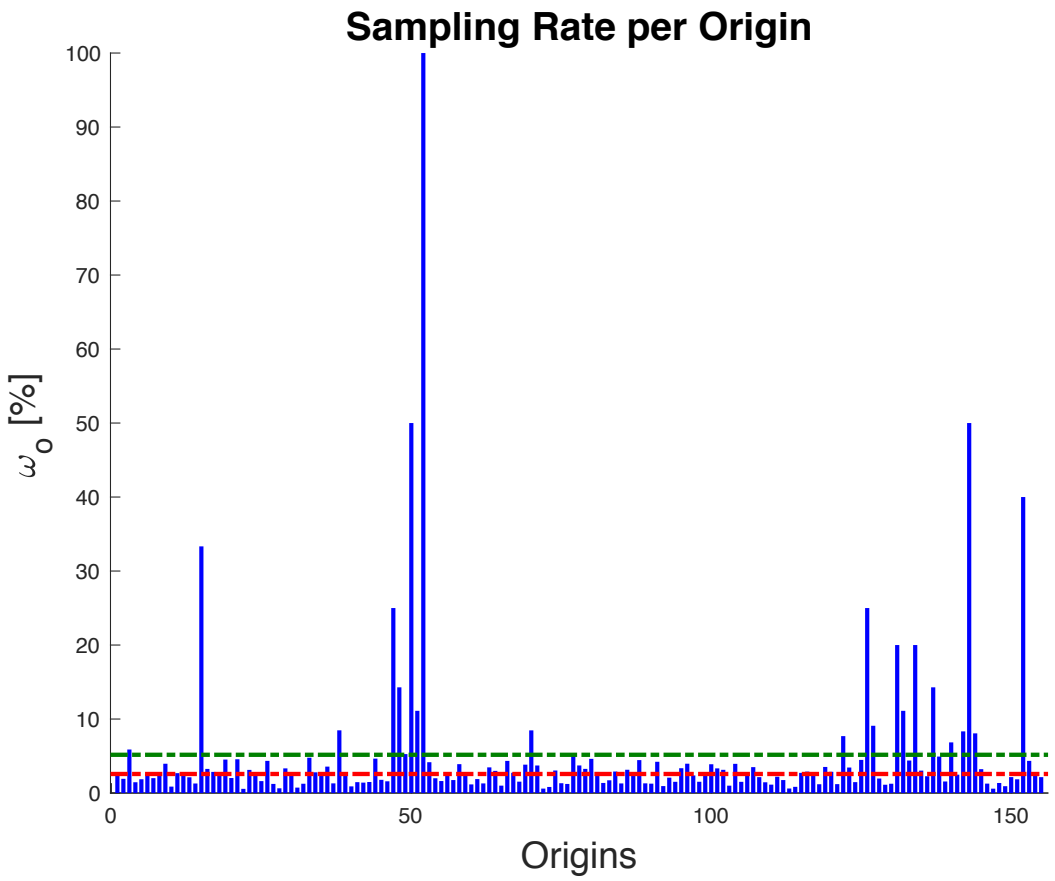


Figure 4.21 Sampling rate per origin zone obtained by comparing trajectory data to ISTAT commuting data collected between 7 and 8 AM of a typical working day, using as zoning system the cities of Campania region depicted in Figure 4.13 (a)

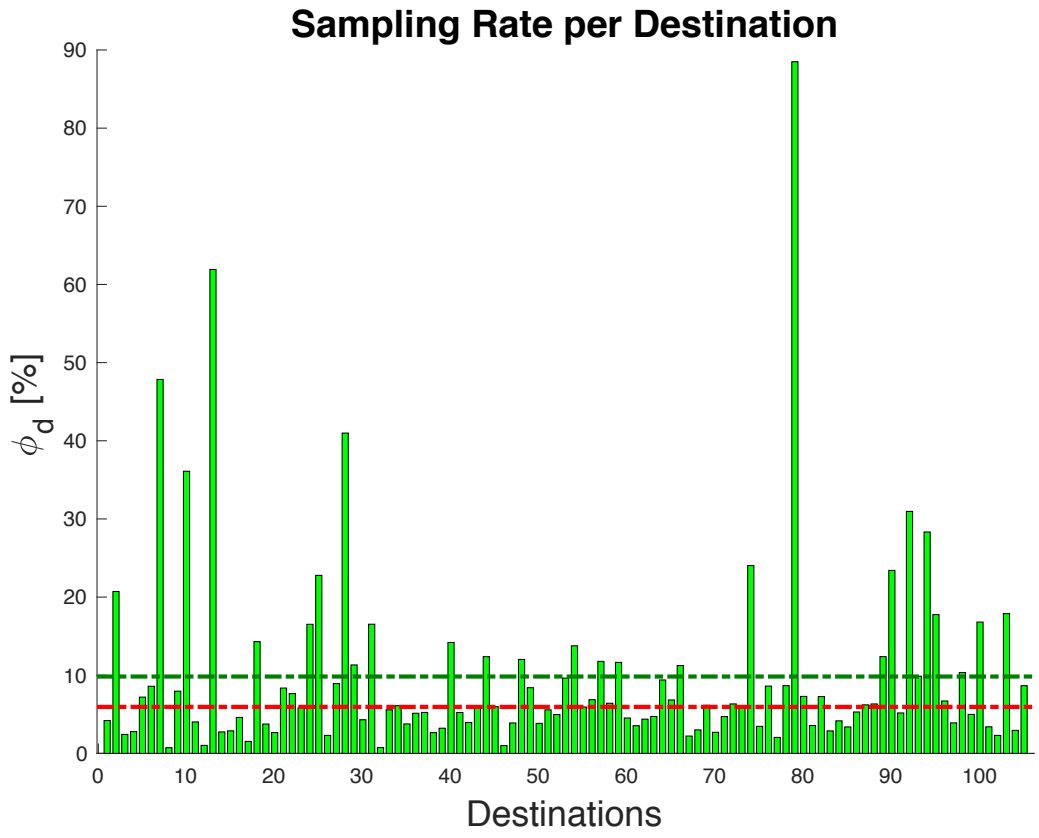


Figure 4.22 Sampling rate per destination zone obtained by comparing trajectory data to ISTAT commuting data collected between 7 and 8 AM of a typical working day, using as zoning system the cities of Campania region in Figure 4.13 (a).

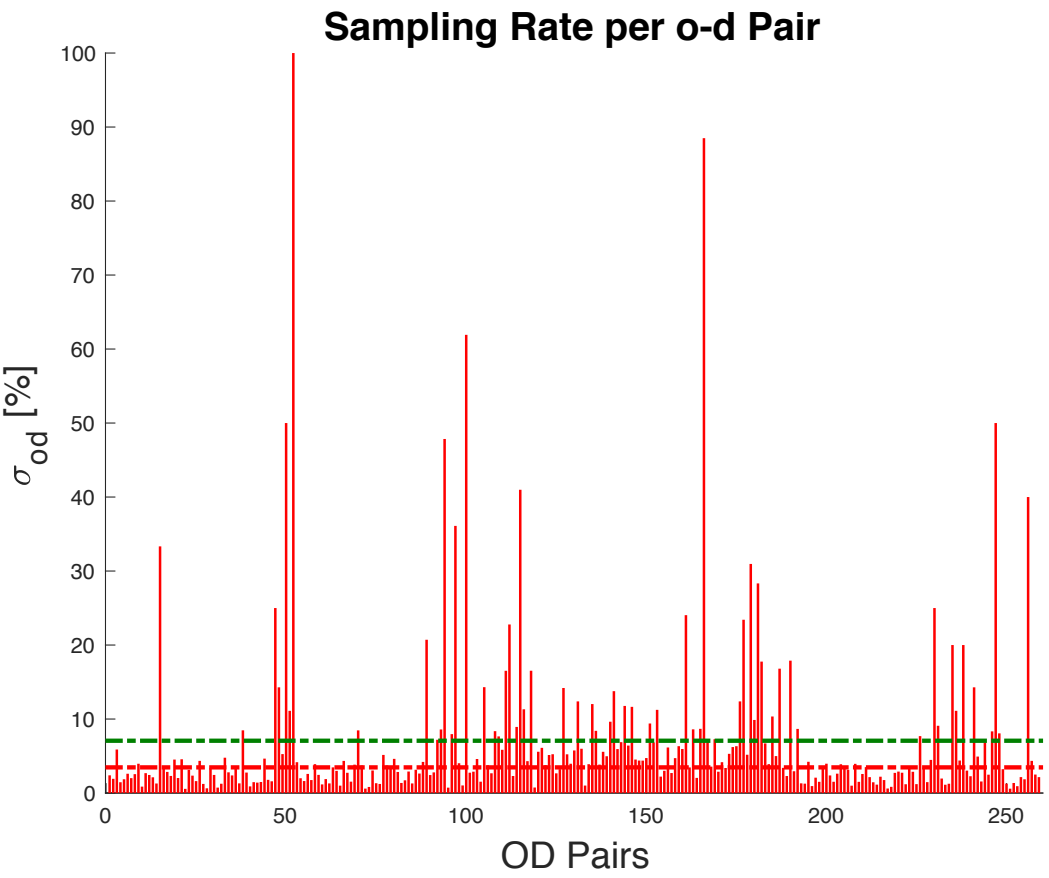


Figure 4.23 Sampling rate per o-d pair obtained by comparing trajectory data to ISTAT commuting data collected between 7 and 8 AM of a typical working day, using as zoning system the cities of Campania region depicted in *Figure 4.13* (a).

	MIN [%]	MEAN [%]	MEDIAN [%]
$\boldsymbol{\omega_o}$	0.56	5.17	2.77
$\boldsymbol{\phi_d}$	0.73	9.84	5.95
$\boldsymbol{\sigma_{od}}$	0.57	7.07	3.47

Table 4.6 relevant statistics on penetration rate defined per origin, destination and o-d pair

To evaluate the evolution and the characteristics of penetration rate defined by o-d pair at urban level, an analogous calculation has been referred to Napoli urban area using as zoning system the macro-census tracts defined by ISTAT. According to this zoning, 30 zones are defined as shown in *Figure* 4.13 (*b*), allowing to have an acceptable granularity and limited sparseness of o-d matrix for the scope of the analysis. The results have been provided for the entire period of observation (21 working days of October 2017). By way of example, only the results referred

to four Wednesdays of October 2017 have been reported in Figure 4.24, Figure 4.25, Figure 4.26 and Figure 4.27 respectively.

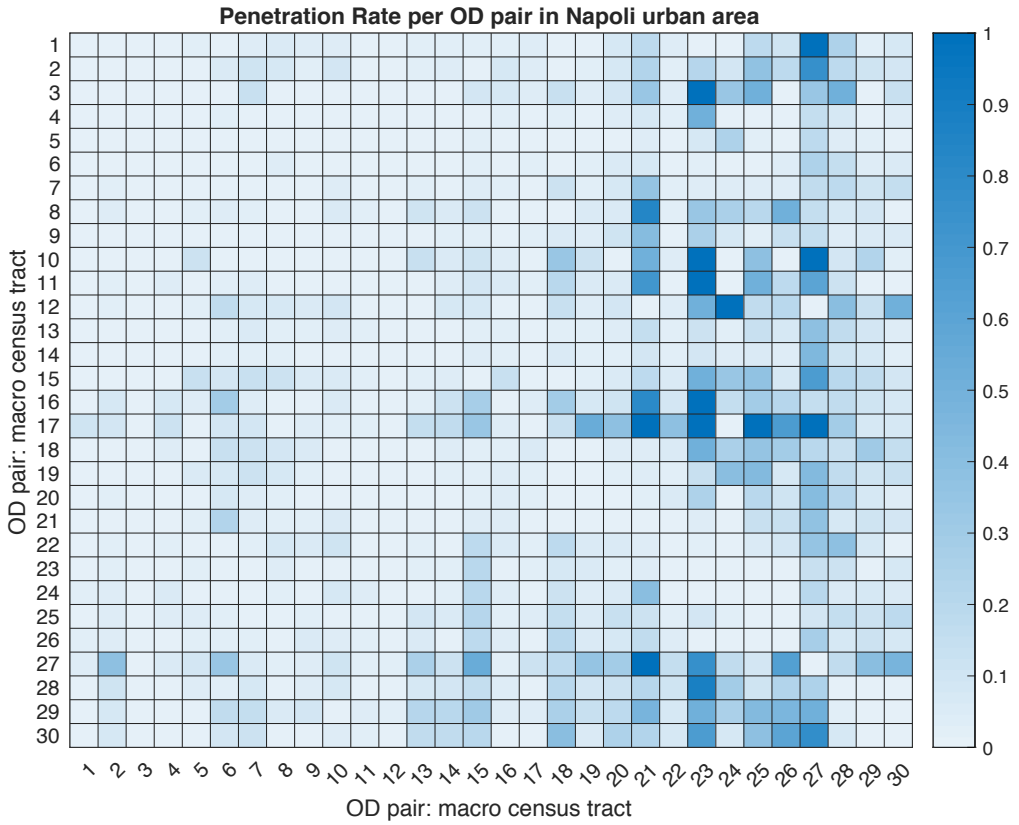


Figure 4.24 Heatmap depicting the level of penetration rate defined by o-d pair at urban level in Napoli city for a typical working day.

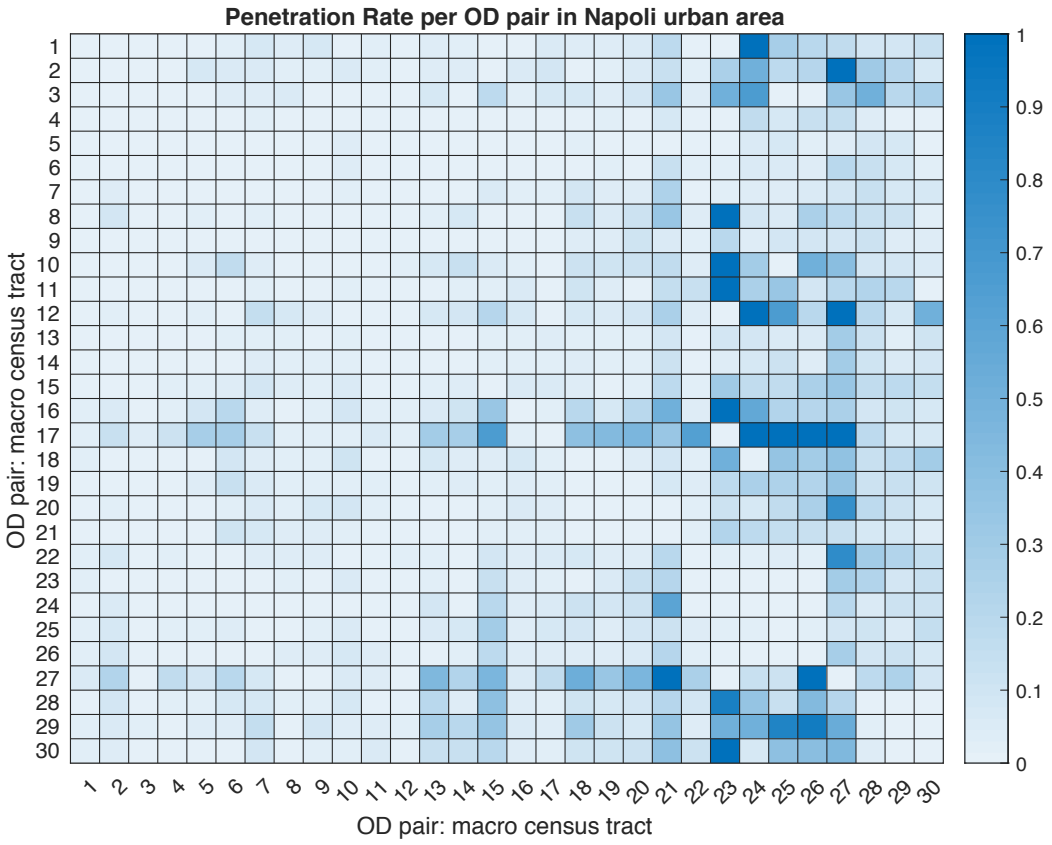


Figure 4.25 Heatmap depicting the level of penetration rate defined by o-d pair at urban level in Napoli city for a typical working day.

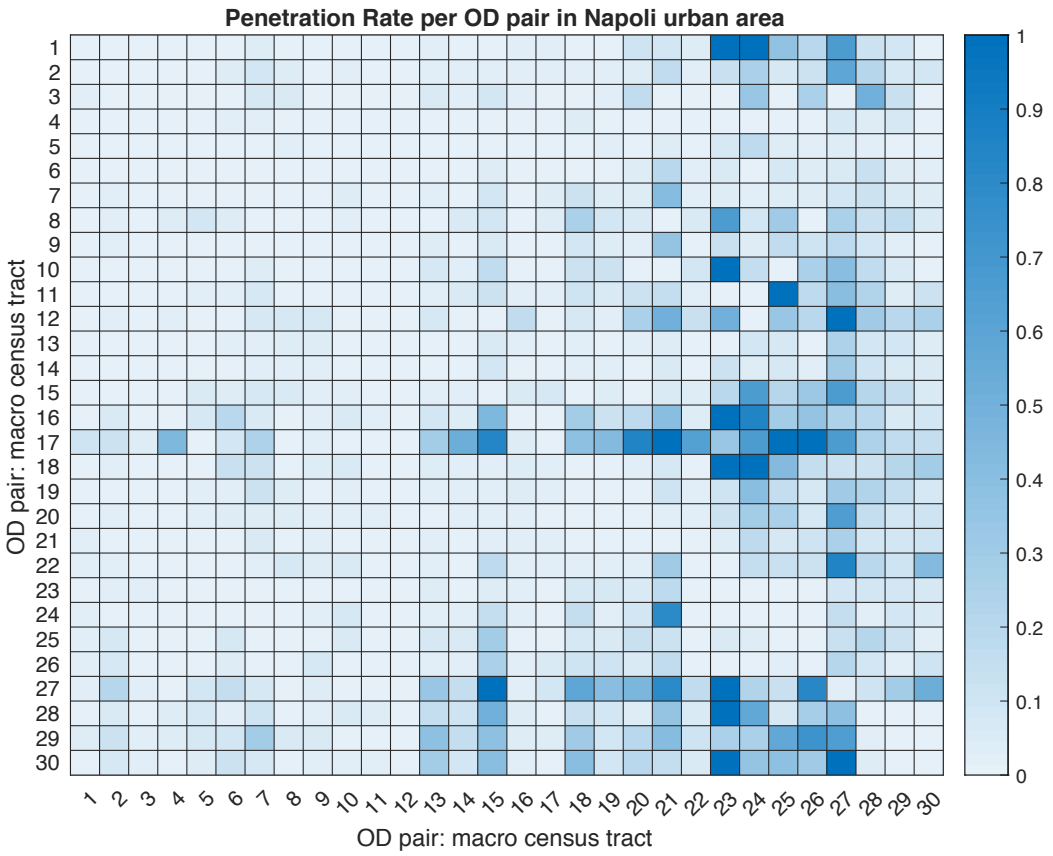


Figure 4.26 Heatmap depicting the level of penetration rate defined by o-d pair at urban level in Napoli city for a typical working day.

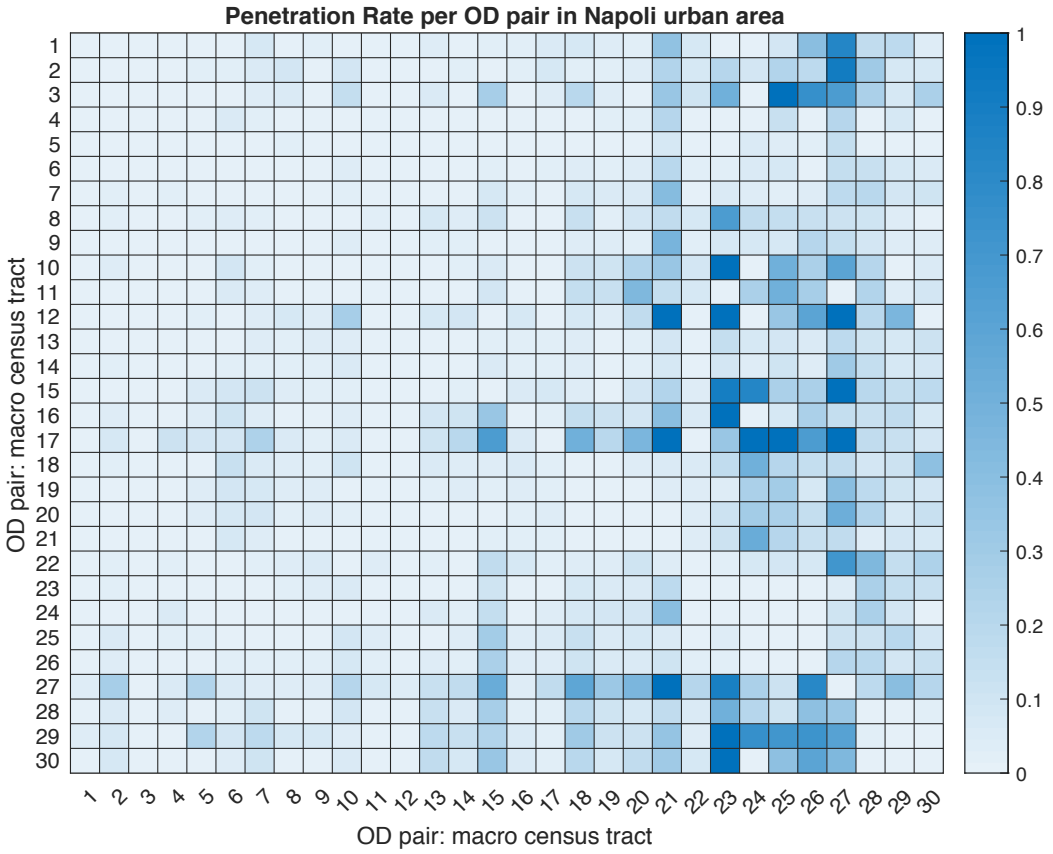


Figure 4.27 Heatmap depicting the level of penetration rate defined by o-d pair at urban level in Napoli city for a typical working day.

Evidence from results suggests a strong variability of the sampling rate defined by o-d pair for both regional and urban level, meaning that for the given sample, the penetration rate distribution is not uniform among o-d pairs, thus the probability of extracting a trip from the population (total number of actual trips) varies according to the considered o-d pair. Clearly, the evidence is similar in the other cases of sampling rate defined by origin zone and destination zone. Notably, for some of the sampled o-d pairs the sampling rate resulted equal to 100% (e.g. see o-d pair number 50 in figure 4.20), this is explained by the fact that some of the o-d flows reported both in the census dataset used as benchmark and in the trajectory dataset are equal to one. Clearly this case is not realistic, therefore these values should not drive the conclusions of this analysis.

Given these considerations, the experimental results have been utilized as a starting point to deepen the analysis on the implications of penetration rate distribution on demand flows accuracy when applying o-d flows estimation methods and inherent scaling techniques. This study, conducted by means of laboratory experiments, is presented in the following Chapters.

4.5 Experimental analysis of the quasi-dynamic assumption in urban context

4.5.1 Methodology

As introduced in Section 2.3.2.1, the quasi-dynamic assumption derives from the intuitive consideration that generation profiles evolution are more within-day dynamic varying compared to distribution shares evolution, basically these two set of variables follow different dynamics: the percentages of distributions vary across a longer period of time, while generation profiles are subject to rapid changes in time. In light of this, according to the quasi-dynamic hypothesis, the distribution profiles can be approximated to their average values and assumed constant over a longer period of variation with the respect to generation profiles reference period of variation. As introduced in equation (2.5), o-d flow demand variables can be expressed as the product of generation profiles and distribution shares: the approximation introduced applying the quasi-dynamic hypothesis yields to a new set of variables, namely the quasi-dynamic o-d flows ($d_{od}^{\theta,qd}$ as in equation 2.5):

$$d_{od}^{\theta} = g_o^{\theta} p_{d|o}^{\theta} \cong g_o^{\theta} p_{d|o}^{\tau(\theta)} = d_{od}^{\theta,qd} \tag{2.5}$$

However, although the quasi-dynamic hypothesis allows to dramatically reduce the number of variables in the o-d flows estimation problem, it introduces an inherent bias into the estimation process known as the 'Intrinsic Error', which is formally reported here for the sake of readability:

$$ie_{od}^{\theta} = d_{od}^{\theta} - d_{od}^{\theta,qd} \tag{2.8}$$

Basically, it represents the distance (error) between the quasi-dynamic o-d flows, computed considering constant distribution shares over a pre-determined period of time, namely the quasi-dynamic interval, and the original o-d flows. Intuitively, the longer is the assumed duration of the quasi-dynamic interval, the larger will be the intrinsic error. Cascetta et al. (2013) assess the magnitude and the variation of the intrinsic error for different duration of the quasi-dynamic interval by means of statistical tests such as chi-squared and likelihood ratio test in the case of a closed motorway system, thus evaluating its variability in uncongested networks with fixed path-choice. Since ground-truth data was available for the test site, the conducted analysis demonstrated that the quasi-dynamic o-d flows follow the same probability distribution of the real o-d flows, meaning that their statistic parameters (mean, variance) represent the same distribution. In this way, the study demonstrated that acceptable goodness-of-fit measures can be obtained even under the hypothesis of a 24h quasi-dynamic interval duration. Specifically the conducted study refers to a highway network, which is a closed system test site, not only implying that ground truth values were available but also that the path choice is fixed and the

congestion phenomena can be considered as negligible (indeed the study is conducted in the case of uncongested network).

In this section, the experimental analysis investigating the intrinsic error magnitude and its distribution has been extended to urban context. Such contexts are normally characterized by more complex o-d flow patterns; however, since a significant percentage of the overall urban mobility is represented by systematic trips, the quasi-dynamic assumption is expected to be sufficiently acceptable considering the peak periods in which it is expected that the majority of trips are due to commuting activities. The approach adopted in this thesis is based on an empirical analysis: the aim is to analyse the intrinsic quasi-dynamicity of the o-d matrices observed from trajectory data, which means evaluating the stability of distribution shares across a pre-specified quasi-dynamic interval, basically quantifying the error between quasi-dynamic trajectory o-d flows and trajectory o-d flows. Two well-known goodness-of-fit measures have been used to assess the quasi-dynamicity of trajectory o-d flows: the coefficient of determination (r^2) expressing the linear fitting between trajectory o-d flows and quasi-dynamic o-d flows (equation 3.5) and the cvRMSE already introduced in equation (3.5). The empirical analysis is propaedeutic to set the basis of synthetic experiments conducted with the scope of analyzing the performance of trajectory data scaling techniques and other o-d flow estimation methods, as described in the Chapters 5 and 7.

4.5.2 **Results**

Consistent with the scope of the analysis, the area of study is bound to Napoli city and the trajectories involved are collected from light weight vehicle trips (private vehicles). The o-d matrices have been calculated discretizing the area in traffic analysis zones matching the 30 macro census tract of Napoli city and referred to 15 minutes time-interval. The spatial and temporal discretization demonstrated to bind the o-d matrices sparsity, limiting the presence of zero values as discussed in Section 4.3. To capture the great percentage of systematic trips, the o-d flows have been calculated for four different working days specifically referring to the morning peak-hour (from 7 to 8 AM) choosing a 60 minutes quasi-dynamic interval. By setting these conditions the vector of percentage of distributions defined per each origin referred to 15 minutes time-interval is approximated to a vector containing their average values over the specified quasi-dynamic interval (one hour), leading to the quasi-dynamic o-d flows as expressed by Equation 2.5 introduced in the previous section.

Figure 4.28, *Figure* 4.29, *Figure* 4.30 and Figure 4.31 depict the scatter plots respectively illustrating the error between 15-minutes trajectory o-d flows and quasi-dynamic o-d flows by trajectory for the four working days considered. Experimental evidence suggests that

distribution shares cannot be properly approximated to their average values over a quasi-dynamic period of one hour, considering that intrinsic error quantified by the $cvRMSE$ calculated on 15-minutes demand flows, is greater than one for each of the considered working days, indeed its average value is equal to 1.31 (see Table 4.5). Therefore, distribution shares evolution does not remain stable along the morning peak hour, thus the quasi-dynamic o-d matrix cannot be used as an adequate approximation of the trajectory o-d matrix, as also confirmed by the values of $cvRMSE$ reported in Table 4.7. Results lead to two different conclusions: either the utilised trajectory data sample penetration is not adequate enough to capture quasi-dynamic evolution of o-d flows, either the evolution of demand in an urban context does not follow a quasi-dynamic trend. Nevertheless, this result has contributed to the generation of the ground-truth population for the synthetic experiments described in the next chapters, being a pivot value to define the dynamic evolution of demand at urban level, (see section 5.2 for details).

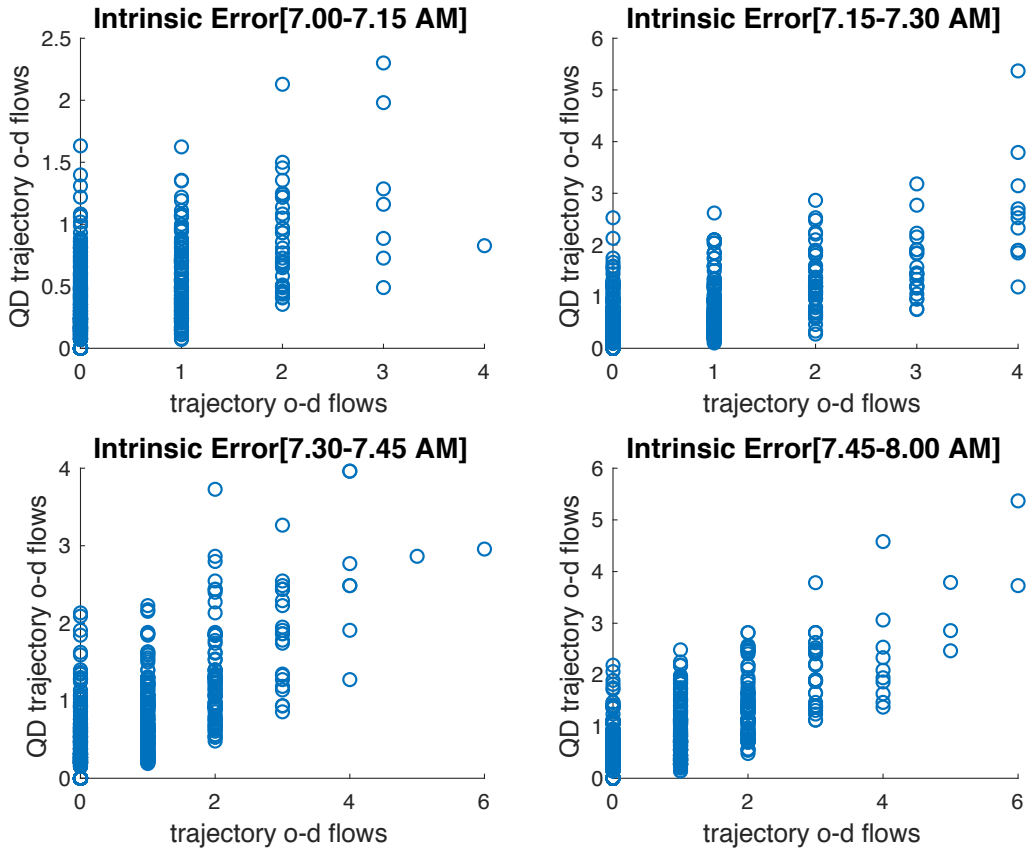


Figure 4.28 linear fitting between trajectory o-d flows and quasi-dynamic o-d flows, τ = 60 min, day 1

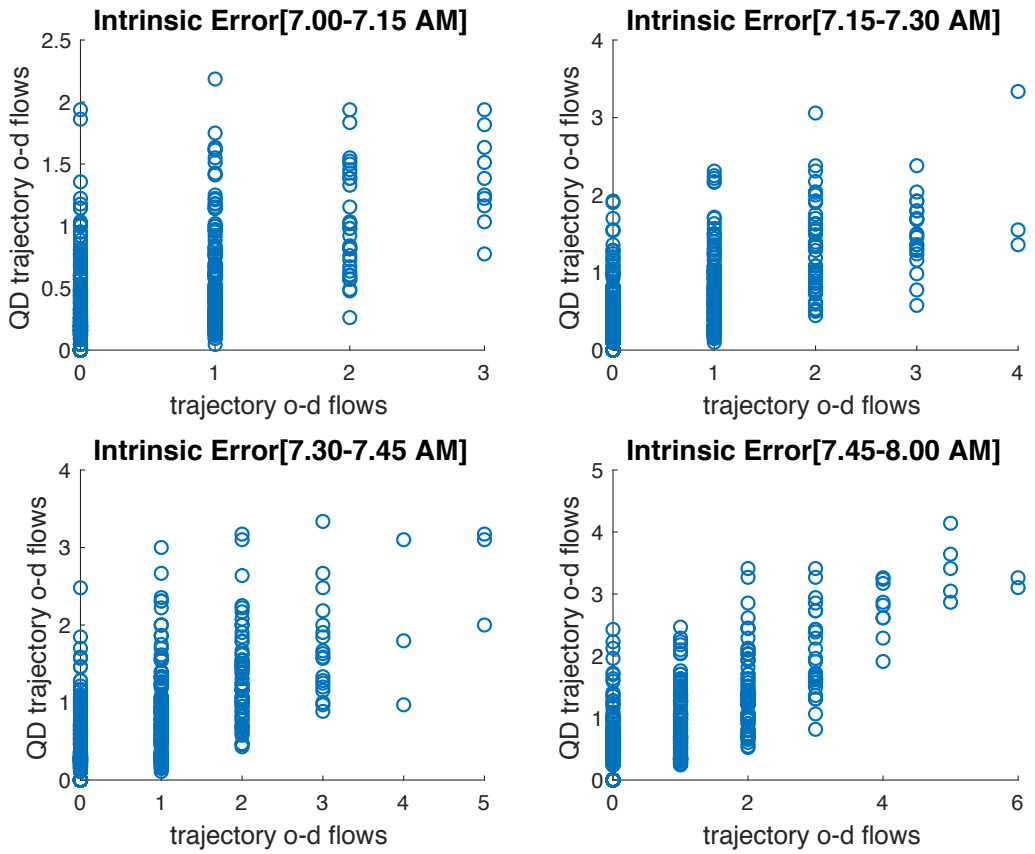


Figure 4.29 linear fitting between trajectory o-d flows and quasi-dynamic o-d flows, τ = 60 min, day 2

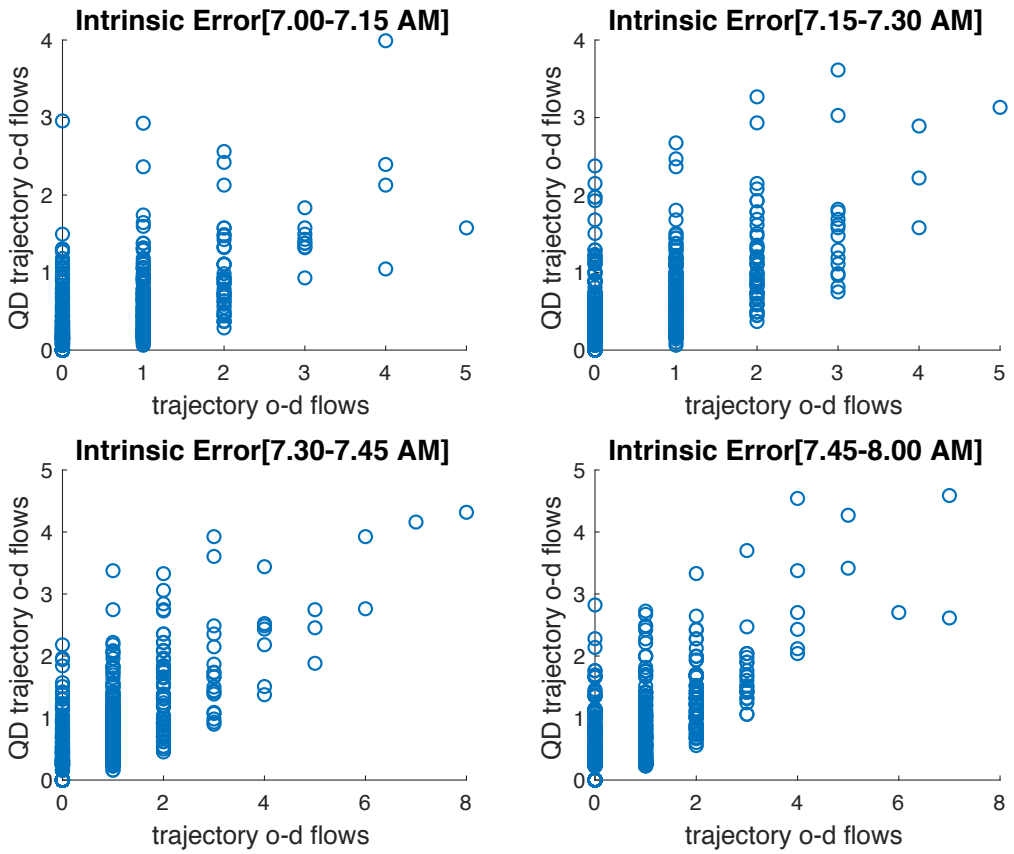


Figure 4.30 linear fitting between trajectory o-d flows and quasi-dynamic o-d flows, τ = 60 min, day 3

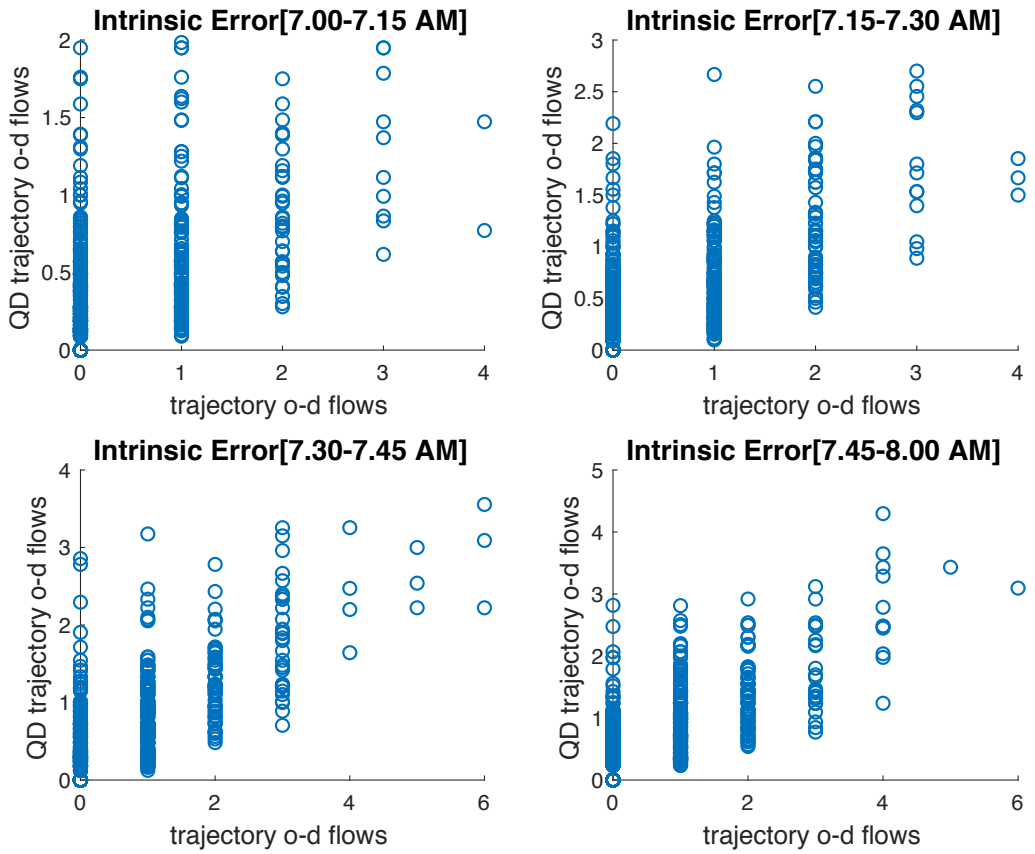


Figure 4.31 linear fitting between trajectory o-d flows and quasi-dynamic o-d flows, τ = 60 min, day 4

INTRINSIC ERROR: CVRMSE	**OCT 4TH 2017**	**OCT 11TH 2017**	**OCT 18TH 2017**	**OCT 25TH 2017**
7.00-7.15 AM	1.64	1.63	1.70	1.62
7.15-7.30 AM	1.37	1.34	1.32	1.38
7.30-7.45 AM	1.15	1.15	1.21	1.25
7.45-8.00 AM	1.13	1.04	1.02	1.04

Table 4.7 trajectory o-d flows intrinsic error - cvRMSE values- (mean value 1.31)

5 Performance Analysis of Direct Scaling

This chapter presents laboratory experiments testing direct scaling techniques providing a sensitivity analysis on trajectory sample representativeness and biases key factors: penetration level, penetration distribution and other sample characteristics combined with other sources of information such as a set of link counts measurements.

5.1 Motivation

The analysis of the literature revealed that few studies explored the opportunities of using trajectory data to enhance or propose alternative o-d estimation methods in diverse manners (see Sections 2.1 – 2.3). Notwithstanding, given that the majority of the provided examples are mainly based on real case studies, the unobservability of o-d flows remains a primary issue for demonstrating the effectiveness of such methods. In light of this, evaluating the quality of o-d estimation methods using real datasets can be troublesome, indeed it is not possible to provide decisive experimental evidence of their effectiveness and advantages over others, except when applied for closed systems such as highway networks in which the true phenomena underlying can be observed. As introduced in Section 1.1, a consolidated procedure to obtain more accurate estimates is to update the a priori estimate of the o-d matrix exploiting a set of traffic measurements (link traffic counts, speed measurements, link travel times, etc.); however, a perfect fit of the observed measurements used for o-d estimation/updating procedures does not necessarily imply that estimated o-d flows closely match the true underlying values. Hence, reliable validation techniques and especially experiments based on synthetic data are crucial to assess o-d estimation methods performance.

Nevertheless, from the analysis of the literature, no evidence of existing laboratory experiments with a remarkably wide range of case studies and experimental setups testing scaling techniques and o-d estimators in presence of trajectory data was found. In light of this, synthetic experiments are extensively applied in this thesis to develop a systematic analysis of trajectory data sample scaling methods and o-d flows updating methods in presence of trajectory data, investigating the impact of trajectory and link counts sample characteristics on o-d estimation

model/scaling technique performances. The main goal is to provide a sensitivity analysis focussing on trajectory sample representativeness and biases key factors: penetration level and penetration rate distribution. To this end, the outcomes from each model/scaling technique are analysed considering different scenarios given by a variety of combinations of the project variables which can be roughly categorised into two main classes: the trajectory sample characteristics (e.g. penetration level and distribution) and the link counts sample characteristics (e.g. number of sensors and sensor locations).

The results from the trajectory data analysis conducted in this study, consistent with other studies found in literature, demonstrated that a generic sample of trajectory data can represent a small percentage (e.g. up to 10%) when compared to the total number trips collected by census surveys, (see Chapter 4). Evidence from the experimental analysis indicates that using a direct scaling method is a primary and essential step to rescale the observed o-d flows by trajectory data and obtain a first crude estimate of the o-d matrix. For the sake of clarity, a direct scaling method is referred as to a procedure by means of which o-d flows derived from trajectory data samples are rescaled or normalised to more realistic values, not involving the application of statistical methods or optimization processes. Additionally, a different type of direct scaling can be defined according to the specification of the upscaling factor(s). The types of direct scaling technique tested in this work are described in Section 5.2.3. Given the essential role of direct scaling methods and the lack of studies presenting synthetic experiments, this part of the study attempts to evaluate to which extent a direct scaling procedure can provide reliable o-d flow estimates by considering a variety of different scenarios. Bringing light on this research question, this work can provide structured guidelines for the use of trajectory data in the o-d flows estimation problem.

5.2 Methodology

To validate and compare different approaches in a variety of settings and conditions, the proposed laboratory experiments are based on the generic structure of the benchmarking platform presented by Antoniou et al. 2017 which aims to provide a reliable test bed examining o-d flows estimation methods and algorithms under equal/standardised conditions. The methodology pivots on the use of synthetic data, basically representing the ground truth data, which allows to set up a full knowledge of the o-d matrix and the underlying phenomena. Although, the methodology applied in this thesis required some necessary adaptations to the specific case. Its structure encompasses three main components as illustrated in Figure 5.1:

- *Synthetic ground-truth case study setup*: defining a true testbed in terms of o-d flows, network characteristics and any other relevant parameter related to demand and supply, including an assignment model performing the interaction between demand and supply which allows to calculate the true link flows and traffic flows characteristics and to ensure mutual consistency between o-d flows and traffic flow characteristics throughout the entire estimation process. The testbed must be consistent with the sensitivity analysis which has to be performed.
- *Design of experimental setup*: mimicking real situations with uncertainties and/or biases of available o-d flow estimates derived by the trajectory data samples. Specifically, the experimental settings refer to a wide range of assumptions on the selection of measurements to simulate realistic scenarios regarding the availability of a subsets of measurements in specific network locations (e.g. link flows measurements) and a subset of o-d flows (e.g. derived from trajectory data samples). This step also includes the definition of key indicators and goodness-of-fit (GOF) measures assessing the estimation methods performances.
- *Testing*: applying the scaling techniques and/or the o-d estimation/updating algorithms under analysis fed by the setup values defined in the second step; the resulting estimated/updated values are compared with the true values defined in the first step of the procedure, using the selected GOFs.

As also pointed out in Antoniou et al. 2017, a key feature in the assessment of o-d flows estimators is to consider a wide experimental plan to test a variety of settings and heterogeneous conditions (network scale, demand volumes, number and location sensors, accuracy and reliability of measurements) consistently defined across methods/algorithms to be tested, yielding to a comprehensive range of possible scenarios.

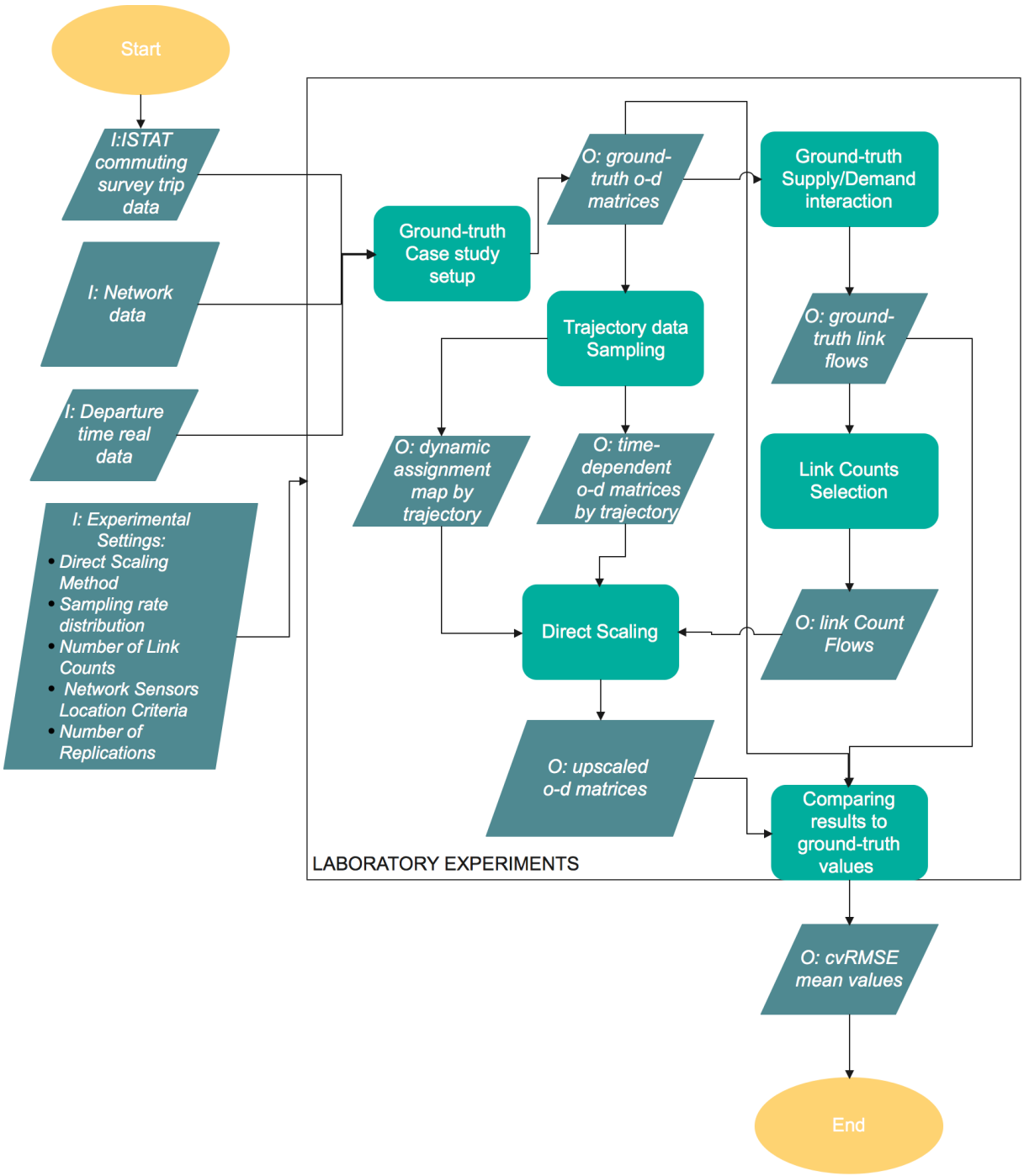


Figure 5.1 Illustrative flowchart of the laboratory experiments testing direct scaling techniques performances

Details regarding the adopted procedure for each step are reported in the next sections.

5.2.1 **Ground-truth case study setup**

To develop the synthetic ground truth case study a set of demand and supply parameters needs to be defined:

- the test site network and the appropriate zoning system;
- the ground truth o-d flows indicating the exact number of trips between origins and destinations;

- the technique to generate the set of trajectories providing the sequence of links traversed by each vehicle and the departure time distribution indicating the time by which each vehicle leaves each origin;

Furthermore, it is necessary to select an assignment model to simulate the interaction between demand and supply and thus obtain the true link flows and the necessary traffic parameters (e.g. travel times, vehicles speed etc.) consistent with true o-d flows.

To cover a wide range of variety of real case studies and develop the sensitivity analysis of trajectory data sampling characteristics, two different test sites were selected composing the ground truth testbeds for small and large scale networks respectively. Experiments on the selected test sites were conducted simulating the demand flows evolution over the morning peak hours which, for the specific test sites, realistically occurs between 6 AM and 10 AM.

5.2.1.1 Small network case study

For the small scale network case study, the test site selected by Yang et al. 2017 was considered, consisting of a road network in the northern part of Maryland State, whose topology is represented in *Figure* 5.2. The network consists of 28 nodes and 74 links (equivalent to 37 bidirectional links) as shown in Figure *5.3* processed by MatLab. The simulation period is set to 4 hours (6-10 AM) corresponding to 16 demand time intervals, choosing 15 minutes long time windows. 40 o-d pairs are selected for the case study and 12 out of the 74 links are equipped with traffic sensors to detect the number of vehicles occupying the link in each time interval. The measured links are chosen according to two criteria extensively described in Section 5.3.1, to test the efficacy of the two correspondent solutions of the sensor location problem (max flow & random selection).

To define the true time-dependent o-d flows, a different criteria has been used to define demand volumes in order to control its dynamics, to define the intrinsic error parameters and evaluate the applicability of quasi-dynamic hypothesis, while the route choice has been derived by Yang et al 2017, selecting the paths whose lengths are shorter than the others. Table 5.1 summarizes the route choice probability of all paths between all o-d pairs.

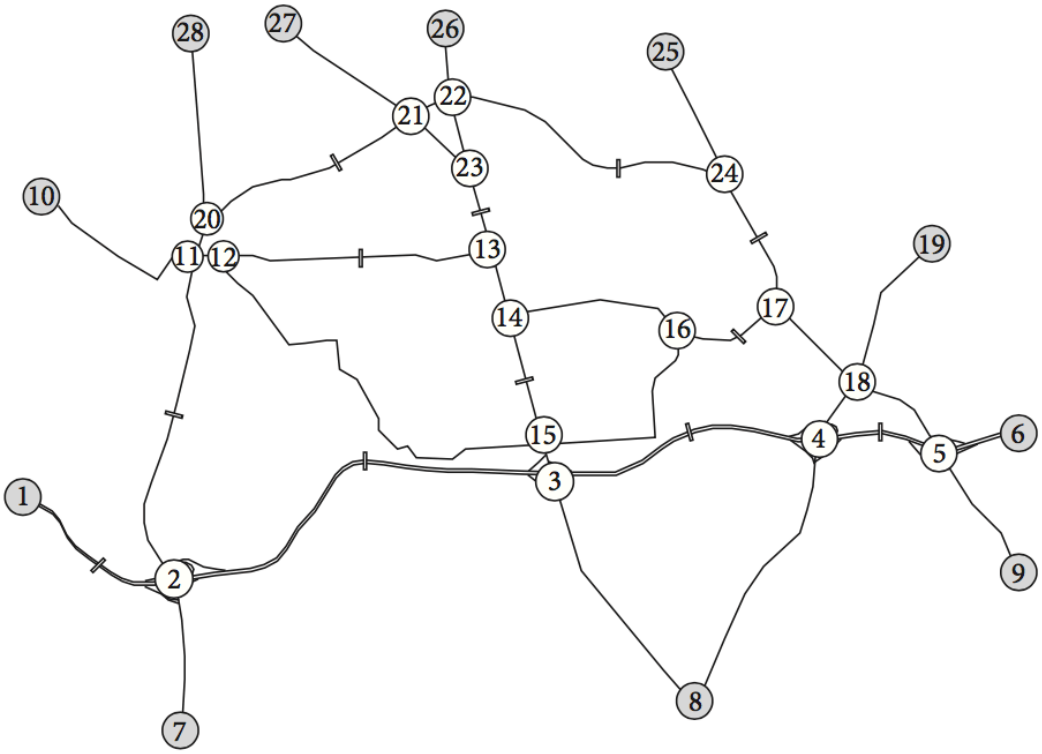


Figure 5.2 small-scale test site Source: Yang et al. 2017

OD	Path	Ratio
1→6	1→2→3→15→16→17→18→5→6	0.05
	1→2→3→15→14→16→17→18→5→6	0.05
	1→2→3→4→5→6	0.9
1→26	1→2→11→12→13→23→22→26	0.3
	1→2→11→20→21→22→26	0.3
	1→2→3→15→14→13→23→22→26	0.4
1→8	1→2→3→8	1.0
1→25	1→2→11→12→13→23→22→24→25	0.3
	1→2→3→15→16→17→24→25	0.3
	1→2→11→20→21→22→24→25	0.4
1→28	1→2→11→20→28	1.0
10→9	10→11→2→3→4→5→9	0.4
	10→11→12→13→14→15→3→4→5→9	0.3
	10→11→12→13→14→16→17→18→5→9	0.3
10→8	10→11→2→3→8	0.5
	10→11→12→15→3→8	0.5
10→26	10→11→12→13→23→22→26	0.5
	10→11→20→21→22→26	0.5
6→1	6→5→4→3→15→14→13→12→11→2→1	0.1
	6→5→4→3→2→1	0.9
6→26	6→5→18→17→16→14→13→23→22→26	0.3
	6→5→18→17→24→22→26	0.3
	6→5→4→3→15→14→13→23→22→26	0.4
6→10	6→5→4→3→15→12→11→10	0.5
	6→5→18→17→16→14→13→12→11→10	0.5
6→8	6→5→4→8	1.0
6→28	6→5→4→3→2→11→20→28	0.4
	6→5→4→3→15→12→11→20→28	0.3
	6→5→18→17→16→14→13→12→11→20→28	0.3
6→1	19→18→17→16→15→3→2→1	0.5
	19→18→4→3→2→1	0.5
19→8	19→18→4→8	1.0
19→27	19→18→17→16→14→13→23→21→27	0.4
9→25	9→5→18→17→24→25	1.0
9→8	9→5→4→8	0.5
	9→5→18→4→8	0.5
9→19	9→5→18→19	1.0
9→1	9→5→4→3→2→1	1.0
9→26	9→5→4→3→15→14→13→23→22→26	0.4
	9→5→18→17→16→14→13→23→22→26	0.3
	9→5→18→17→24→22→26	0.3
9→27	9→5→18→17→24→22→21→27	0.5
	9→5→4→3→15→14→13→23→21→27	0.5
28→8	28→20→11→2→3→8	0.6
	28→20→11→12→13→14→3→8	0.2
	28→20→11→12→15→3→8	0.2
28→7	28→20→11→2→7	1.0
27→9	27→21→23→13→14→15→3→4→5→9	0.6
	27→21→22→24→17→18→5→9	0.4
27→1	27→21→20→11→2→1	1.0

OD	Path	Ratio
26→10	26→22→23→13→12→11→10	0.3
	26→22→21→20→11→10	0.7
26→8	26→22→23→13→14→15→3→8	1.0
26→19	26→22→23→13→14→15→3→4→18→19	0.3
	26→22→24→17→18→19	0.7
7→28	7→2→11→20→28	1.0
7→25	7→2→3→4→18→17→24→25	0.4
	7→2→11→20→21→22→24→25	0.3
	7→2→3→15→14→16→17→24→25	0.3
7→9	7→2→3→4→5→9	1.0
8→25	8→3→15→16→17→24→25	0.5
	8→4→18→17→24→25	0.5
8→26	8→3→15→14→13→23→22→26	1.0
8→1	8→3→2→1	1.0
8→28	8→3→2→11→20→28	0.5
	8→3→15→14→13→12→11→20→28	0.5

Table 5.1 Turning ratio via different paths. Source: Yang et al. 2017

5.2.1.2 Large network case study

For the large-scale network case study, the laboratory experiments were carried out with reference to the real network of the Caserta province. The road network is based on the OpenStreetMap (OSM) topology, consisting of 4.879 nodes and 12.671 links, hierarchically clustered into four mutually exclusive sets of network levels depending upon specific road attributes such as link travel time and link monetary costs, available from previous studies. In the following, the links whose hierarchy is lower than 3 are referred as relevant links (depicted in Figure 5.3 in black). The relevant links are considered for specific considerations on the calculation of the Mean Absolute Percent Error (*MAPE*) and the Root Mean Square Error (*RMSE*) indicators (see Section 5.4.1). The selected traffic analysis zones correspond to the 104 municipalities of the Caserta province whose boarders are depicted in Figure *5.3* in green, yielding to 10816 o-d pairs.

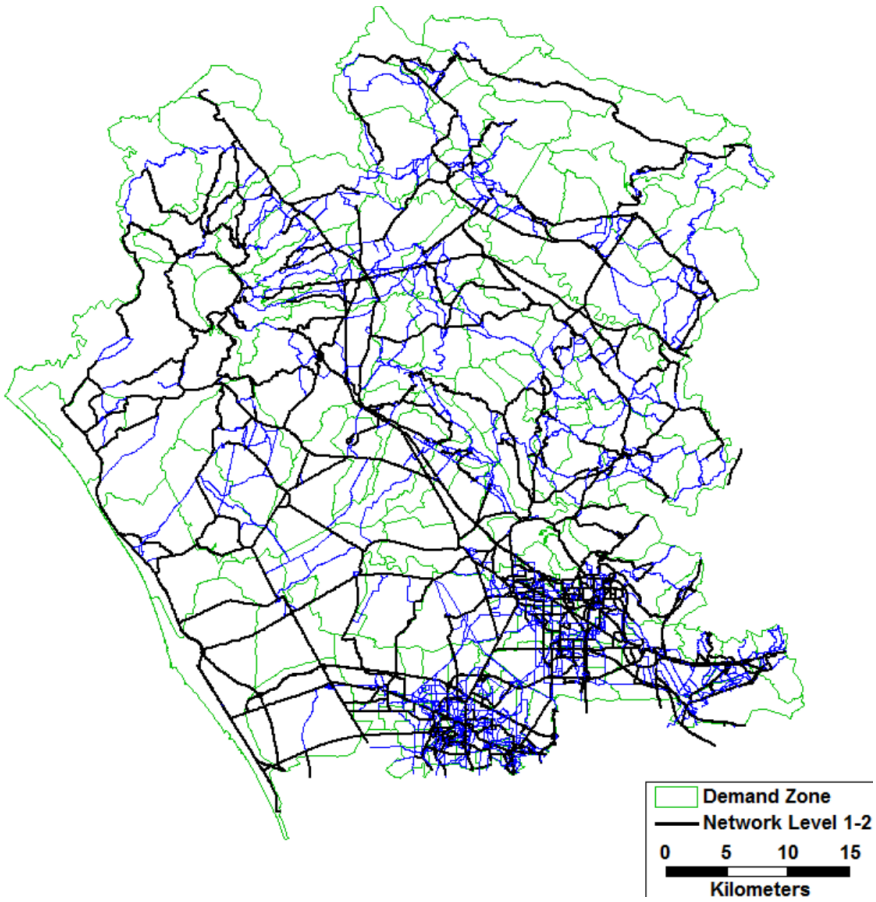


Figure 5.3 Caserta province road network: hierarchy levels and traffic analysis zones

To examine the steady state context, ground-truth demand values were derived from ISTAT (Italian National Institute of Statistics) 2011 commuting survey, which reports the internal flows of Caserta province relative to back and forth trips of the same day to go to or to come from the work/study place during the morning peak hours (approximately 7.15 – 9.15 AM) of a typical working day (usually Wednesday). Consistently with the resulting o-d matrix, a population of travellers between each o-d pair has been generated. These operations allowed to set up a synthetic but realistic case study, wherein the travel demand is known.

To model the route choice of each traveller of the population and thus generate a set of trajectories consistent with the o-d matrix, a random utility model based on the generalized perceived route costs has been adopted. Specifically, the utility function of each individual *i* is expressed as a linear combination of the travel time and the monetary cost, whose values have been derived from previous studies. Formally, the random utility is expressed as:

$$U_t^i = \beta_t^i t_r + \beta_c^i c_r + \varepsilon_r^i \quad (5.1)$$

Wherein:

- U_t^i is the perceived utility of route *r* for the user *i*;
- t_r is the travel time of the route *r*;

- c_r is the monetary cost of the route *r*;
- β_t^i and β_c^i are the utility function coefficients for the user *i*;
- ε_r^i is the random term in the perceived utility for the route *r* and user *i*;

To simplify the case study, congestion phenomena are considered negligible and, since link costs are assumed as additive, route costs can be calculated as the sum of the corresponding link costs. The travel time and monetary cost coefficients of each individual (β_t^i and β_c^i) , have been respectively drawn from two different mono-variate Normal distributions to introduce heterogeneity among the individuals of the synthetic population. The ratio between the two coefficients represents the value of time given by the individuals. Since the distribution of the ratio of two Normal random variables is unknown, their average values are set such that the value of time (*VOT*) is equal to 10 €/h. Furthermore, although the Normal distribution is theoretically unbounded, the dispersion factors for the distributions of β_t^i and β_c^i have been set such that the results did not yield to unrealistic positive values. The random term ε_r^i has been drawn upon a multivariate normal distribution, with a covariance matrix consistent with the Daganzo and Sheffi (Daganzo and Sheffi, 1977) assumption, as recalled in the following:

$$c_l^p \sim N(c_l, \xi \cdot c_l) \tag{5.2}$$

Wherein:

- c_l^p represents the perceived cost of the link *l*;
- c_l is its expected value;
- ξ is a proportionality factor to be estimated.

Drawing upon the Daganzo and Sheffi assumption, the covariance matrix $\boldsymbol{\Sigma}_r$ of the perceived costs of the routes is given by:

$$\boldsymbol{\Sigma}_r = \xi \cdot \boldsymbol{Q}^T \cdot \boldsymbol{\Sigma}_l \cdot \boldsymbol{Q} \tag{5.3}$$

where $\boldsymbol{Q}$ is the link-paths incidence matrix and $\boldsymbol{\Sigma}_l$ is the diagonal variance-covariance matrix of the link costs consistent with the Daganzo-Sheffi assumption. The constant of proportionality ξ has been computed per each o-d pair *s* setting the value of variation coefficient *cv* according to the following expression:

$$\xi = cv^2 \cdot C_{od,min} \tag{5.4}$$

being $C_{od,min}$ the minimum cost among all the paths connecting the considered o-d pair. Starting from the previous assumptions, a sequence of links defining a path has been assigned to each user travelling between all o-d pairs such that the total number of users of the synthetic population was equal to the total number of trips and consistent with each actual od matrix entry.

5.2.1.3 Dynamic evolution of o-d flows for the large-scale network case study

To study the dynamic evolution of the travel demand in the case of large-scale network, a different set of ground truth o-d flows values has been defined. Specifically, the time-dependent o-d flows were referred to the morning peak period indicated with *T* occurring from 6 AM to 10 AM. Demand values were referred to 16 time intervals corresponding to 15 minutes long time slice θ, composing the entire time-horizon *T*.

Real traffic data reporting the distribution of departure time during the morning peak has been used to reconstruct realistic ground-truth values of time-dependent o-d flows. Specifically, the traffic data contained the percentages of travellers departing every 5 minutes with respect to the travel demand referred to entire day of Caserta urban area (percentage of trips detected every 5 minutes).

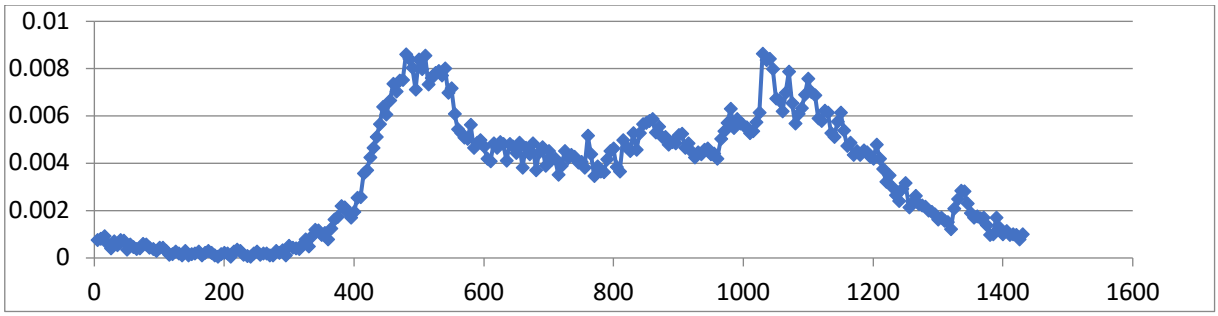


Figure 5.4 x: time of the day (min); y: percentage of departures. Real data from urban loop detectors.

Processing the data, the percentages of trips departing every hour during the time horizon *T* ($p_{h|T}$) and every time slice θ per each hour *h* of the time horizon ($p_{\theta|h}$) have been calculated for the time-horizon *T* (6-10 AM). Results are shown in *Table* 5.2:

Peak Hour	$p_{h\|T}$	$p_{\theta_{1\|h}}$	$p_{\theta_{2\|h}}$	$p_{\theta_{3\|h}}$	$p_{\theta_{4\|h}}$
6-7 AM	0.10	0.17	0.18	0.25	0.31
7-8 AM	0.29	0.23	0.24	0.26	0.24
8-9 AM	0.36	0.23	0.27	0.24	0.23
9-10 AM	0.25	0.37	0.31	0.25	0.22

Table 5.2 Percentage of departure during the morning peak period for each hour h of the time horizon ($p_{h|T}$) and for each time slice θ of each hour h of the time horizon T ($p_{\theta|h}$)

Three sets of information were derived from the o-d matrix reported by the ISTAT commuting survey with reference to the peak hour 7-8 AM:

- the total number of trips ($N^{ISTAT\ (7-8\ AM)}$) ;
- the generated flow from each origin ($g_o^{7-8\ AM}$);
- the percentages of distribution ($p_{d|o}^{ISTAT\ (7-8\ AM)}$) ;

Given the percentage of departures during the same time interval reported by the real data from traffic sensors as in *Table* 5.2, the total number of trips referring to the time-horizon *T* has been calculated as:

$$N^T = \frac{N^{ISTAT\ (7-8\ AM)}}{p_{7-8\ AM}}; \tag{5.5}$$

Consequently, the number of trips per each hour of the time horizon T has been was derived as:

$$N^h = N^T \cdot p_{h|T};\ \forall h \in \mathrm{T} \tag{5.6}$$

Wherein $p_{h|T}$ indicates the percentage of generated demand during each hour *h* of the time horizon *T*. The demand flows leaving each origin during each hour of the time-horizon (g_o^h) have been proportionally amplified or reduced using as benchmark values the flows generated from each origin of the set of origin nodes *O* during the reference period 7-8 AM ($g_o^{7-8\ AM}$). To simulate the travel demand fluctuations among different time-slices and to introduce a dispersion factor, the generated flow from each origin has been perturbed drawing upon a standard normal distributed random variable (identified as *z*) using a pre-specified coefficient of variation cv_{gen}:

$$g_o^h = \left(\frac{N^h}{N^{ISTAT\ (7-8\ AM)}}\right) \cdot g_o^{7-8\ AM} + \left(cv_{gen} \cdot z\right);\ \forall o \in O, \forall h \in \mathrm{T}_h \tag{5.7}$$

Wherein:

- $z \sim N(0, 1)$
- cv_{gen}=0.1

Starting from these values and exploiting the percentage of travellers departing every 15 minutes over each hour of the time-horizon reported in *Table* 5.2 ($p_{\theta|h}$), the final flows leaving each origin during each time-slices have been calculated according to the following:

$$g_o^\theta = g_o^h \cdot p_{\theta|h};\ \forall \theta \in \mathrm{T}_\theta \tag{5.8}$$

The percentages of distribution $p_{d|o}^\theta$ expressing the portion of demand generated leaving from origin *o* and heading to the destination *d* of the set of destinations *D* during time-slice θ were assigned perturbing the percentages of distribution derived from the hourly-based o-d matrix

reported by ISTAT ($p_{d|o}^{ISTAT\ (7-8\ AM)}$), extracting pseudo-random values from a standard normal distribution using a pre-specified coefficient of variation cv_{dis}, leading to:

$$p_{d|o}^{\theta} = p_{d|o}^{ISTAT\ (7-8\ AM)} + (cv_{dis} \cdot z);\ \forall d \in D, \forall \theta \in \mathrm{T}_{\theta} \tag{5.9}$$

The perturbed percentages of distribution have been subsequently scaled such that their sum was equal to 1, satisfying the following condition:

$$\sum_{d \in D'} p_{d|o}^{\theta} = 1\ \forall o \in O, \forall \theta \in \mathrm{T}_{\theta} \tag{5.10}$$

Therefore, the final time-depending o-d flow values composing the ground-truth values for each o-d pair of the set *OD* were derived as:

$$d_{od}^{\theta} = g_{o}^{\theta} \cdot p_{d|o}^{\theta}\ \forall od \in OD \tag{5.11}$$

The values of the coefficients of variation for both perturbations have been defined in order to obtain an intrinsic error consistent with the average value found in the experimental analysis of the quasi-dynamic assumption in urban context described in Section 4.5.

Given the o-d flows values, a set of trajectories has been generated according to the same path choice model used for the steady state context, yielding to a total number of trajectories equal to 165,237. Details of each test bed are summarised in *Table* 5.3.

	Small Network ***Yang et al. 2017***	**Large Network** ***Simonelli et al. 2019***
Nodes	28	4,879
Links	74	12,671
Origins/Destinations	11	104
o-d Pairs	40	10,816*
Trips	19,185	165,237

Table 5.3 Test beds for the laboratory experiments

Considering the higher complexity of the dynamic state context and the consequent increasing number of variables to be estimated, a smaller set of o-d pairs with respect to the steady state context has been selected according to a max-flow criterion. Therefore, to lean the estimation process only o-d flows greater than 20 have been considered, reducing the number of o-d pairs from 10816 to 394 (*). Consequently, the total number of time-dependent o-d flows to be estimated was given by:

$$n_{od} = n_{od}^{\theta} \cdot n_{\theta} = 394 \cdot 16 = 6304;\ \ n_{od}: d_{od} > 20 \tag{5.12}$$

Wherein n_{od}^{θ} is the number of o-d flows to be estimated for each time slice θ and n_{θ} is the number of time-slices composing the simulation period. Considering a smaller set of o-d pairs

with respect to the steady-state context test bed, yields a reduction of the total number of paths connecting the o-d pairs and consequently to a reduction of the links travelled by each user of the network.

The experiments are conducted considering the case of uncongested network for both test sites: while for the small scale network case studies path choice probabilities are fixed and known, the assignment model utilised to simulate the interaction between demand and supply parameters in the large scale network case study involves a flow propagation model and a route choice model assuming a linear structure. Therefore, since the link costs are assumed as additive and congestion phenomena are considered as negligible, the considered assignment model relies on a linear approximation of the assignment map identifying the relationship between o-d flows and link flows. A dynamic network loading algorithm has been used to identify the flow propagation on the network for the dynamic state conditions. To operate in the dynamic context, the algorithm determines the time-dependent link volumes, together with link and path travel times, given the time-varying path flow departure rates over a finite time horizon.

5.2.2 **Design of Experimental Setup**

The second main implementation task to develop the laboratory experiments is to provide a comprehensive design of the experimental setup in terms of:

- The trajectory sample and its characteristics, defining the set of direct measures of o-d flows;
- the set of traffic flow measurements and locations, composing the indirect measures to estimate the demand flows;
- the definition of the direct scaling technique to test (Section 5.2.3);
- the choice of GOFs (Section 5.2.4).

A delicate step of designing the experimental setup consists of generating the trajectory data sample and defining its characteristics. As recalled in Section 1.1.2, there are some critical aspects challenging trajectory data exploitation in transportation studies. Indeed, collected data should have a sufficient scale to apply inference techniques, thus the penetration level of the sample is a key factor for deriving accurate demand estimates; furthermore, trajectory data are usually acquired from specialised vendors, thus sample variety can be compromised and the collected sample can be biased.

In light of this, it is necessary to conduct an extended analysis focussing on two important characteristics associated to the trajectory data sample:

- The penetration level: also known as sampling rate, it can be expressed as a random variable, whose average value indicates whether the collected sample is representative of the true underlying o-d flows. By definition, during a laboratory experiment in which ground truth values are available, an estimate of the penetration rate can be obtained comparing the total number of collected trips with the ground truth trip total number.
- The distribution of the penetration rate among the different elements of the network e.g. o-d pairs, origins, destinations. Since the trajectory data sample is usually not collected according to a systematic sampling, the data collection yields to biased samples across space and time e.g. among different traffic analysis zones and time of the day (concerning the within-day dynamic context).

To explore the range of variability and conduct a robust sensitivity analysis of these two factors, different techniques to generate trajectory samples simulating a data collection process have been proposed. The range of variability of the penetration rate, its temporal and spatial distributions have been set according to the values observed from the analysis of a real trajectory data sample, as described in Chapter 4. Section 4.4.2 reports the average value of the penetration rate obtained confronting the total number of sampled trajectories and the total number of trips registered in the last commuting survey referred to a typical working day. Its value resulted around 6%, and its maximum value resulted around 30%, not considering outlier (non-realistic) values. Therefore, the analysed range in the laboratory experiments concerning some of the o-d flow updating procedures was set to [5% , 30%], while for the assessment of direct scaling technique performances and to validate the method, the entire range of variability has been considered [1% -100%] with a step range of 5%, such that it was possible to verify whether the error was equal to zero in the case of maximum possible sampling rate (100%).

To assess the impact of sampling rate distribution on the level of accuracy of demand flows, two sampling process have been used to extract a sample of trajectories from the synthetic population corresponding to the two types of trajectory data sample introduced in Section 1.1.2: opportunistic data and purpose-oriented data. Concerning the former case, in light of the strong variability of the sampling rate among different o-d pairs which arose from the analysis of the opportunistic data sample conducted in Chapter 4, the trajectory sample has been generated assuming a different probability of extraction for each o-d pair (as illustrated in. Figure *4.23*). To simulate the trajectory data sampling process, each o-d pair is randomly associated with a different weight, thus the algorithm picks the trajectories with higher weights until the desired total number of trajectories is extracted. While, to mimic the latter case deriving from a well-structured sampling process and designed with a specific purpose for transport research and

applications, a uniform distribution of the sampling rate has been assumed. In this case, the sampling rate is assumed homogeneous among o-d pairs, thus the probability of extracting trajectories from a certain o-d pair is equal among all o-d pairs. Therefore the algorithm selects the trajectories from the ground-truth trajectory list drawing upon a uniform random distribution.

The dimension of the link counts sample has been set according to a plausible value consistent with existing applications in literature and based on the typical number of available loop traffic sensors installed in an urban area; the penetration of link traffic sensors, meaning the ratio of the number of measured links to the total number of the links of the is commonly around 1% out of the total number of links. The analysis evaluates the sensitivity of the results considering a range varying from approximately 0.5 to 2 % out of the total number of links. The maximum value (2%, corresponding to 254 link counts) has been considered to eventually observe any improvement of the technique performances when a greater number of link counts is available. To define link traffic sensors locations and thus to select the set of link traffic measurements, two criteria have been applied: the k-max flow and random selection criteria. According to the k-max flow criteria, the set of the available link counts contains the measurements of the k links interested by the major flows. The link count sections have been located on the network such that the equations describing the relation between link traffic measurements and o-d flows were linearly independent. This aspect is critical to feed models with non-redundant information and to guarantee the applicability of GLS-based estimators. For the within-day dynamic context, the k maximum link flows composing the sample have been selected considering the time-slice interested by the maximum demand volume (i.e. the total number of travelling vehicles).

Conversely, according to the random selection criteria, the set has been defined extracting pseudorandom locations of the sensors guaranteeing the linear independency of measurement equations. For each experiment deriving from the combination of the experimental settings shown in Table 5.4, ten replications were generated to attain stable results on the goodness of fit measure mean values.

Observations	Experimental Settings			Number of Replications
Link Counts	Max Flows	Random	Number of Link Counts {0.5, 1, 2} %	**10**
Trajectory	Uniform	OD-based	Max Sampling Rate 1÷100 %	

Table 5.4 Values of the project variables used to test direct scaling techinques by means of laboratory experiments

5.2.3 **Direct scaling techniques: hypothesis on upscaling factors**

Several direct scaling techniques can be tested according to the specification of the upscaling factor by means of which o-d flows derived by trajectory data sample can be scaled; some of them have been already introduced in Section 2. The first direct scaling method tested in this work was proposed by Van Aerde et al. (1993) and applied by Yang et al. (2017) for the o-d flows estimation problem; its formulation is reported here for the sake of readability:

$$\gamma^{\vartheta} = \frac{\sum_{l=1}^{nlc} f_l^{\vartheta}}{\sum_{l=1}^{nlc} f_l^{traj,\vartheta}} \quad \forall\, \vartheta \in T \tag{2.47}$$

As introduced in Section 2, the method consists of rescaling the trajectory o-d matrix by means of an upscaling factor defined for some period ϑ (γ^{ϑ}) obtained as the ratio of the total number of vehicles observed from link traffic counts at time-interval ϑ to the total number of tracked vehicles traversing link l at time-interval ϑ. The method defines a scalar upscaling factor per each time-slice, implying that the dynamic evolution of the upscaling factor follows the same demand fluctuations.

The second method is a particularization of the technique above, involving a scalar upscaling factor for all the o-d pairs and for the entire time-horizon. Therefore, the upscaling factor is independent of the o-d pair and the time slice considered. This value can be easily calculated as the ratio of the sum of counted link flows to the sum of link flows corresponding to the o-d flows derived by the trajectory sample. As illustrated in *Figure* 5.1, to perform the dynamic traffic assignment, dynamic traffic assignment map entries are directly derived from the

trajectory data sample according to the procedure described in Section 2.2.3. The upscaling factor is formally expressed by:

$$\gamma = \frac{\sum_{\vartheta=1}^{n_\vartheta} \sum_{l=1}^{nlc} f_l^\vartheta}{\sum_{\vartheta=1}^{n_\vartheta} \sum_{l=1}^{nlc} f_l^{traj,\vartheta}} \tag{5.13}$$

A novel formulation which, by the author's best knowledge at this time, has not been introduced in literature involves a time-dependent upscaling factor defined by link count section. Therefore, this hypothesis implies a different value of the upscaling factor per link count section and per each time interval, leading to the following formulation:

$$\zeta_l^\vartheta = \frac{f_l^\vartheta}{f_l^{traj,\vartheta}} \; \forall \, \vartheta \; \in T, \forall \; l \; \in LC \tag{5.14}$$

Wherein *LC* is the subset of links of the network equipped with a detector.

The coefficient ζ_l^ϑ is defined as the ratio of the traffic flow of link *l* measured at time interval *t* to the link traffic flow deriving from the loading of o-d flows by trajectory onto the network at the same time interval. To rescale the o-d flows derived by the trajectory sample, the coefficient ζ_l^ϑ must be referred to the set of o-d pairs: this calculation is performed by considering a matrix of the same dimensions of the dynamic sub-assignment map (i.e. referred to link counts only) deriving from the trajectory data sample (Λ^{traj}). The generic term of the Λ^{traj} matrix $\lambda_{odl}^{traj,\theta'\theta}$ is equal to one if the measured link flow entering the link *l* at time slice ϑ belongs to the path *k* connecting the o-d pair *od* whose corresponding o-d flow is generated at at the time slice θ' and equal to zero otherwise:

$$\lambda_{odl}^{traj,\theta'\theta} = \begin{cases} 1; \;\; if \; l \; \in k_{|od} \\ 0; \;\; otherwise \end{cases} \tag{5.15}$$

Consistently multiplying element by element each upscaling factor ζ_l^ϑ by the entries of the Λ^{traj} matrix and computing the average values for each column of the matrix, it is possible to obtain an upscaling factor per each o-d pair od and each time slice ϑ:

$$\xi_{od}^\vartheta = \frac{\sum_{l=1}^{nlc \cdot n\vartheta} \zeta_l^\vartheta \lambda_{odl}^{traj,\theta'\theta}}{nlc \cdot n\vartheta} \; \forall \, \vartheta \; \in T, \forall \; od \; \in OD \tag{5.16}$$

Clearly, the average values are calculated omitting the Λ^{traj} matrix entries equal to zero and all the entries such that the link flow deriving from trajectory data entering the link *l* at time-slice ϑ is equal to zero.

It is worth of notice that $\lambda_{odl}^{traj,\theta'\theta}$ are direct estimates of the time-depending path choice sets and path choice probabilities derived from the collected trajectory data sample: thus,

a significant bias is introduced in the upscaling method. To quantify and investigate this bias a set of laboratory experiments have been conducted to assess the accuracy level of direct estimates of assignment map entries (see Chapter 6).

5.2.4 **Goodness-of-fit measures**

In order to compare the results of the analysis, the *cvRMSE* indicator has been analysed as the sampling rates changes, whose expression is introduced in Section 3. The goodness of fit measure is calculated comparing the upscaled o-d flows, the corresponding link count flows, hold-out flows, all link flows, assignment map entries and upscaling factors to the ground-truth values. The outcomes of the experiments are described in detail in the following Section.

5.3 Experimental Results

The combination of all the possible values of the project variables shown in Table 5.4 have led to 384 different experiments testing each scaling method presented in Section 5.2.3. Overall, the total number of experiment instances (considering the number of replications) was equal to 11520. To simplify the analysis of the results, the experiments were classified by the trajectory sampling rate distribution and by the network sensors localization criteria; the classification yielded to twelve different scenarios, which are reported in the following table (Table 5.5):

Scenario	Sampling Rate Distribution	network sensors localization criterion	Upscaling Horizon/Method
U-MF-TH	Uniform	Max Flow	Time-Horizon based
U-RND-TH		Random	Time-Horizon based
U-MF-TS		Max Flow	Time-Slice based
U-RND-TS		Random	Time-Slice based
U-MF-LC		Max Flow	Link-Count based
U-RND-LC		Random	Link-Count based
OD-MF-TH	OD-based	Max Flow	Time-Horizon based
OD-RND-TH		Random	Time-Horizon based
OD-MF-TS		Max Flow	Time-Slice based
OD-RND-TS		Random	Time-Slice based
OD-MF-LC		Max Flow	Link-Count based
OD-RND-LC		Random	Link-Count based

Table 5.5 Scenarios summarizing the experimental plan of direct scaling performance analysis

Table 5.5 is used as reference throughout the description of the results. All the conclusions were deduced observing the trend of *cvRMSE* mean values varying the sampling rate and the number of link counts. The *cvRMSE* mean values were calculated among the results deriving from the multiple replications of the same experiment corresponding to each scenario listed above (Table 5.5). Results shown in this section refer to the case of large-scale network described in Section 5.2.1.2.

Conducting a structured and purpose oriented sampling process, thus assuming a uniform sampling rate distribution, the minimum sampling rate threshold above which the coefficient of variation of the *RMSE* calculated for o-d flows is less than 0.30 is equal to 35% . This value

is reached assuming a max flow link counts sample and a time-horizon based upscaling factor (see Table 5.6, Scenario ***U-MF-TH***).

Max Sampling Rate/ Scenario	***5%***	***10%***	***20%***	***30%***	***40%***	***50%***	***60%***	***70%***	***80%***	***90%***	***100%***
U-MF-TH	0.96	0.65	0.44	0.33	0.27	0.22	0.17	0.14	0.11	0.07	0.00
U-RND-TH	0.97	0.66	0.44	0.34	0.27	0.22	0.18	0.14	0.11	0.07	0.00
U-MF-TS	0.97	0.66	0.44	0.34	0.27	0.22	0.18	0.15	0.11	0.07	0
U-RND-TS	0.97	0.66	0.44	0.34	0.27	0.22	0.18	0.15	0.11	0.07	0
U-MF-LC	0.97	0.74	0.57	0.47	0.40	0.33	0.27	0.21	0.15	0.09	0.00
U-RND-LC	1.17	0.98	0.80	0.68	0.58	0.48	0.39	0.30	0.20	0.12	0.00
OD-MF-TH	1.74	1.53	1.37	1.32	1.17	1.03	0.88	0.75	0.60	0.43	0.06
OD-RND-TH	1.50	1.30	1.18	1.09	1.03	0.93	0.80	0.71	0.59	0.42	0.07
OD-MF-TS	1.76	1.55	1.37	1.32	1.17	1.02	0.88	0.74	0.60	0.43	0.06
OD-RND-TS	1.56	1.34	1.24	1.13	1.05	0.94	0.82	0.71	0.59	0.41	0.06
OD-MF-LC	1.45	1.23	1.11	1.03	0.98	0.87	0.77	0.67	0.53	0.37	0.05
OD-RND-LC	1.39	1.27	1.14	1.09	0.99	0.88	0.76	0.69	0.54	0.36	0.04

Table 5.6 cvRMSE mean values calculated comparing upscaled o-d flows and ground-truth values per each scenario and sampling rate considering 127 link counts

Each figure in the following reports four diagrams respectively showing the trend of *cvRMSE* values on o-d flows, link count flows, hold-out flows and all link flows arranged in clockwise order as the sampling rate varies on a scale ranging from 0 to 1 (100%). The colour of each line depicted in each diagram identifies the dimension of the link counts sample (i.e. blue for 63, red for 127 and yellow for 254 link count sections).

As illustrated in Figure 5.5, Figure 5.6, Figure 5.7, Figure 5.8, Figure 5.9 and Figure 5.10 the frequent overlapping of the *cvRMSE* on o-d flows trends suggests that the scenarios assuming a uniform distribution do not show any appreciable improvement in their outcomes augmenting the numerosity of the link count sample and varying link count sections locations when applying the time-horizon and time-slice based direct scaling methods. Conversely, using a different scaling factor per each link count section and per each time-slice leads to worse performances compared to the other scaling methods, which are likely to improve if a greater number of link traffic measurements is available (see Scenario ***U-MF-LC*** and Scenario ***U-RND-LC*** in Figure 5.9 and Figure 5.10 respectively).

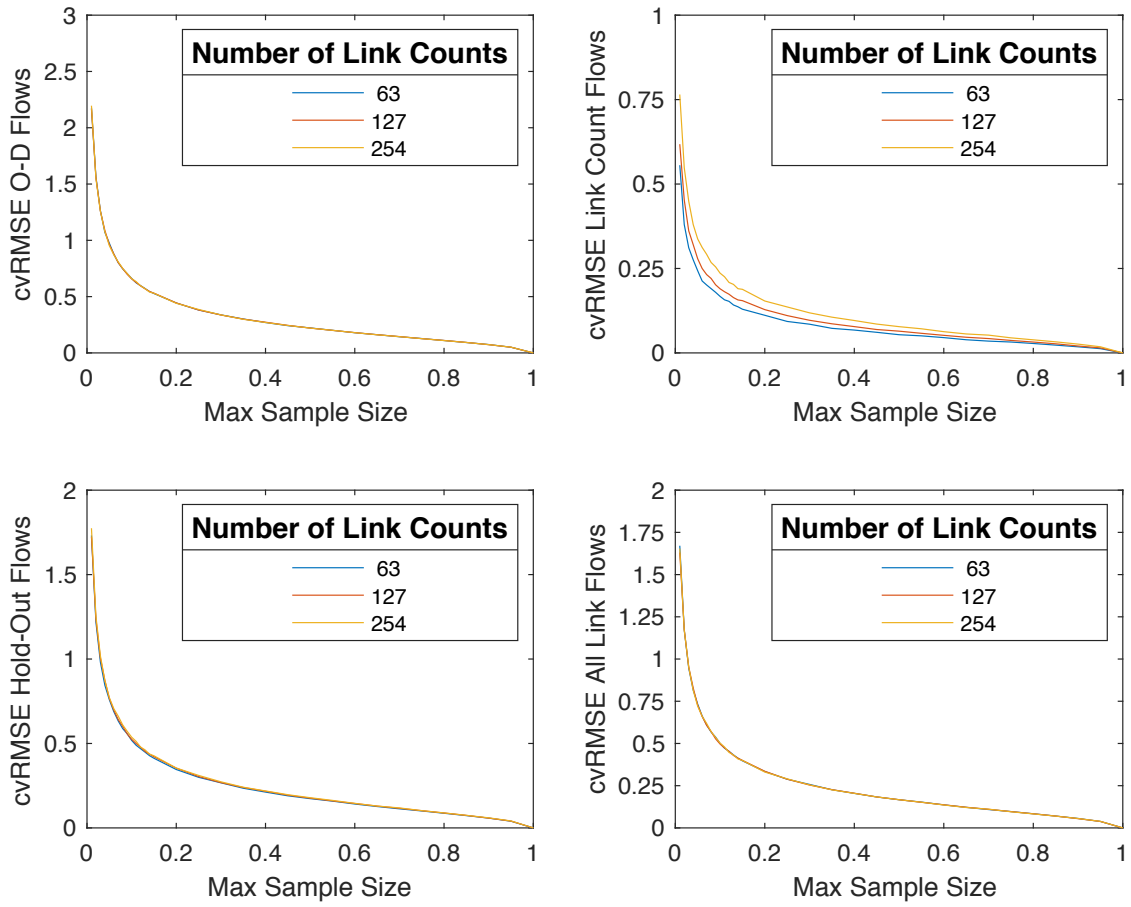


*Figure 5.5 Scenario **U-MF-TH**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

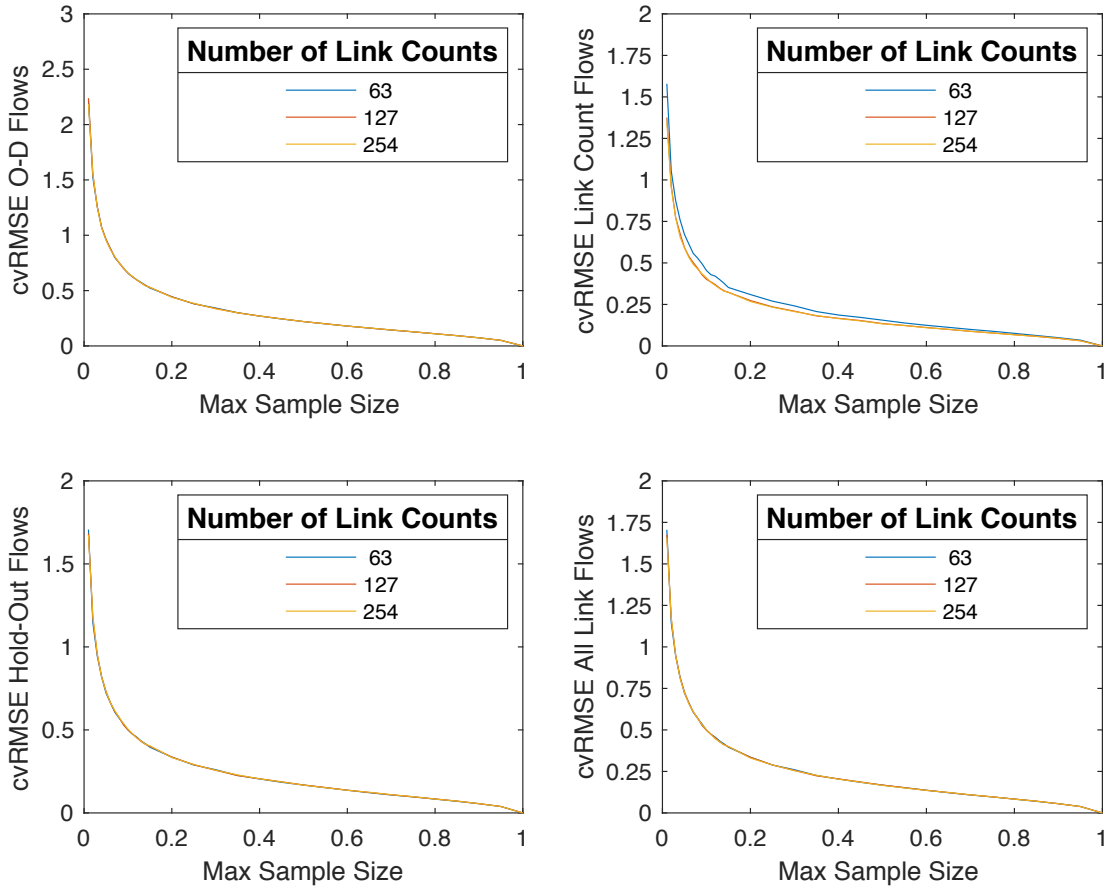


*Figure 5.6 Scenario **U-RND-TH**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

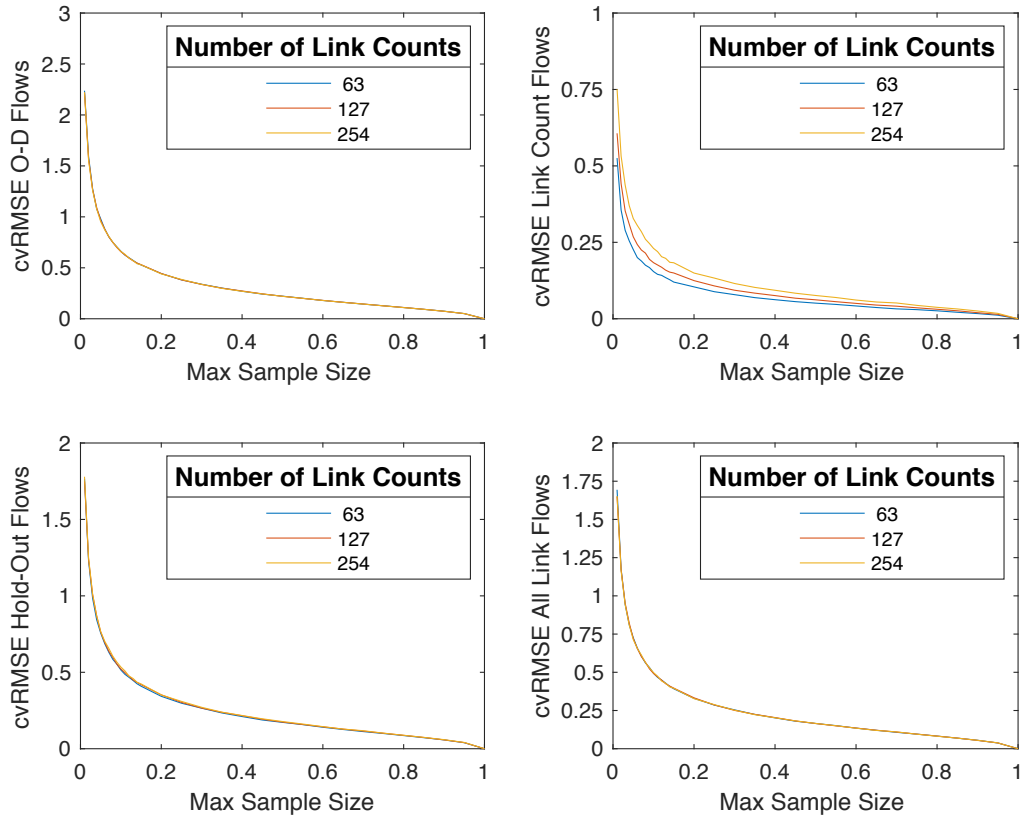


*Figure 5.7 Scenario **U-MF-TS**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

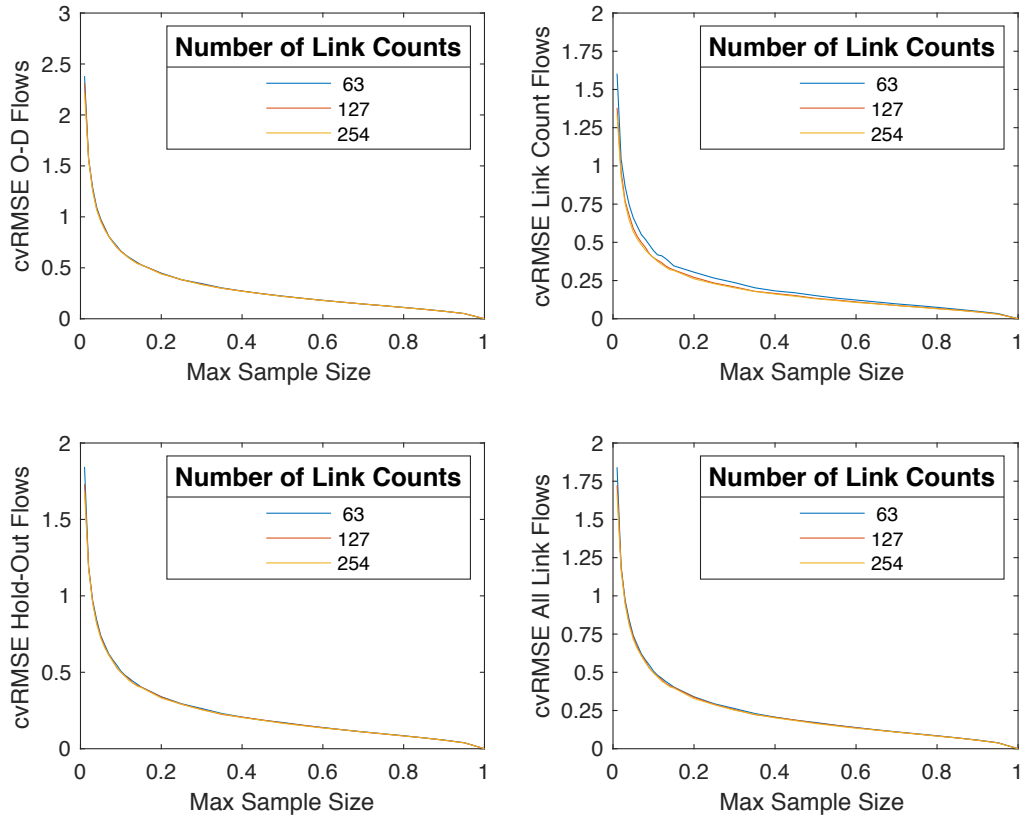


*Figure 5.8 Scenario **U-RND-TS**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

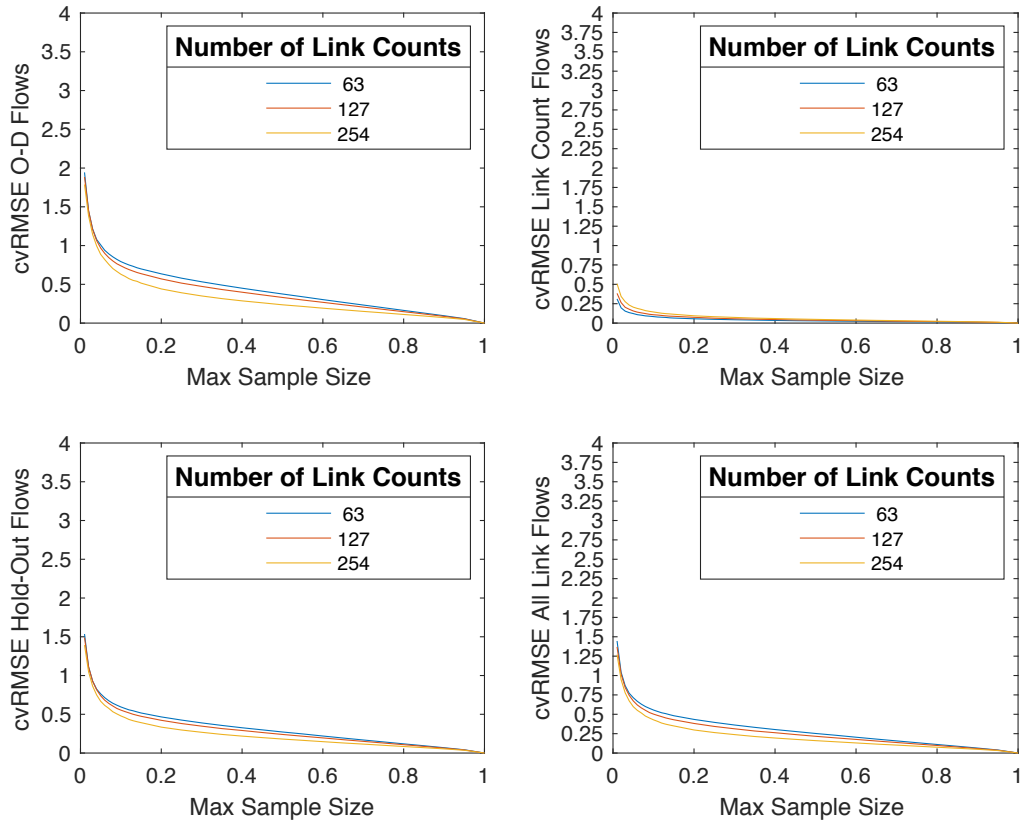


*Figure 5.9 Scenario **U-MF-LC**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

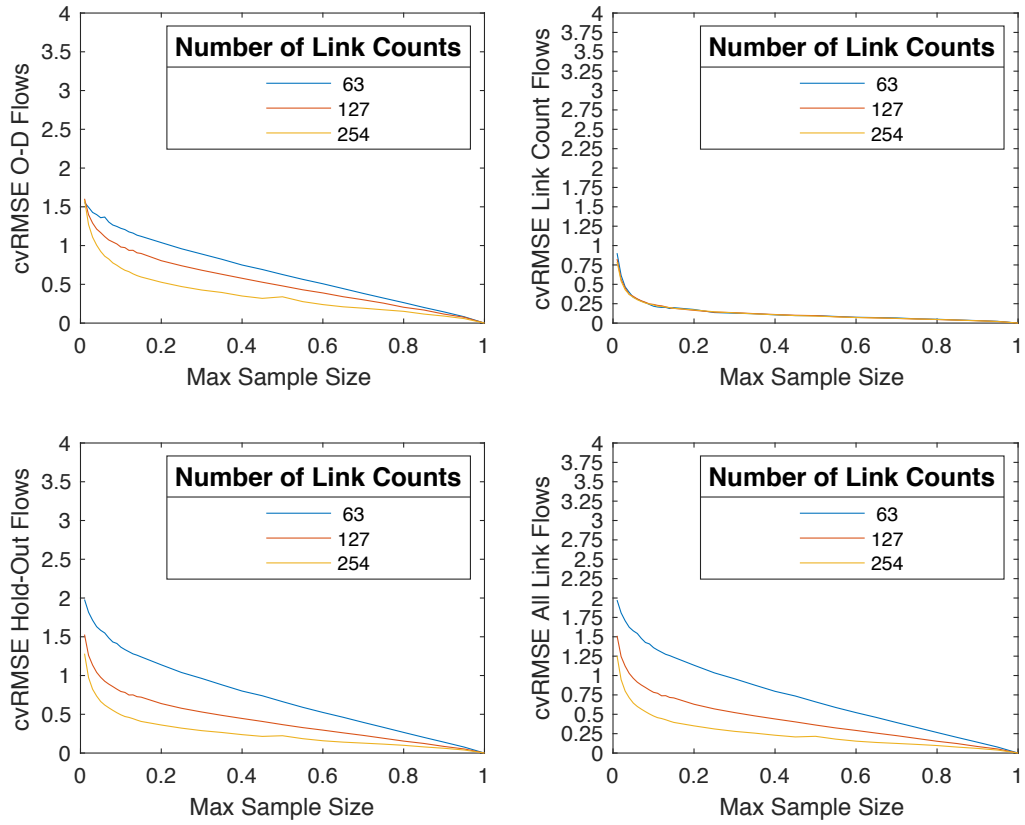


*Figure 5.10 Scenario **U-RND-LC**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

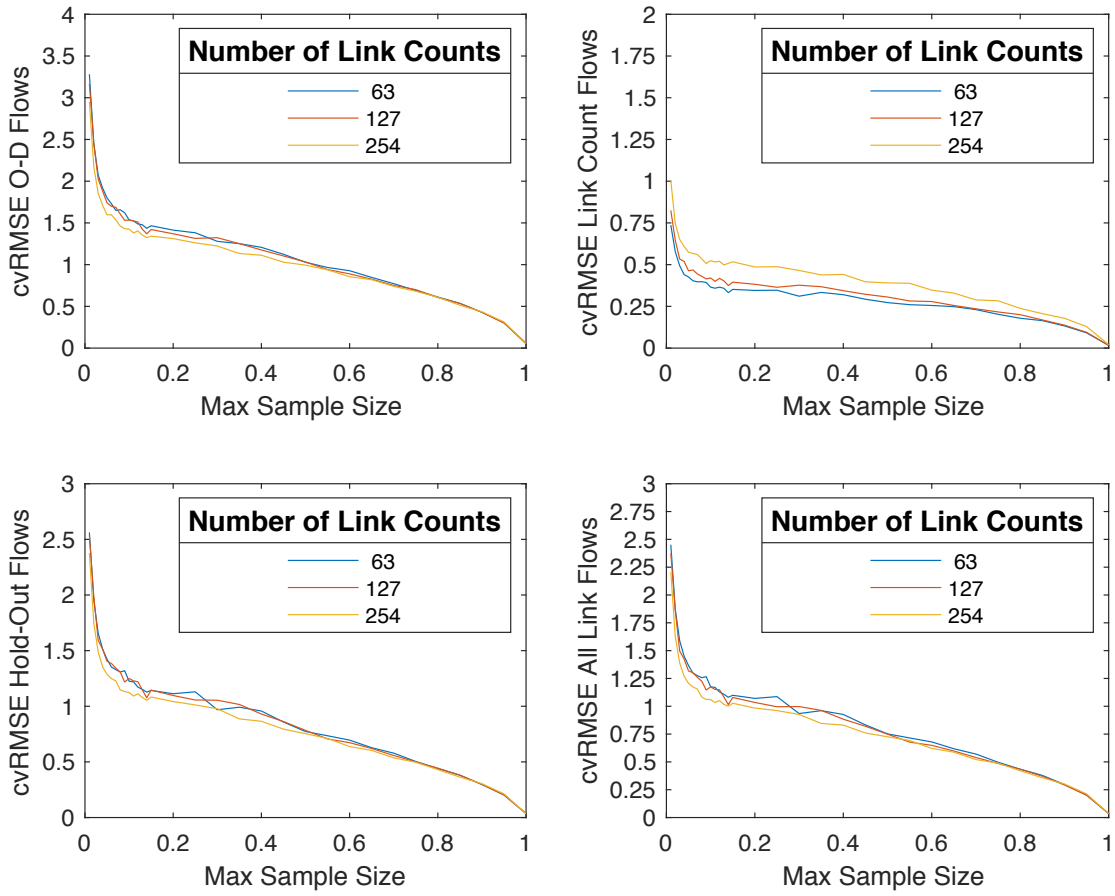


*Figure 5.11 Scenario **OD-MF-TH**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

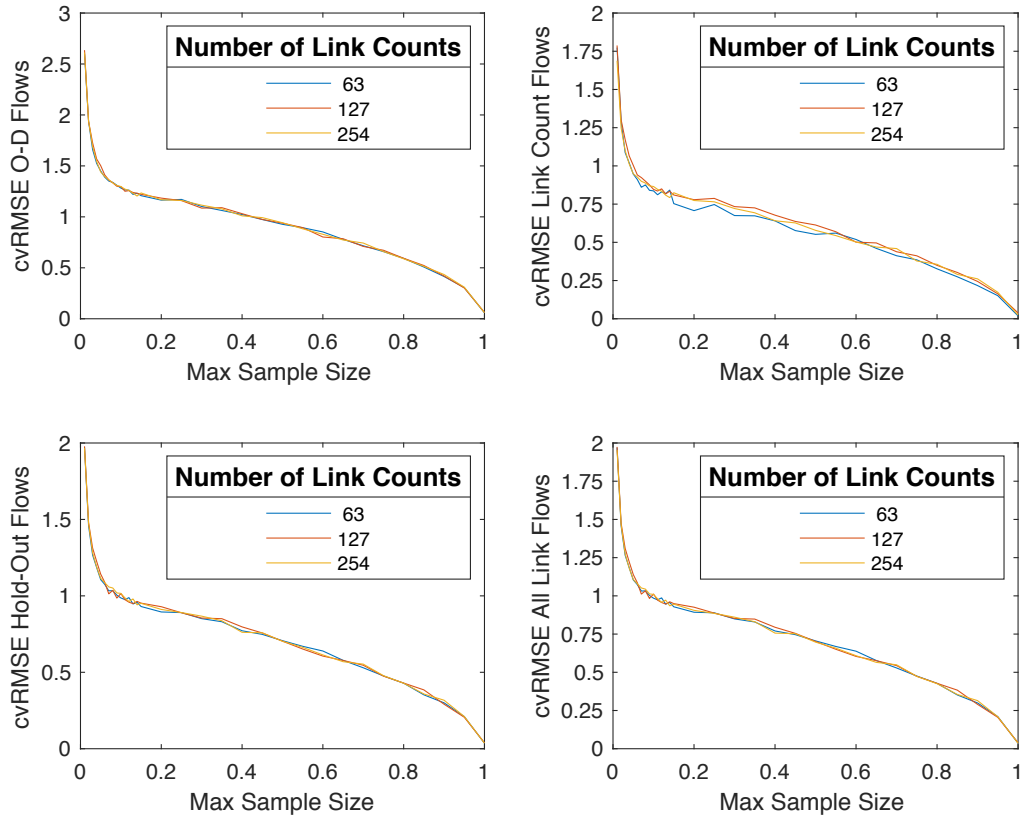


*Figure 5.12 Scenario **OD-RND-TH**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

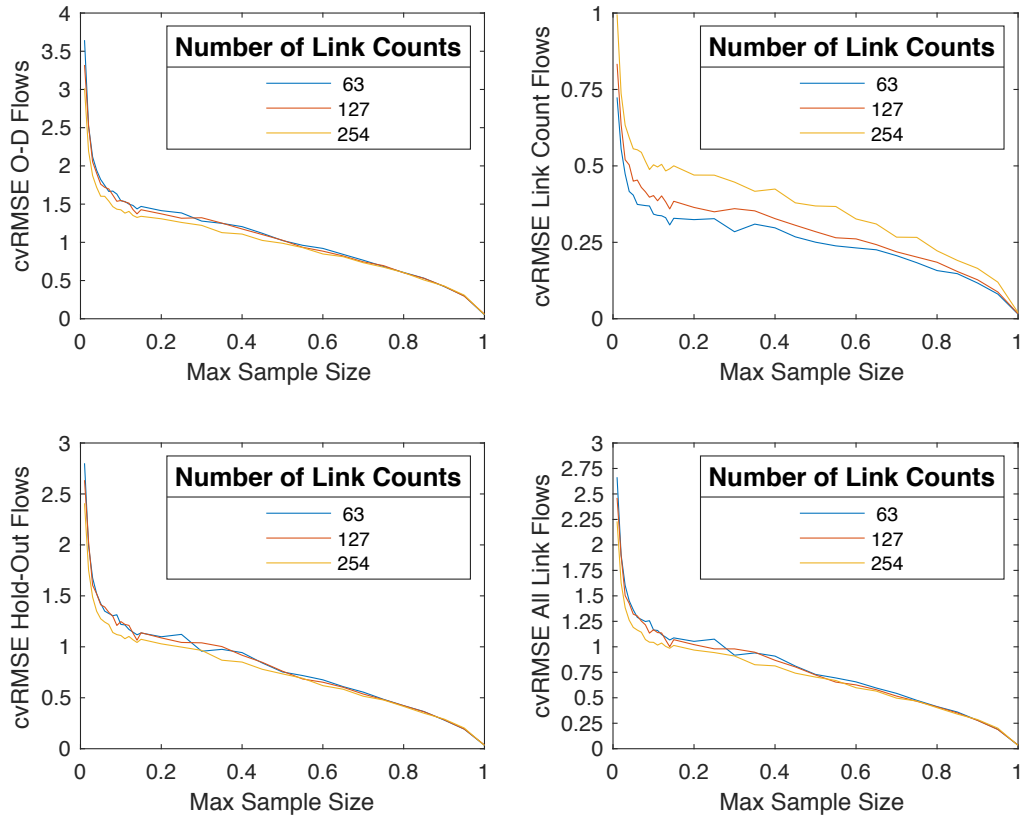


*Figure 5.13 Scenario **OD-MF-TS**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

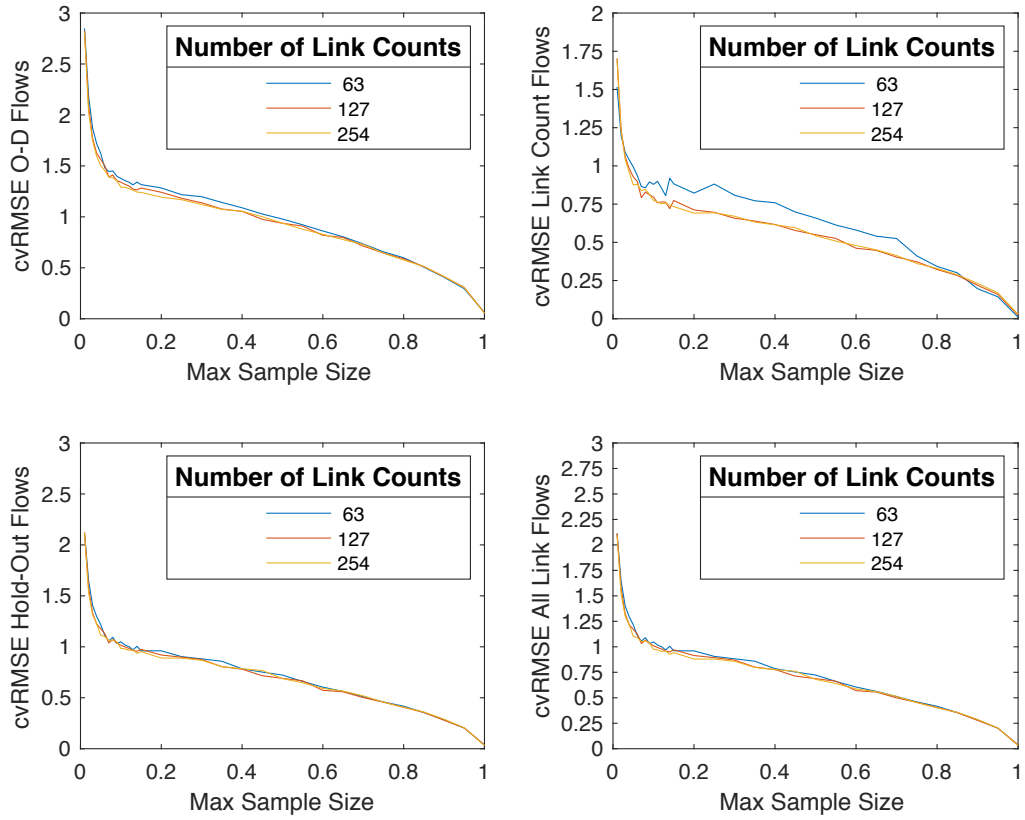


*Figure 5.14 Scenario **OD-RND-TS**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

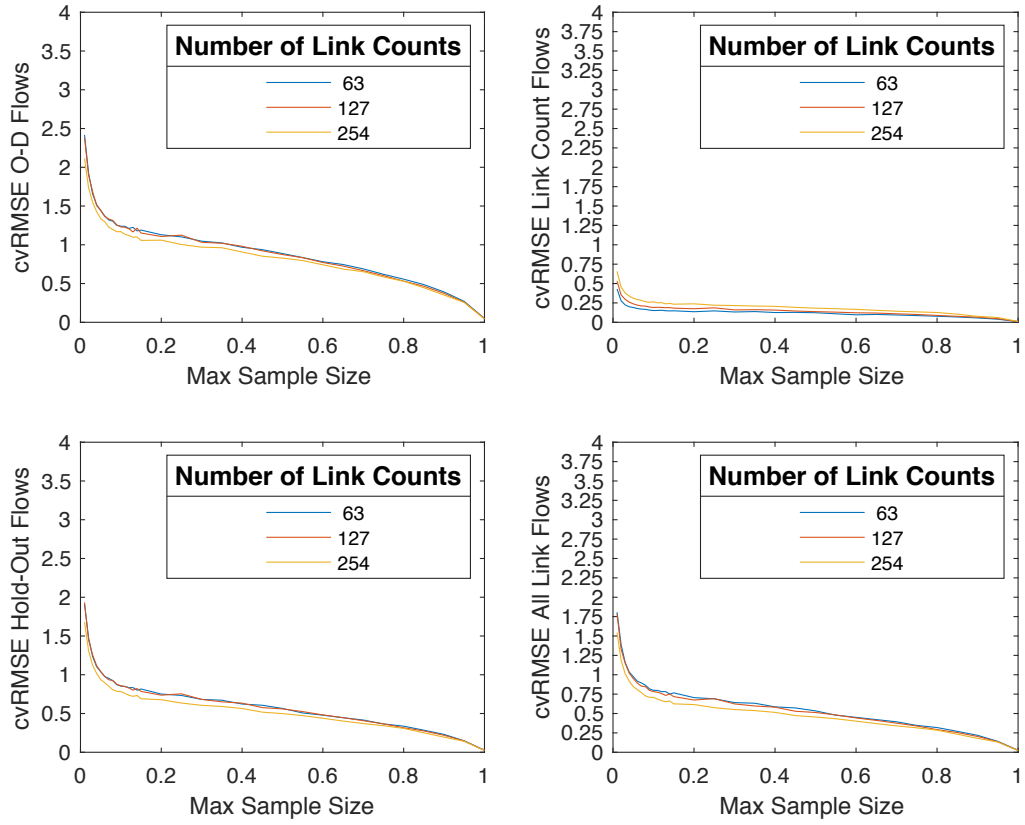


*Figure 5.15 Scenario **OD-MF-LC**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

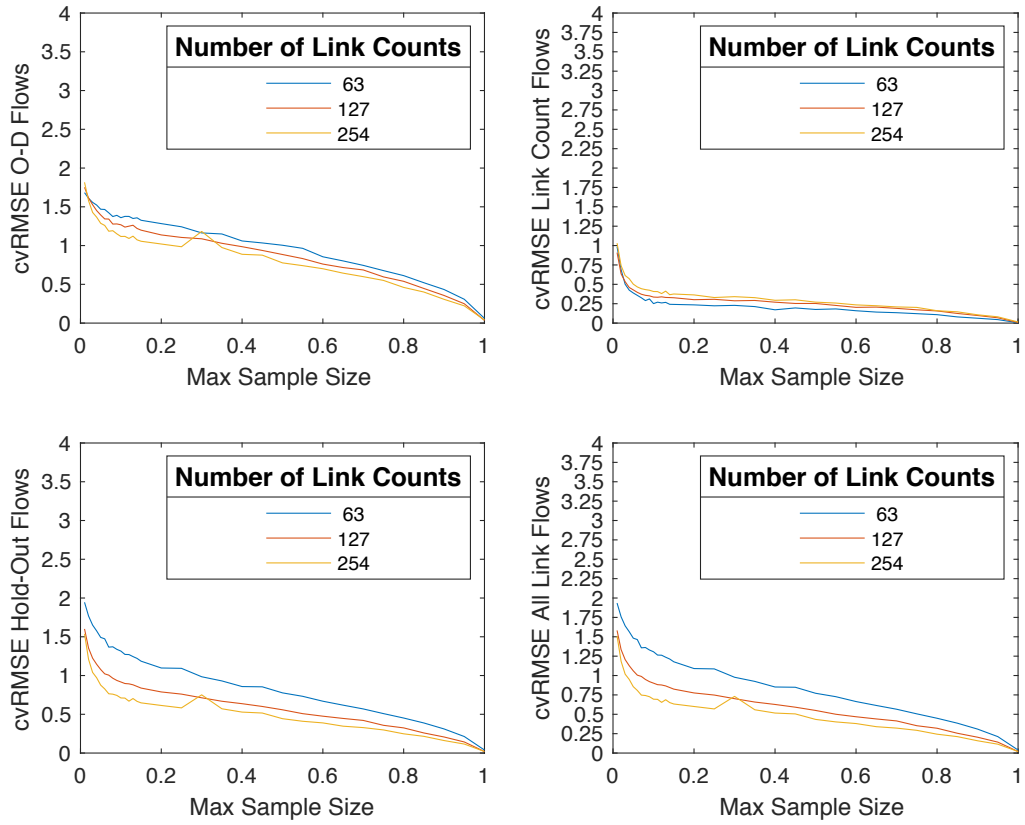


*Figure 5.16 Scenario **OD-RND-LC**: cvRMSE average values on o-d flows, link count flows, Hold-Out flows, All link flows (clokwise order).*

As expected, observing the results in Table 5.7, a random localization of the sensors negatively affects the results of *cvRMSE* trend calculated for link count flows compared to the scenarios assuming a max-flow localization criterion, regardless of all the values of the other project variables.

Max Sampling Rate/ Scenario	***5%***	***10%***	***20%***	***30%***	***40%***	***50%***	***60%***	***70%***	***80%***	***90%***	***100%***
U-MF-TH	0.28	0.19	0.13	0.10	0.08	0.06	0.05	0.04	0.03	0.02	0.00
U-RND-TH	0.59	0.40	0.27	0.21	0.17	0.13	0.11	0.09	0.07	0.05	0.00
U-MF-TS	0.27	0.18	0.12	0.09	0.08	0.06	0.05	0.04	0.03	0.02	0.00
U-RND-TS	0.59	0.40	0.27	0.21	0.17	0.13	0.11	0.09	0.07	0.05	0.00
U-MF-LC	0.16	0.11	0.07	0.06	0.05	0.04	0.03	0.02	0.02	0.01	0.00
U-RND-LC	0.35	0.25	0.18	0.14	0.11	0.10	0.08	0.06	0.05	0.03	0.00
OD-MF-TH	0.46	0.42	0.38	0.38	0.34	0.31	0.28	0.24	0.20	0.14	0.02
OD-RND-TH	1.00	0.85	0.78	0.73	0.68	0.61	0.50	0.44	0.35	0.24	0.04
OD-MF-TS	0.45	0.40	0.36	0.36	0.33	0.28	0.26	0.22	0.18	0.13	0.02
OD-RND-TS	0.93	0.80	0.71	0.66	0.62	0.55	0.46	0.40	0.32	0.22	0.02
OD-MF-LC	0.24	0.19	0.17	0.16	0.16	0.14	0.12	0.11	0.09	0.07	0.01
OD-RND-LC	0.43	0.34	0.30	0.29	0.27	0.25	0.20	0.19	0.16	0.10	0.01

Table 5.7 cvRMSE mean values calculated comparing link count flows by assigning upscaled o-d flows by trajectory and ground-truth values per each scenario and sampling rate considering 127 link counts

Regarding the scenarios assuming an o-d weighted sampling rate distribution, it is worth of notice that the *cvRMSE* on o-d flows corresponding to a sampling rate equal to 100% is not equal to zero due to the fact that the algorithm extracting trajectories from ground-truth values utilizes an approximation to obtain integer numbers (i.e. number of trajectories to extract per each o-d pair); this implies that the effective sampling rate deriving from the sampling process is lower compared to the value set by the user. In light of this, results are referred to the maximum possible sampling rate values. Comparing the scenarios assuming the different sampling rate distributions (uniform and o-d based), for the same value of the sampling rate threshold found for the uniform distribution cases, in the case of o-d weighted sampling rate distribution the value of the *cvRMSE* on o-d flows is tripled; specifically, applying the time-horizon and time-slice based direct scaling methods the value associated to the threshold is greater than one (Scenarios ***OD-MF-TH***, ***OD-RND-TH***, ***OD-MF-TS***, ***OD-RND-TS***), while slightly better results are obtained when applying a link-count based scaling method in which cvRMSE on o-d flows is around one (***OD-MF-LC*** and Scenario ***OD-RND-LC***). Indeed, a straightforward conclusion can be deduced from these results: the link count based direct scaling leads to the

best performances when dealing with opportunistic trajectory data (see Figure 5.17 for an overall comparison among all the scenarios).

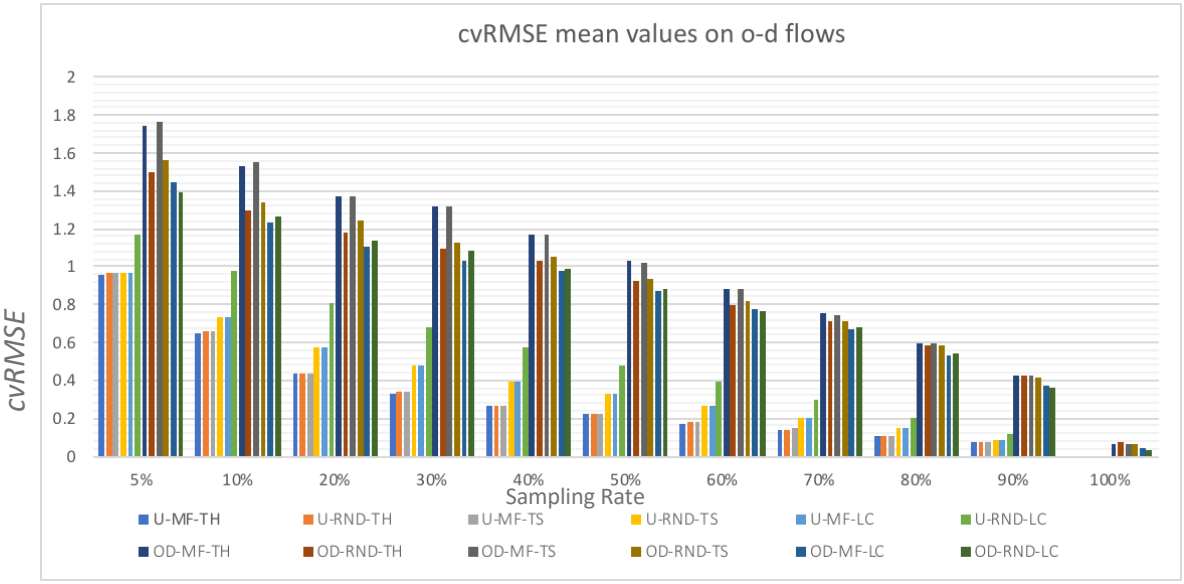


Figure 5.17 Overall comparison among scenarios: cvRMSE mean values on o-d flows (y axis) per sampling rate (x axis) for all scenarios considering 127 link counts

From an overall point of view, given the experimental results obtained from the trajectory data analysis described in Chapter 4, considering the average value of the penetration rate obtained applying Equation 4.6 (about 6%) and considering the range of trajectory sample penetration rate reported in literature ([0.1-10] %, FHWA 2016), it is evident that by using opportunistic trajectory data, direct scaling performances alone are not satisfactory to obtain reliable o-d flows estimates. In light of this, upscaled o-d flows need further updating methods to reach an acceptable level of accuracy. On the basis of this finding, an extended analysis has been carried out to eventually evaluate any significant improvement updating upscaled o-d flows by means of GLS-based estimators: indeed, in line with the experiments conducted by Yang et al. (2017) and Mitra et al. (2020), upscaled o-d flows can be used as prior o-d estimates and updated exploiting a set of link traffic measurements. As supported by numerous studies, reliability of the target matrix is crucial to ensure better performances of o-d updating methods, thus using a link-count based direct scaling method could be recommended in this phase. The complete analysis is reported in Chapter 7.

Another relevant aspect which required further analysis concerned the direct estimation of traffic assignment map. Indeed, Chapter 6 describes an extended analysis identifying and quantifying the effect of the errors affecting traffic assignment map estimation on traffic simulation and o-d estimation.

5.4 Conclusions

This chapter deals with the crucial issue of sample representativeness arising with o-d flows estimation methods specifically investigating the potential of direct scaling methods which provides a first estimate of o-d flows upscaling trajectory o-d flows. To understand how the level of sample representativeness and other characteristics of the trajectory sample affect direct o-d flows estimates accuracy, the performances of three direct scaling methods described in Section 5.2.3 are analysed by means of laboratory experiments; these methods have demonstrated to be an essential step for ameliorating o-d flows estimates derived by trajectory data. The proposed experiments, consistently with the benchmarking platform proposed by Antoniou et al. (2107), cover a wide experimental plan accounting for the range variability and variety of trajectory data sample and traffic measurements sample characteristics. Final considerations were deduced comparing upscaled o-d flows and related link traffic flows to the ground-truth values, therefore they might provide useful guidelines for researchers and practitioners dealing with various types of trajectory data sample and conducting o-d related applications. Specifically, conclusions can be referred to the two types of trajectory data classified according to the adopted trajectory data generation process: opportunistic data and purpose-oriented data (see Section 1.1.2).

The direct scaling performance analysis revealed that, when a sample of opportunistic data is available, to obtain reliable o-d flows estimates further updating of the upscaled o-d flows is essential. Indeed, acceptable errors on o-d flows can be attained only if the sampling rate reaches values greater that 80% (in the best case), which is way above the penetration of commonly available samples. Nonetheless, as also supported by literature (see Chapter 2), the upscaled o-d flows may constitutes a reliable starting point for updating procedures. This investigation is extensively developed in Chapter 7. Furthermore, given the central role of dynamic traffic assignment in o-d estimation/updating procedures, a further analysis is required to evaluate the reliability of traffic assignment map derived by trajectory (see Chapter 6). On the other hand, if the data collection process can be generated in a controlled environment, direct scaling methods can provide satisfactory results even with (relatively) low values of the sampling rates.

6 Laboratory experiments to assess the reliability of traffic assignment map

This Chapter is the revised version of the article: Simonelli F., Tinessa F., Marzano V., Papola A. and Romano A., "Laboratory experiments to assess the reliability of traffic assignment map," 2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2019, pp. 1-9.

6.1 Motivation and background

The assignment models, representing the relationship between travel demand and flows on the network, play a key role in both transport network planning and management applications. Although synthetic approaches have recently proposed, e.g. based on deep learning (Vlahogianni et al 2014, Csikos et al. 2015, Gallo et al. 2016), and/or based on data driven models (Oh et al. 2015, Chen et al. 2001, Clark 2003, Innamaa 2003, van Lint et al. 2002, Tak et al. 2014), an explicit representation and interpretation of the travel behavior is an indispensable tool for transport networks design and simulation and for the implementation of effective traffic management operations, especially when coping with unexpected changes of travel demand and/or supply, typical of non-recurring situations.

Although there are countless specifications of the assignment models for both steady-state and within-day dynamic contexts, their reliability and the statistical properties of their variables are rarely discussed in the literature. Specifically, the reliability of the assignment models depends upon the dispersion of the aggregate traffic measurements (e.g. the trace of the traffic measurements covariance matrix) and to possible distortions of the estimator itself, e.g. a direct and/or a model-based estimator might introduce inherent biases, while the statistical properties of these models refer to the statistical distribution of the assignment matrix, whose entries can be provided by a given assignment model and/or estimated from experimental data. Since in most cases information on the reliability and on the statistical properties of the assignment models is not available, the statistical properties of the assignment matrix are often neglected assuming the assignment matrix entries as deterministic variables and assuming “error-free” traffic measurements. Therefore, neglecting both the assignment matrix simulation error (i.e.

the error deriving from approximating the real phenomenon) and the error affecting traffic measurements implies remarkable impacts on the representation of travellers' route choice and consequently on the reliability of the o-d matrix estimation process.

As mentioned in Chapter 2, concerning the o-d estimation models based on traffic measurements, several assumptions have been proposed to represent the relationship between travel demand and traffic measurements. The diverse assumptions on the statistical properties of travel demand, traffic measurements and the assignment models lead to different estimators such as minimum information/maximum entropy (Van Zuylen and Willumsen, 1980), maximum likelihood (Bell, 1983), generalized least squares (GLS) (Cascetta, 1984) or Bayesian (Baher, 1983). The approaches proposed in steady state conditions have been extended to the within day dynamic context (Cascetta et al. 1993) and models suitable for online applications have been introduced (Ashok 1996, Ashok and Ben-Akiva 200), while other works deal with the use of other traffic measurements (Antoniou et al. 2006) or address algorithmic aspects related to the computational efficiency (Cipriani et al. 2011, Lu et al. 2015). A comprehensive literary review on o-d estimation methods is reported in Chapter 2.

Another research field in which the hypothesis of error-free assignment matrix can strongly affect the results is strictly related to o-d matrix estimation applications and deals with the NSLP (Network Sensor Location Problem), which consist of defining the optimal locations of the sensors to collect the traffic measurements used to estimate the o-d flows (Yang and Zhou 1998, Gan et al. 2005, Bierlaire 2002, Zhou and List 2010). Nevertheless, even when a random error is introduced in the measurement equation, implying a stochastic relation between the problem variables and the measurements, the random term is generally related to the measurements error due to sensor failure, while the effect of the assignment matrix simulation error is usually neglected. This aspect will be addressed by means of laboratory experiments with reference to a hypothetic but realistic case study, to assess the reliability of the assignment map derived from a trajectory data based survey and/or model calibration. The study aims at identifying and quantifying the effect of the errors affecting traffic assignment map estimation on traffic simulation and o-d estimation.

6.2 Methodology

6.2.1 **Description of the experiments**

The assignment model involves a non-linear route choice model in the problem variables and a flow propagation model which, in the case of uncongested network or fixed travel times, can be expressed by means of a linear relationships between o-d flows and traffic network variables.

Route choice models' state of the art embraces several decades of literature since the first specific application of the Multinomial Logit (MNL) model for this purpose (Dial 1971). Successive contributions proposed the application of different random utility models to route choice, with the aim of obviating the limitations of the MNL, mainly due to the I.I.A (Independence of Irrelevant Alternatives). property and the impossibility to reproduce the substitution pattern across the alternatives. A well- established assumption on route choice correlations is due to Daganzo and Sheffi 1977, which directly relates the correlations among the unobservable components of the route utilities to their physical overlapping, identifying in the Multinomial Probit (MNP) model the perfect model for accommodating this target.

Unfortunately, the MNP does not entail a closed-form expression of the choice probabilities, thus it requires burdensome simulations to be performed. In light of this, several authors proposed the application of other route choice models, following three main research directions. The former accounts for the physical overlaps among the routes by introducing a deterministic correction penalty factor to compute the systematic utilities of overlapping routes (see the C-Logit (Cascetta et al. 1996, Russo and Vitetta 2003, Zhou et al. 2012) and the Path-Size Logit Ben-Akiva and Ramming 1998, Ramming 2001, Hoogendoorn-Lanser et al. 2015, Bovy et al. 2009). The second one introduces different distribution shapes for the random term of the perceived utility. The distributions are mainly GEV type, allowing for closed-form expressions for choice probabilities (see the Cross Nested Logit: Vovsha and Bekhor 1998, Prashker and Bekhor 1998, Bekhor and Prashker 2001, Prashker and Bekhor 2004, Bekhor et al. 2008; the Pair Combinatorial Logit: Gliebe et al. 2009; the Network GEV: Papola and Marzano 2013) and for both choice probabilities and correlations (see the Combination of Nested Logit model: Papola 2016, Tinessa et al. 2017, Papola et al. 2018). The third research direction involves the application of the Mixed Logit for route choice (Bekhor et al. 2002, Frejinger and Bierlaire, 2009) which does not avoid integral simulation, but allows for simplifying it. Recently, different paradigms inheriting their theoretical background from other disciplines are taking hold (see the recursive models derived from the dynamic programming: Fosgerau et al. 2013, Mai et al. 2015, Mai 2016; the mental representation items e.g. Kazagli et al. 2016; the machine learning e.g. Yang et al. 1993 or the fuzzy logic: e.g. Lotanand and Koutsopoulos, 1993). However, their applications are currently a few and their properties need to be further investigated.

Considering the non-linearity, a theoretical approach to estimate the dispersion of the route choice model outputs appears cumbersome; therefore in this study, this dispersion can be numerically evaluated by means of laboratory experiments which allow for operating in a controlled environment. Specifically, with reference to a real context concerning the network

topology, the travel demand and the flow propagation, several populations of travellers are generated according to diverse route choice models. Each corresponding model is in turn calibrated on the basis of several samples of individuals extracted from the population according to different sampling rate. Based on the assumed data collection technique, the sample may represent a set of users whose entire route is known or alternatively, a set of floating cars with partial location data collected along their route.

In light of this, the sampling rate can be varied both in terms of percentage of users detected from the entire population and in terms of complete or partial coverage of the path followed. Given a sample of users travelling through the network and known their corresponding o-d pair and route, the assignment matrix could be inferred by considering the percentage of users in the sample choosing a certain route, therefore the estimation is performed without using any model or behavioural hypothesis. Depending upon the sampling rate, this kind of estimation which is named in the following "direct estimate", does not guarantee the full o-d pairs and network coverage, i.e. if no observation of user travelling between some o-d pair or using some link is collected in the sample, there is no information about the assignment submatrix related to that o-d pair or, in the latter case, the link flow estimate on that link is equal to zero. This aspect highlights the need of using a model, capable of balancing missing information in the sample. For each sampling rate, several draws and calibrations are carried out to numerically evaluate the dispersion of the route choice model parameters (i.e. those entering its underlying utility function) and especially of its outputs (route choice probabilities) which, combined with flow propagation models, provide the assignment matrix. In this context, it is also possible to evaluate to which extent the errors affecting the assignment model impact on link flows estimation, given the true travel demand and/or vice versa, on the o-d flows estimation when a set of link counts is available.

Furthermore, the effect of the spatial disaggregation level of the travel demand can be investigated. Specifically, given a higher level of spatial aggregation (e.g. larger traffic analysis zone system), it is possible to define the extent to which the corresponding aggregated travel demand and the corresponding routes, limited to hierarchically higher links in the network, lead to a greater or lower level of uncertainty of the assignment matrix, which in turn affects the link flows estimation process and the o-d matrix estimation process based on traffic measurements.

6.2.2 **Experimental Setup**

The laboratory experiments were carried assuming the steady state conditions with reference to the real network of the Caserta province, depicted in Figure *5.3*. The experimental setup are described in detail in Section 5.2.1.2. To model the route choice of each traveller of the

population, the random utility model based on the generalized perceived route costs described in Section 5.2.1.2 has been adopted.

6.2.3 **Direct Estimates**

Given a random sample *k* of users, the assignment matrix M_a^s could be estimated by considering the percentage of users in the sample travelling between each o-d pair *s* that uses the link *a*. Formally:

$$M_{k,a}^s = \frac{n_{k,a}^s}{n_k^s} \tag{6.1}$$

being n_k^s the number of users in the sample *k* travelling between the o-d pair *s*, while $n_{k,a}^s$ the number of users included in the sample *k* travelling between the o-d pair *s* and using the link *a*; The total sampled link flows $f_{k,a}$ are obtained as:

$$f_{k,a} = \sum_s n_{k,a}^s \tag{6.2}$$

It is worth noting again that, if no users travelling between some o-d pairs are detected in the sample, the direct estimation of the corresponding assignment submatrix cannot be performed. In other words, it is assumed that a set of users is traced along their journey and, for each sampling rate, the set is randomly extracted regardless of the o-d pairs (each o-d pair has the same probability of extraction from the entire set of o-d pairs). Moreover, two sampling scenarios are considered: the former assumes a full link coverage of observed routes, whilst the latter assumes a GPS polling frequency leading to partial route coverage.

6.2.4 **Models estimate**

Part of the study aims at comparing the results of the direct estimation with the outcomes of the MNL and the C-Logit models. Therefore, an explicit path enumeration has been performed by generating a choice set with the double random method.

The estimation of the model parameters is performed through the maximum log-likelihood method, expressing the likelihood in terms of reproduction of observed link choice probabilities by the model.

The general explicit formulation for the route choice probability is:

$$p(r) = \frac{\exp\left(\beta_t \cdot t_r + \beta_c \cdot c_r + \beta_{CF} \cdot CF_r\right)}{\sum_{r'} \exp\left(\beta_t \cdot t_r + \beta_c \cdot c_r + \beta_{CF} \cdot CF_r\right)} \tag{6.3}$$

Where β_{CF} is the coefficient of the commonality factor CF_r, defined according to Cascetta et al. 1996 as:

$$CF_r = \ln \sum_{r'} \left[\frac{C_{rr'}}{\sqrt{C_r \cdot C_{r'}}} \right]^{\gamma} \tag{6.4}$$

being $C_{rr'}$ the sum of link costs which are common to r and r'.

Noticeably, substituting $\beta_{CF} = 0$ in Eq. (6.3), the basic MNL formulation can be obtained.

The generic assignment matrix element reproduced by such models can be obtained as:

$$M_{k,a}^{s} = \frac{f_a^s}{d_s} \tag{6.5}$$

being f_a^s the link flow obtained by the network loading of the demand d_s.

6.3 Results

In order to compare the results of the direct estimate and the route choice model, the Mean Absolute Percent Error (*MAPE*) and the Root Mean Square Error (*RMSE*) indicators are shown according to different values of the sampling rates; the indicators have been calculated with reference to the single $M_{k,a}^{s}$ values and to the link flows f_a. Specifically, the figures illustrate the trend of the average value and the coefficient of variation of the MAPE and the RMSE indicators, computed on 100 experiments for each value of sampling rate.

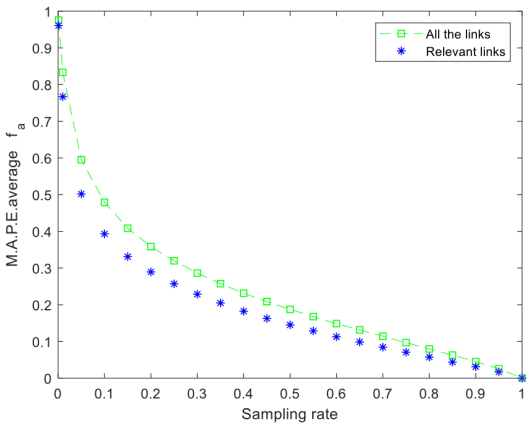


Figure 6.1Average values of MAPE on link flows..

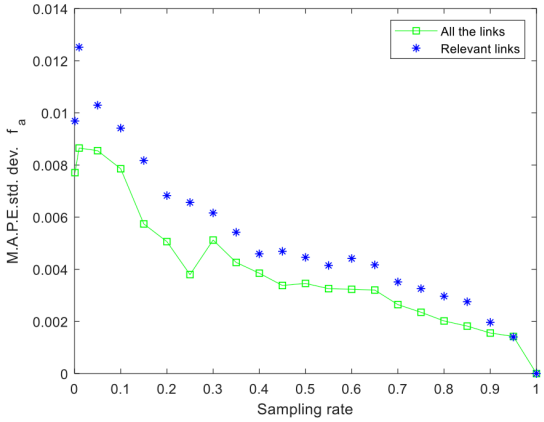


Figure 6.2 Standard Deviation of MAPE on link flows.

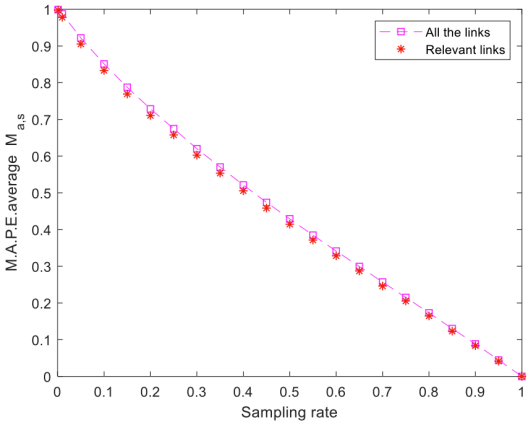


Figure 6.3 Average values of MAPE on Assignment map entries.

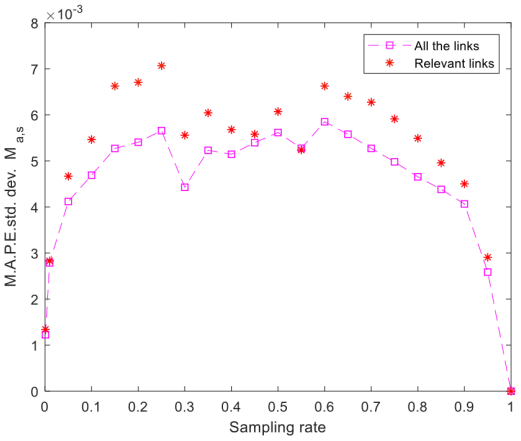


Figure 6.4 Standard Deviation of MAPE on Assignment map entries.

As can be seen in Figure 6.1 the MAPE on the link flows reaches average values in the range between 0.40 to approximately 1, for values of the sampling rates lower than 10%, although it rapidly decreases with an upwards concavity. A sampling rate of about 30% should be reached to observe errors on the aggregated flows of 30% on average, while a sampling rate lower than 20% is enough to obtain an acceptable error for the estimation of the relevant links.

Conversely, concerning the assignment map entries, the concavity is less pronounced (the trend is approximately linear), and only great values of the sampling rates can guarantee an acceptable estimation error, as shown in Figure 6.2. Furthermore, no substantial differences are observed when considering only relevant links. However, the errors on the assignment map entries seem

to auto-compensate for low sampling rates: for example, a value of 20% of the sampling rate corresponds to an average value of the MAPE equal to 70% on the assignment map entries, while the value in terms of aggregated link flows on relevant links is smaller than 30%.

The link flows error trend is further confirmed by the RMSE values shown in Figure 6.5.

In this case, greater errors observed on relevant links can be easily justified by the fact that, unlike the MAPE, the RMSE draws upon absolute values and relevant links are generally interested by the major link flows on the network. However, the errors are smaller in terms of percentage, as shown inFigure 6.3. The standard deviation on link flows tends to decrease as the sampling rate increases, suggesting an asymptotical stability of the link flow estimator, which implies an increasing reliability of the link flows estimates as the sample dimension augments (Figure 6.2 and Figure 6.6).

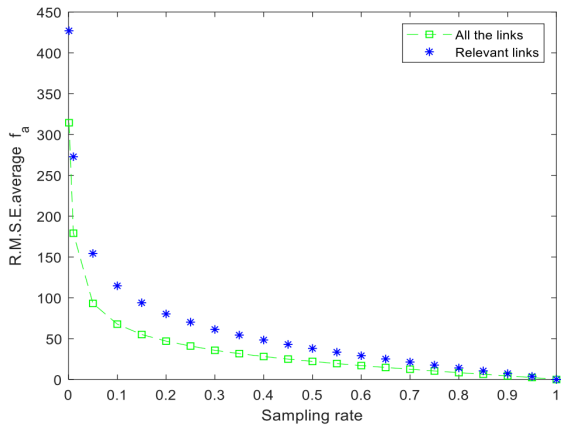


Figure 6.5 Average values of RMSE on link flows.

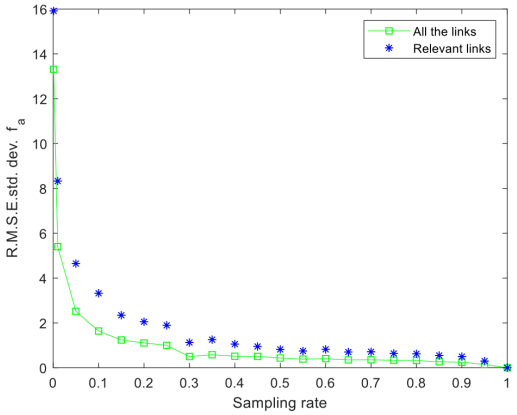


Figure 6.6 Standard Deviation of RMSE on link flows.

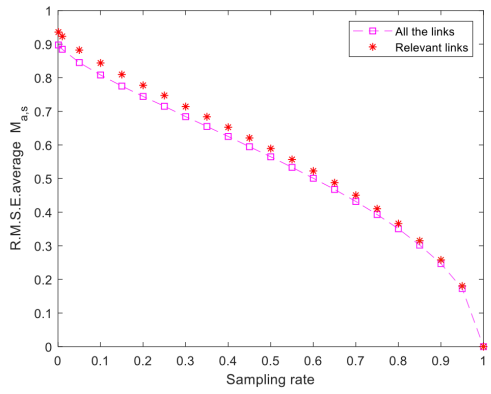


Figure 6.7 Average values of RMSE on Assignment map entries.

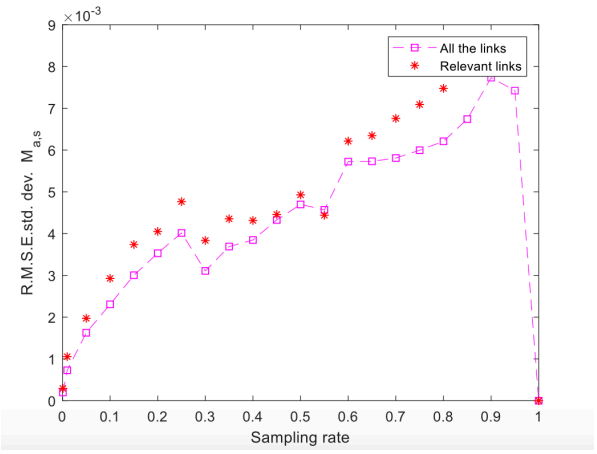


Figure 6.8 Standard Deviation of RMSE on Assignment map entries.

As introduced in Section 6.2.1, an interesting analysis has been carried out evaluating the effect of zonal (or o-d pairs) aggregations on direct estimates, in order to assess the extent to which link flows estimates and assignment entries estimated are affected by the spatial discretization of origins and destinations.

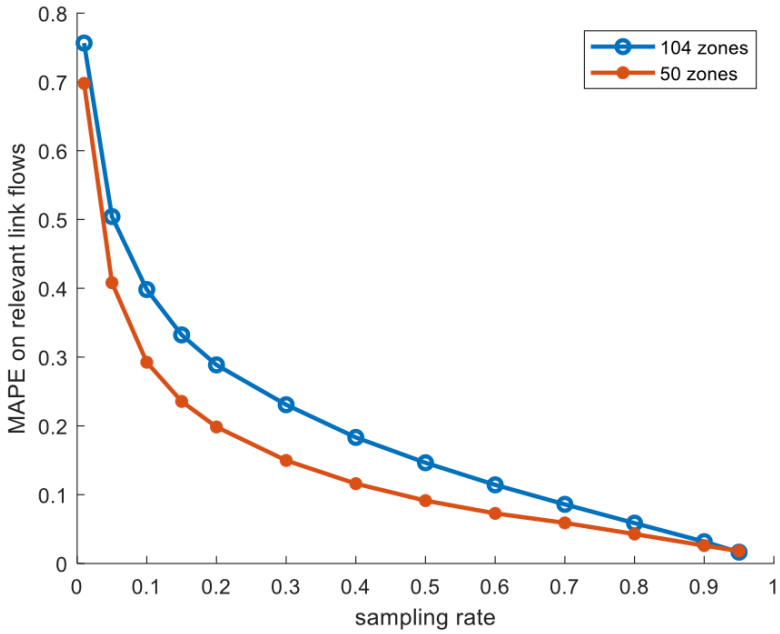


Figure 6.9 Average value of the MAPE on link flows for aggregation levels.

To this aim, a 50-zones zoning has been defined from the initial 104-zones zoning.

Results are reported in Figure 6.9, showing a noticeable improvement – around 10% on average – in the capability of reproducing relevant link flows. In order to assess the effect of the errors in matrix assignment entries on o-d matrix estimation updating based on link flows measurements, further experiments have been carried out. Considering a set of 500 link counts and the assignment map derived from a direct estimation over the entire range of variability of the sampling rate, the a priori o-d matrix has been updated by means of a GLS estimator.

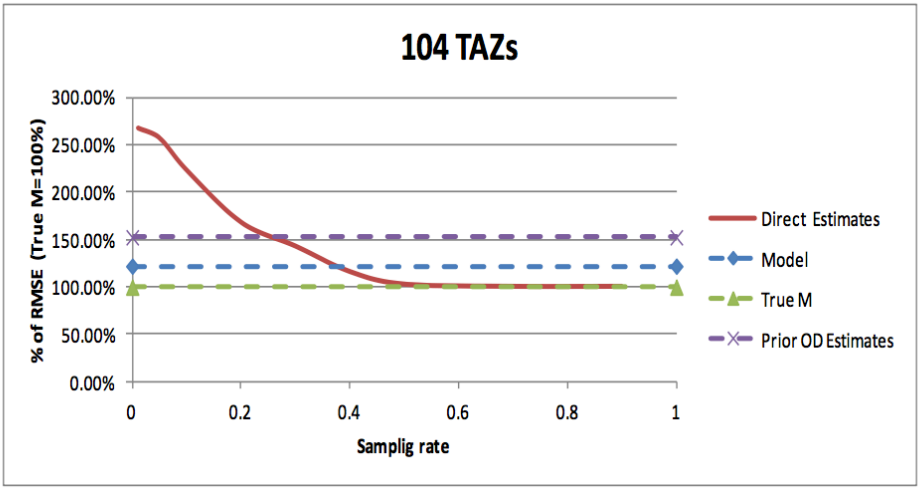


Figure 6.10 percentage error of RMSE on link flows – benchmark: RMSE obtained by using true assignment map in the o-d matrix updating (104 zones)

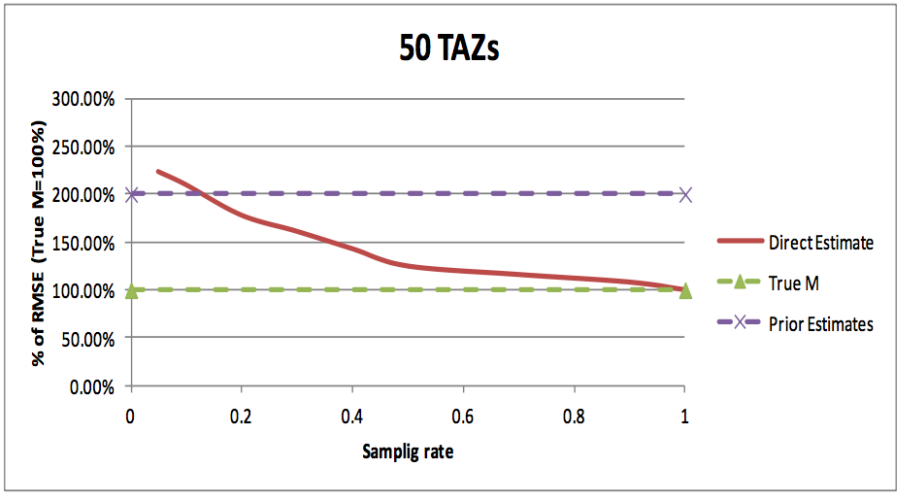


Figure 6.11 percentage error of RMSE on link flows – benchmark: RMSE obtained by using true assignment map in the o-d matrix updating (50 zones)

Figure 6.10 shows the trend of the percentage error of RMSE on link flows computed considering as benchmark the value of the RMSE obtained by using the true assignment map in the o-d matrix updating process. The RMSE decreases up to a value of 50% of the sampling rate, nearly reaching the value obtainable using the true matrix assignment in the updating process. Interestingly, for small sampling rates, the error exceeds the RMSE value associated to the prior o-d matrix (e.g. less than 25%). This result proves that the o-d flows estimates obtained using direct estimates of the assignment map relative to small sampling rates may report higher errors than the prior o-d estimate, which usually is obtained by survey and/or a model. This consideration holds for a smaller range of the sampling rate (e.g. less than 15%)

in the case of a higher level of spatial aggregation (50 zones), as illustrated in Figure 6.11, showing the same trend as Figure 6.10.

Concerning the model-based estimates (MNL and C- Logit models), trends are totally different. Indeed, the values of MAPE and RMSE for both models do not appreciably depend upon the sampling rate. Conversely, when comparing their performance only on relevant links, the MAPE significantly decreases for both models. Regarding the assignment map, differently from the direct estimate case, a reduced error occurs when considering only relevant links (Figure 6.12 and Figure 6.16); furthermore, results on the RMSE confirms this trend (Figure 6.14 and Figure 6.18). These outcomes can be substantially explained by the design of experimental setup, wherein the generation of the synthetic population and the specification of the route choice model utility are inherently consistent: this setup has been motivated in order to demonstrate that, even in ideal conditions, a model-based estimation tends to perform poorly, regardless of the sampling rate. In general, a comprehensive assessment of this aspect will require a more exhaustive set of experiments, which is left as research prospect. However, an interesting tendency seems to appear: a direct-based estimate requires a significant number of sampled trajectories, unlikely available even considering the current market penetration of sensing devices. In addition, for relatively low sampling rates (i.e. 0.4 – 0.5) and in optimal conditions (i.e. a synthetic population consistent with a simple choice context), model-based estimates yield comparable results with direct-based estimates, albeit both high in absolute terms. Furthermore, comparing the results on MAPE of direct estimate with the ones of MNL (Figure 6.12) and CL (Figure 6.16), to obtain a comparable error among the two procedures, the direct estimate involves a sampling smaller than 10%.

Concluding, reliable estimates can be only obtained by exploiting trajectory samples with substantial values of the sampling rate (about 20% to have mean percentage errors of 20% on aggregated measurements, e.g. link flows), which by far, are still outside the usual practical possibilities (considering this population, a sampling rate of 20% means more than 10’000 observed trajectories).

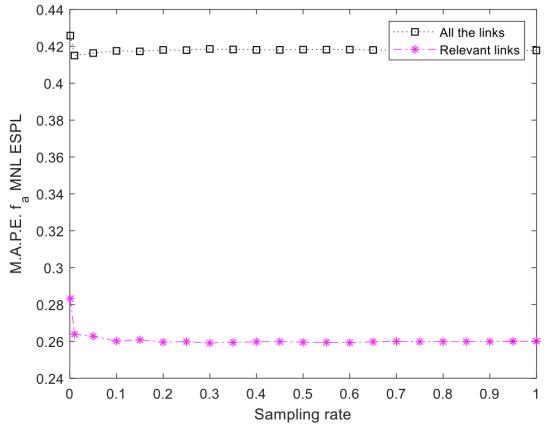


Figure 6.12 MAPE on link flows (MNL with explicit path enumeration).

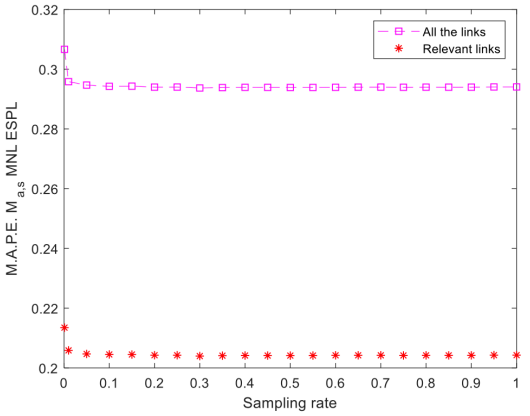


Figure 6.13 MAPE on the assignment map entries (MNL with explicit path enumeration).

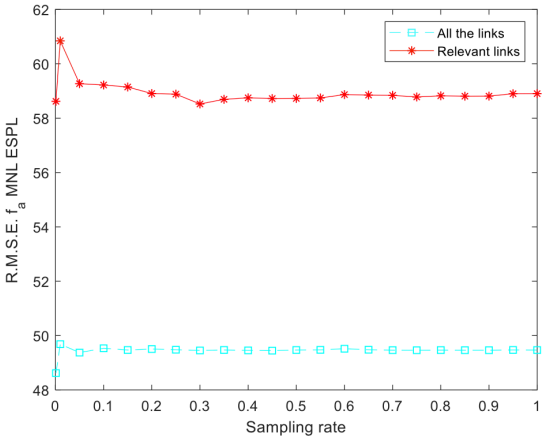


Figure 6.14 RMSE on the link flows (MNL with explicit path enumeration).

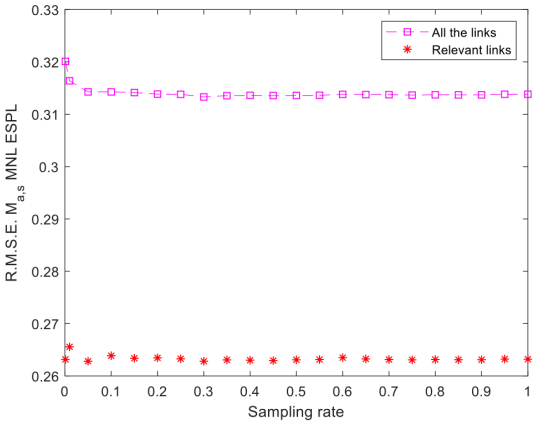


Figure 6.15 RMSE on the assignment map entries (MNL with explicit path enumeration).

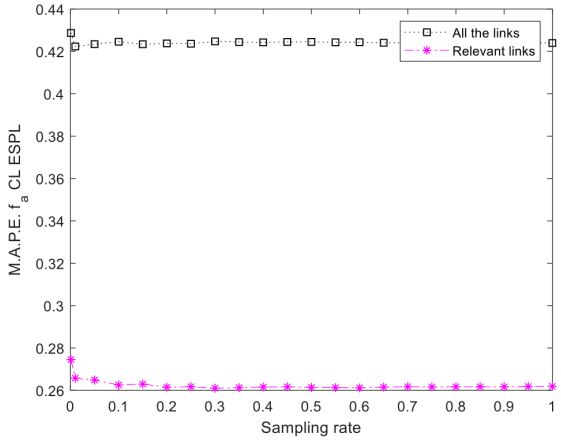


Figure 6.16 MAPE on link flows (CL with explicit path enumeration).

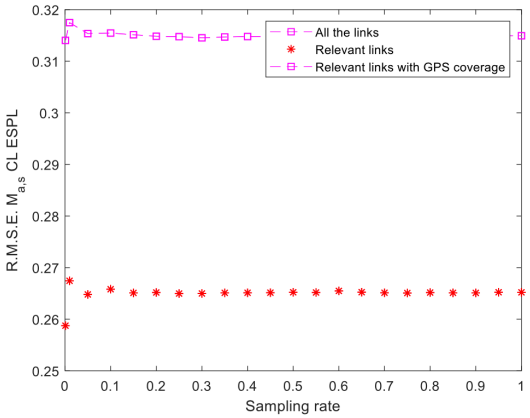


Figure 6.17 MAPE on the assignment map entries (CL with explicit path enumeration).

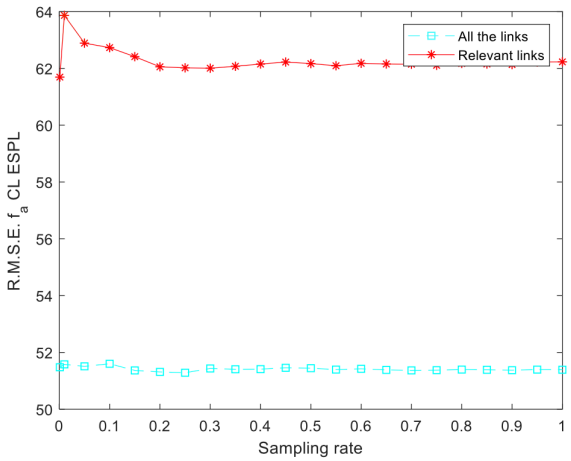


Figure 6.18 RMSE on link flows (CL with explicit path enumeration).

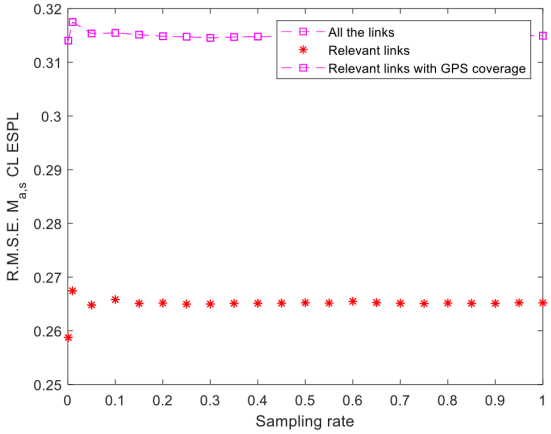


Figure 6.19 RMSE on the assignment map entries (CL with explicit path enumeration).

6.4 Conclusions

This part of the work aims at evaluating the reliability of standard assignment procedures (e.g. model-based) and at comparing their outcomes with direct estimation performances by means of synthetic experiments based on a ground-truth population of more than 50'000 individuals. The evaluation is developed in terms of total link flows and single assignment map entries, using as reference values the ground-truth values of the simulated population. Results show that direct estimates produce an error on link flows estimates rapidly decreasing with the sampling rate, but less rapidly when compared to the error trend of assignment map entries estimates. Regarding the link flows, a sampling rate of about 20% is necessary to obtain an acceptable level of error. Consequently, obtaining reliable estimates by inference using trajectory data can be a burdensome procedure and the necessary sampling rates enabling comparable performances with model-based assignment procedures (MNL, CL) can be outside the usual practical possibilities. Concerning the tested models, their performance result stable over the entire range of the sampling rates values. Indeed, the models reach their maximum performance stability even for values ranging from 1 to 5% of sampling rates, both in terms of link flows and in terms of assignment map entries. The gap between direct estimation outcomes and assignment models outputs is not negligible in terms of assignment map, although evidence showed a compensation in terms of aggregated link flows.

Therefore, o-d flows estimation/updating problem performances can be heavily affected by replacing assignment models with direct estimates of the traffic assignment map, specifically introducing not negligible errors into the estimation process.

Overall, an interesting tendency seems to appear: a direct-based estimate requires a significant number of trajectories, unlikely available even considering the current market penetration of sensing devices. In addition, for relatively low sampling rates and considering optimal conditions (i.e. a synthetic population consistent with a simple choice context), model-based estimates yield comparable results with direct-based estimates, although both relatively high in absolute terms.

7 Testing o-d flows estimation/updating methods in presence of trajectory data

This chapter aims at testing the simultaneous GLS and QD-GLS performances in presence of trajectory data by means of laboratory experiments, implemented on the basis of the results obtained from the studies described in the previous Chapters. Two key inputs of the GLS estimators formulations are directly estimated from a synthetic trajectory dataset: the a priori matrix, obtained according to the methods described in Chapter 5, and the assignment map whose estimation procedure is described in Chapter 6.

7.1 Motivation

Few studies from literature dealing with the o-d estimation problem have demonstrated that trajectory data can ameliorate o-d flows estimation/updating methods performances (Iqbal et al. 2014;Yang et al. 2017; Mitra et al. 2020). Currently, information derived from a sample of trajectory data (see Section 2.2) enriches o-d flows updating methods (e.g. GLS- estimators), indeed opportunities from exploiting trajectory data are mainly related to novel and alternative formulations of the o-d flow updating problem itself. Although a few studies on the topic are available, to the author's knowledge no systematic assessment of the potential of these approaches has been developed yet. Furthermore, amongst methods proposed in the literature for o-d flows estimation/updating in presence of trajectory data, no assessment of the potential of the quasi-dynamic framework has been proposed, as also explicitly mentioned by Yang et al. (2017).

In light of this, to investigate the potential of existing and new formulations of o-d flows updating methods in presence of trajectory data, synthetic experiments have been developed on the basis of the results and considerations deduced from precedent studies (see Section 2.2), from the analysis of direct scaling methods performances conducted in Chapter 5 and fromm the results regarding the reliability of trajectory assignment map obtained in Chapter 6. Results from the analysis in Chapter 5 revealed that, according to the type of trajectory data sample at hand, the numerosity of traffic measurements sample and their location on the network,

upscaled o-d flows can reach different levels of accuracy, which unfortunately in most of the cases are not satisfactory. Nevertheless, as also introduced by Yang et al. (2017) and Mitra et al. (2020), upscaled o-d flows can serve as target matrix for o-d flows updating procedures. Furthermore, numerous studies demonstrate the enormous advantages of substituting DTA models with a direct estimation of assignment map, although accepting to introduce considerable errors in the estimation process as demonstrated in Chapter 6. These aspects are extensively investigated by means of laboratory experiments conducted in this part of the work.

7.2 Methodology

7.2.1 Description of the experiments

The laboratory experiments test the performances of two GLS estimators enriched with information from a trajectory data sample generated from synthetic ground-truth data. The considered estimators are the simultaneous GLS estimator (Eq. 2.4) and the quasi-dynamic GLS estimator (Eq. 2.9) whose standard formulations are reported in the following:

$$\{\mathbf{d}^{*1},\dots,\mathbf{d}^{*\theta},\dots,\mathbf{d}^{*n_\theta}\} = \underset{\mathbf{x}^\theta \geq 0 \ \forall \theta \in T}{\arg\min} \left\{ \sum_{\theta=1}^{n_\theta} \sum_{od=1}^{n_{od}} \frac{\left(x_{od}^\theta - \hat{d}_{od}^\theta\right)^2}{\sigma_{od}^\theta} + \sum_{\theta=1}^{n_\theta} \sum_{l=1}^{n_l} \frac{\left(\sum_{\theta'=\theta_l}^{\theta} \sum_{od=1}^{n_{od}} m_{odl}^{\theta'\theta} x_{od}^{\theta'} - \hat{f}_l^\theta \right)^2}{\sigma_l^\theta} \right\} \quad (2.4)$$

Wherein:

- $\mathbf{x}^\theta = \{x_1^\theta,\dots,x_{n_{od}}^\theta\} \ \forall \theta \in T$ represents the unknown demand vectors;
- $\mathbf{d}^{*\boldsymbol{\theta}} = \{d_1^{*\theta},\dots,d_{n_{od}}^{*\theta}\} \forall \theta \in T$ *is* the corresponding optimal solutions
- $\hat{\boldsymbol{d}}^\theta$ the $(n_o \cdot n_d)$ matrix of the prior demand estimates $\hat{d}_{od}^\theta$ for the time slice θ;
- $\hat{\boldsymbol{f}}^\theta$ the $(n_{lc} \cdot 1)$ vector of the observed link counts $\hat{f}_l^\theta$ for the time slice θ.
- $m_{odl}^{\theta'\theta}$ is the generic term of the dynamic assignment map linking time-dependent o-d flows with time-dependent link flows ang(i.e. it represents the fraction of o-d flow generated at the time slice θ' being on link l at the time slice θ);
- σ^θ_{od} and σ^θ_l are related to the dispersion matrix of the demand and of the counted flows distribution respectively;
- θ_l is the farthest time slice whose generated demand contributes to the link flows on θ.

The implementation of quasi-dynamic estimator requires an additional setting related to the quasi-dynamic interval, that is the time-period in which distribution shares are approximated to their average values. In these experiments the quasi-dynamic time-interval was set equal to 60

minutes. This value is chosen in order to balance the number of variables to be estimated and the number of equations (e.g. deriving from linearly independent link flow measurements), (see Section 2.1.3.1).

$$\{\boldsymbol{g}^{*^1}, \dots, \boldsymbol{g}^{*^\theta}, \dots, \boldsymbol{g}^{*^{n_\theta}}; \boldsymbol{p}^{*^1}, \dots, \boldsymbol{p}^{*^\tau}, \dots, \boldsymbol{p}^{*^{n_\tau}}\} \tag{2.9}$$

$$= \arg \min_{\substack{g^1 \dots g^{n_\theta} \in S_g \\ p^1 \dots p^{n_\tau} \in S_p}} \left\{ \sum_{\theta=1}^{n_\theta} \sum_{od=1}^{n_{od}} \frac{(g_o^\theta \cdot p_{d|o}^{\tau(\theta)} - \hat{d}_{od}^\theta)^2}{\sigma_{od}^\theta} + \sum_{\theta=1}^{n_\theta} \sum_{l=1}^{n_{lc}} \frac{(\sum_{\theta'=\theta_l}^{\theta} \sum_{od=1}^{n_{od}} m_{odl}^{\theta'\theta} g_o^{\theta'} \cdot p_{d|o}^{\tau(\theta')} - \hat{f}_l^\theta)^2}{\sigma_l^\theta} \right\}$$

s.t.

$g^1 \dots g^{n_\theta} \in S_g: g_o^\theta \geq 0 \; \forall o, \forall \theta \in T$

$p^1 \dots p^{n_\tau} \in S_p: 0 \leq p_{d|o}^\tau \leq 1 \; \forall p_{d|o}^\tau \in \boldsymbol{p}_{d|o}^\tau \; \forall \tau \in T \; ; \sum_d p_{d|o}^\tau = 1 \; \forall o, \forall \tau \in T$

Wherein:

- $\boldsymbol{g}^\theta$ is the $(n_o \cdot 1)$ vector of the generated demands g_o^θ for a given time slice θ;
- $\boldsymbol{p}^\tau$ is the $(n_o \cdot n_d)$ matrix of the distribution probabilities $p_{d|o}^\tau$ for a given sub-period τ;
- $\widehat{\boldsymbol{d}}^\theta$ the $(n_o \cdot n_d)$ matrix of the prior demand estimates $\hat{d}_{od}^\theta$ for the time slice θ;
- $\hat{\boldsymbol{f}}^\theta$ the $(n_{lc} \cdot 1)$ vector of the observed link counts $\hat{f}_l^\theta$ for the time slice θ.

For the scope of the analysis, both models are informed replacing two fundamental inputs:

- the target demand flows $\hat{d}_{od}^\theta$ are replaced with $\hat{d}_{od}^{UP,\theta}$, obtained by using one of the upscaling method described in Section 5.2.3;
- the entries of the dynamic assignment map are replaced with observed values $m_{odl}^{traj,\theta'\theta}$ consisting of direct estimates calculated according to Equation 2.54.

To develop a consistent comparison between the results of the updating process, the direct scaling methods used to produce the a priori o-d flows in this set of experiments refer to Equation 2.47 and Equation 5.16. Indeed, both methods refer to a time-slice based upscaling horizon. The method in Equation 2.47 consists of rescaling the trajectory o-d matrix by means of an upscaling factor defined per each period ϑ (γ^ϑ) of the time horizon:

$$\gamma^\vartheta = \frac{\sum_{l=1}^{nlc} f_l^\vartheta}{\sum_{l=1}^{nlc} f_l^{traj,\vartheta}} \quad \forall \vartheta \in T \tag{2.47}$$

Therefore, an upscaling factor defined for some period ϑ (γ^{ϑ}) obtained as the ratio of the total number of vehicles observed from link traffic counts at time-interval ϑ to the total number of tracked vehicles traversing link l at time-interval ϑ. Considering this, he upscaled a priori o-d flows are equal to:

$$\hat{d}_{od}^{UP,\theta} = \gamma^{\vartheta} \cdot \hat{d}_{od}^{traj,\theta} \; \forall \; od \in OD, \forall \; \vartheta \in T$$

While, the second upscaling method used for this set of experiments involves a time-dependent upscaling factor by link count section which, according to the procedure described in Section 5.2.3 leads to the following o-d time-dependent upscaling factor:

$$\xi_{od}^{\vartheta} = \frac{\sum_{l=1}^{nlc \cdot n\vartheta} \zeta_l^{\vartheta} \lambda_{odl}^{traj,\theta'\theta}}{nlc \cdot n\vartheta} \; \forall \; od \in OD \tag{5.16}$$

and to the following a priori o-d flows:

$$\hat{d}_{od}^{UP,\theta} = \xi_{od}^{\vartheta} \cdot \hat{d}_{od}^{traj,\theta} \; \forall \; od \in OD, \forall \; \vartheta \in T$$

The final estimates obtained from the updating procedure are compared to ground truth data to provide goodness of fit (GOF) indicators. The GOF measures used in this part of the thesis are *MSE* and *cvRMSE*, already introduced in Section 3.3.

Figure 7.1 illustrates the global flowchart of the entire process underlying the laboratory experiments conducted in this part of the work.

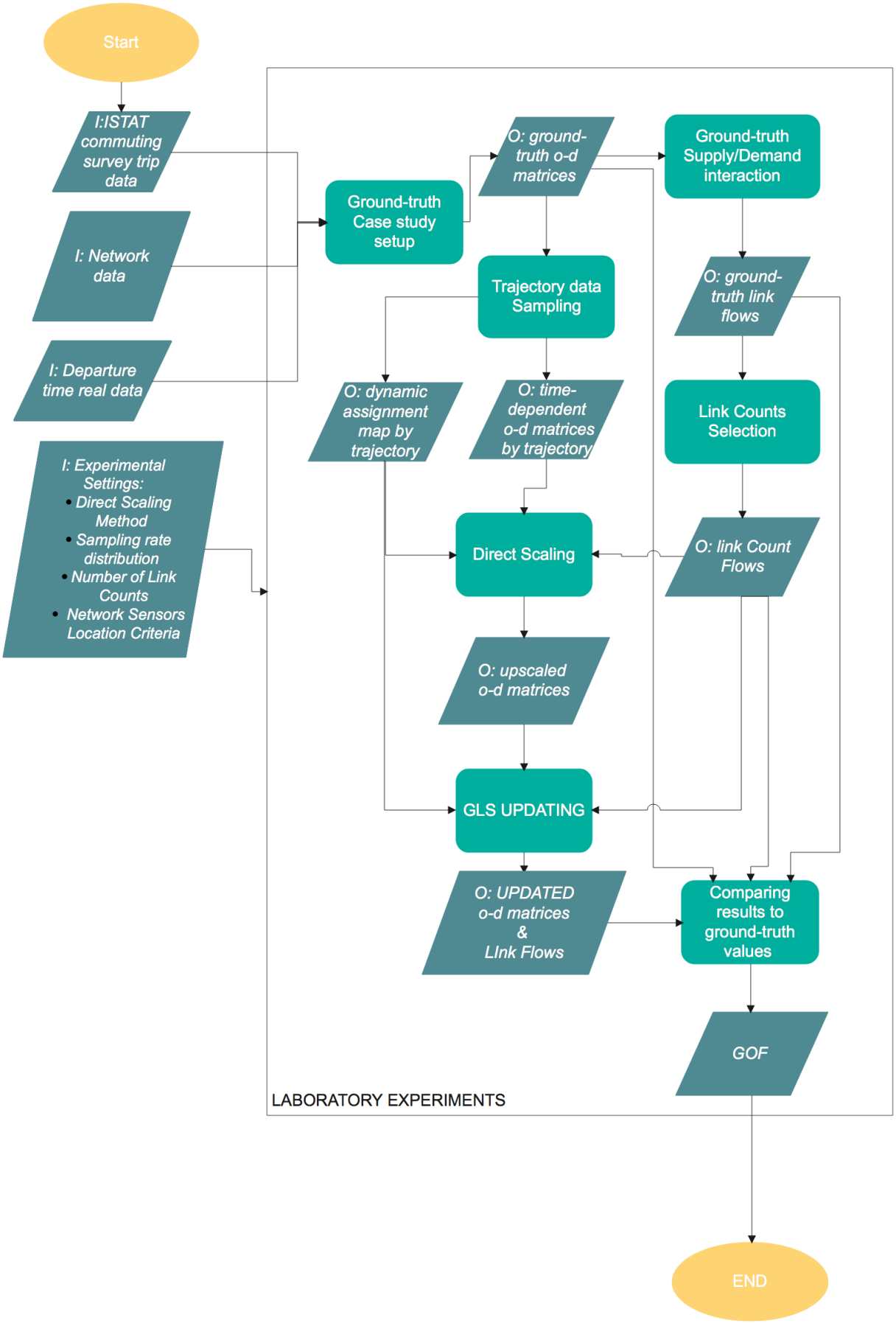


Figure 7.1 Flowchart of laboratory experiments testing GLS updating methods in presence of trajectory data.

7.2.2 Experimental Setup

The laboratory experiments are based on the same ground-truth testbeds setup for the performance analysis of direct scaling methods conducted in Chapter 5, thus the reader can refer to Section 5.2.1 for an extended description of synthetic ground-truth data setup.

The experimental plan on which this new set of experiments is based, is consistent with the experimental settings presented in Chapter 5 (see Table 5.4), although some necessary adaptations were adopted to account for the high computational burden, typical of dynamic o-d flows updating problem. The experimental settings involves the characteristics of the trajectory data sample and the link traffic measurements sample used to inform the updating model.

Concerning the trajectory data sample, experimental settings consist of different values of the sample penetration level (sampling rate) and of its distribution among the element of the network (Uniform, o-d weighted), while regarding the link counts sample, experimental settings refer to the sample numerosity and the location of the counting sections on the network(Max Flows, Random). A detailed description and the reasons underlying the definition of each setting can be found in Section 5.2.2. To account for the significant computational burden, the experimental plan explores a smaller range of sampling rate (5% to 80%) and only one value for the link counts sample numerosity, which is set to the realistic value of 1% of the total number of links.

Observations	**Experimental Settings**		
Link Counts	Max Flows	Random	Number of Link Counts 1%
Trajectory	Uniform	OD-based	Max Sampling Rate 5÷80 %

Table 7.1 Experimental Settings to test GLS estimators in presence of trajectory data

7.3 Results

Results presented in this section are referred to the large-scale network of Caserta province introduced in 5.2.1.2. For the description of the results, scenarios explored in this part of the work are reported in the following table (

Scenario	Sampling Rate Distribution	network sensors localization criterion	Upscaling Horizon/Method
U-MF-TS	Uniform	Max Flow	Time-Slice based
U-RND-TS		Random	Time-Slice based
U-MF-LC		Max Flow	Link-Count based
U-RND-LC		Random	Link-Count based
OD-MF-TS	OD-based	Max Flow	Time-Slice based
OD-RND-TS		Random	Time-Slice based
OD-MF-LC		Max Flow	Link-Count based
OD-RND-LC		Random	Link-Count based

Table 7.2) keeping the same reference of the scenarios as in Table 5.5.

Scenario	Sampling Rate Distribution	network sensors localization criterion	Upscaling Horizon/Method
U-MF-TS	Uniform	Max Flow	Time-Slice based
U-RND-TS		Random	Time-Slice based
U-MF-LC		Max Flow	Link-Count based
U-RND-LC		Random	Link-Count based
OD-MF-TS	OD-based	Max Flow	Time-Slice based
OD-RND-TS		Random	Time-Slice based
OD-MF-LC		Max Flow	Link-Count based
OD-RND-LC		Random	Link-Count based

Table 7.2 Explored scenarios to test updating methods in presence of trajectory data

7.3.1 Performances of simultaneous GLS estimator

Initial values of the *cvRMSE* reported in Table 7.3 are calculated comparing upscaled (a priori) o-d flows to ground-truth values. Analogously, final values are obtained comparing updated o-d flows to ground-truth values.

Comparing final values to initial values of the *cvRMSE* on o-d flows in Table 7.3 referred to the scenarios involving a uniform distribution of the sampling rate and a time-slice based upscaling method, a slight improvement on demand flows accuracy can be noted. A similar trend is observed for scenarios ***OD-MF-LC*** and ***OD-RND-LC***. This remark holds regardless of the link counts sample characteristics (numerosity and sensors location). Conversely, using a link-count based upscaling method (scenarios ***U-MF-LC*** and ***U-RND-LC***), it is evident that no appreciable improvement can be obtained implementing the updating process (see also percentages of reduction in Table 7.4). This trend is explained by the fact that, link count measurements cannot further inform the updating model, indeed initial values of *cvRMSE* on link counts are already very low (see Table 7.5).

Max Sampling Rate/ Scenario		*5%*	*10%*	*20%*	*30%*	*40%*	*50%*	*60%*	*70%*	*80%*
U-MF-TS	I	0.96	0.68	0.44	0.34	0.28	0.22	0.18	0.15	0.11
	F	0.92	0.65	0.41	0.31	0.25	0.20	0.17	0.13	0.10
U-RND-TS	I	0.95	0.69	0.43	0.34	0.28	0.22	0.18	0.15	0.11
	F	0.89	0.65	0.41	0.31	0.26	0.20	0.17	0.14	0.10
U-MF-LC	I	0.97	0.74	0.57	0.47	0.40	0.33	0.27	0.21	0.15
	F	0.97	0.74	0.56	0.47	0.40	0.32	0.26	0.20	0.15
U-RND-LC	I	1.11	0.94	0.75	0.66	0.56	0.45	0.36	0.28	0.19
	F	1.11	0.95	0.75	0.65	0.56	0.45	0.36	0.28	0.19
OD-MF-TS	I	1.88	1.53	1.43	1.21	1.07	0.95	0.81	0.71	0.63
	F	1.78	1.41	1.28	1.07	0.95	0.83	0.70	0.60	0.51
OD-RND-TS	I	1.62	1.36	1.22	1.11	0.97	0.87	0.79	0.68	0.61
	F	1.53	1.26	1.09	1.01	0.89	0.78	0.71	0.60	0.57
OD-MF-LC	I	1.49	1.28	1.12	1.01	0.94	0.83	0.74	0.65	0.57

	F	1.48	1.22	1.05	0.92	0.87	0.76	0.67	0.57	0.49
OD-RND-LC	I	1.45	1.27	1.21	1.03	0.96	0.86	0.73	0.67	0.58
	F	1.45	1.26	1.19	1.01	0.94	0.83	0.71	0.65	0.57

Table 7.3 Inital (I) and final (F) values of cvRMSE on o-d flows updated with the simultaneous GLS from all the scenarios for each sampling rate

Max Sampling Rate/ Scenario	*5%*	*10%*	*20%*	*30%*	*40%*	*50%*	*60%*	*70%*	*80%*
U-MF-TS	-4.7%	-5.2%	-6.5%	-7.7%	-8.1%	-7.3%	-7.2%	-8.3%	-7.6%
U-RND-TS	-6.3%	-6.3%	-6.8%	-7.7%	-7.7%	-7.5%	-7.1%	-7.2%	-7.4%
U-MF-LC	-0.2%	0.0%	-1.2%	-1.3%	-1.4%	-1.1%	-1.2%	-1.6%	-1.7%
U- RND-LC	0.3%	0.4%	-0.4%	-0.2%	-0.4%	-0.3%	-0.3%	-0.3%	-0.3%
OD-MF-TS	-5.5%	-7.7%	-10.3%	-11.4%	-10.7%	-12.0%	-13.7%	-15.2%	-18.6%
OD-RND-TS	-5.5%	-7.0%	-10.2%	-9.1%	-9.1%	-10.5%	-10.2%	-11.4%	-6.8%
OD-MF-LC	-11.6%	-14.9%	-22.9%	-16.8%	-23.2%	-27.2%	-22.5%	-24.6%	-14.9%
OD-RND-LC	0.3%	-0.9%	-1.6%	-2.2%	-2.2%	-3.7%	-2.8%	-3.1%	-2.0%

Table 7.4 percentages of reduction of cvRMSE values after updating o-d flows

Max Sampling Rate/ Scenario	*5%*	*10%*	*20%*	*30%*	*40%*	*50%*	*60%*	*70%*	*80%*
U-MF-TS	I	0.29	0.19	0.12	0.10	0.08	0.07	0.05	0.04	0.03
	F	0.08	0.05	0.03	0.02	0.02	0.01	0.01	0.01	0.01
U-RND-TS	I	0.58	0.39	0.26	0.20	0.16	0.13	0.11	0.09	0.06
	F	0.26	0.17	0.10	0.07	0.06	0.05	0.04	0.03	0.02
U-MF-LC	I	0.15	0.11	0.07	0.06	0.05	0.04	0.03	0.03	0.02
	F	0.07	0.05	0.03	0.02	0.02	0.01	0.01	0.01	0.01
U-RND-LC	I	0.30	0.22	0.14	0.11	0.09	0.08	0.06	0.05	0.04
	F	0.21	0.13	0.07	0.05	0.04	0.04	0.03	0.02	0.02
OD-MF-TS	I	0.47	0.40	0.40	0.30	0.27	0.27	0.23	0.22	0.20
	F	0.13	0.08	0.06	0.04	0.04	0.03	0.03	0.03	0.02
OD-RND-TS	I	0.93	0.75	0.67	0.60	0.56	0.54	0.44	0.39	0.30
	F	0.40	0.27	0.24	0.17	0.16	0.14	0.12	0.11	0.08
OD-MF-LC	I	0.24	0.18	0.17	0.17	0.13	0.14	0.12	0.11	0.11
	F	0.12	0.07	0.05	0.04	0.04	0.03	0.03	0.03	0.02

OD-RND-LC	I	0.44	0.33	0.31	0.24	0.26	0.25	0.19	0.18	0.14
	F	0.31	0.19	0.15	0.10	0.10	0.10	0.07	0.06	0.04

Table 7.5 Inital (I) and final (F) values of cvRMSE on link count flows updated with the simultaneous GLS from all the scenarios for each sampling rate

As illustrated in Figure 7.2, results from the two different distribution of the sampling rate can be clearly distinguished: o-d based distribution of the sampling rate leads to higher values of the *cvRMSE* with respect to the uniform case. In the former group of experiments, fixing a certain value of the sampling rate, best performances of the simultaneous GLS are obtained when a max-flow criterion and a link-count based upscaling method are used to produce the a priori o-d flows (see the blue line denoting ***OD-MF-LC*** scenario), while for the latter group, the *cvRMSE* trends from scenarios ***U-MF-TS*** and ***U-RND-TS*** basically overlap, suggesting that concerning a uniform distribution of the sampling rate, a time-slice based upscaling method yields to the best performance of the estimator.

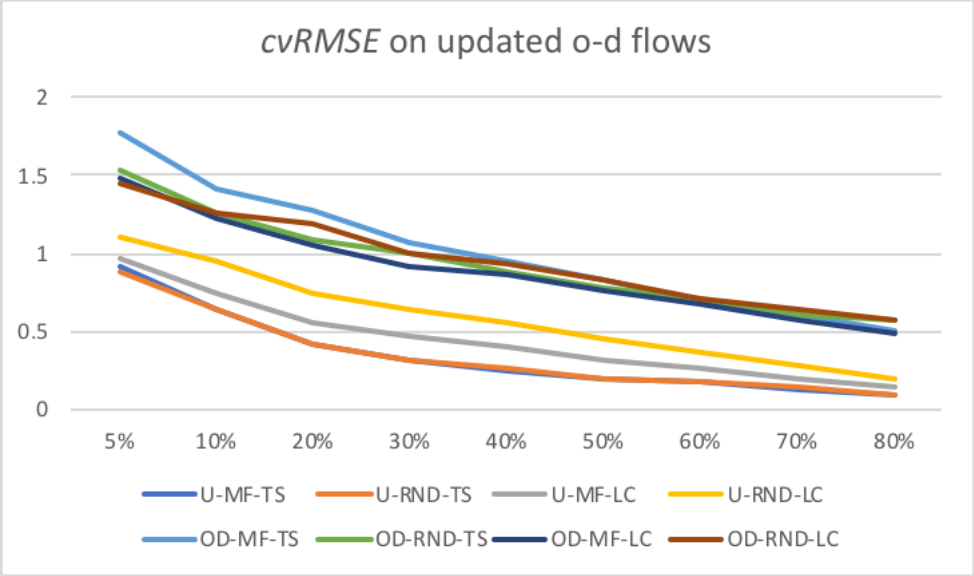


Figure 7.2 cvRMSE on o-f flows per sampling rate obtained applicating the simultaneous GLS estimator - all scenarios

7.3.2 **Performances of quasi-dynamic GLS estimator**

Considering the scenarios involving a uniform distribution of the sampling rate and for sampling rates greater than 30%, the outcomes of the QD-GLS estimator did not bring any further improvement of the level of accuracy of demand flows, instead the model seems to overfit the measurements data. This consideration holds regardless of the characteristics of trajectory data sample and link count sample. While, for the scenarios involving a different probability of extraction per each o-d pair (o-d based), a slight improvement can be observed (SEE Figure 7.3). It is worth of notice that, results of laboratory experiments testing the

reliability of assignment map derived from trajectory data have demonstrated that using such estimates can introduce a significant error into the estimation process. This could be the main explanation of the poor performances of tested methods, despite high levels of sampling rates. Nevertheless, as shown in Figure 7.4 and Figure 7.5, QD-GLS do not outperform the simultaneous estimator. This result must be evaluated accounting for the intrinsic error of ground-truth demand values, which on purpose has been set according to the empirical results obtained analysing real trajectory data to investigate the applicability of quasi-dynamic assumption in urban context (see Section 4.6). Therefore, conducting this analysis on closed networks such as highway networks may lead to very different results.

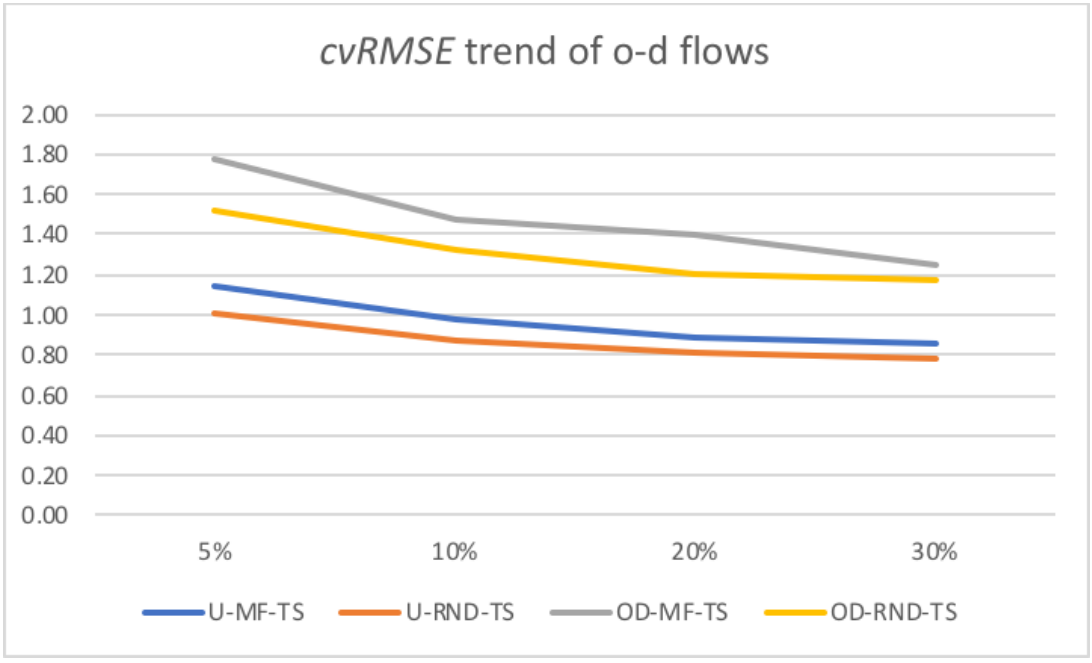


Figure 7.3 performances of the QDGLS estimator: cvRMSE trends per sampling rate

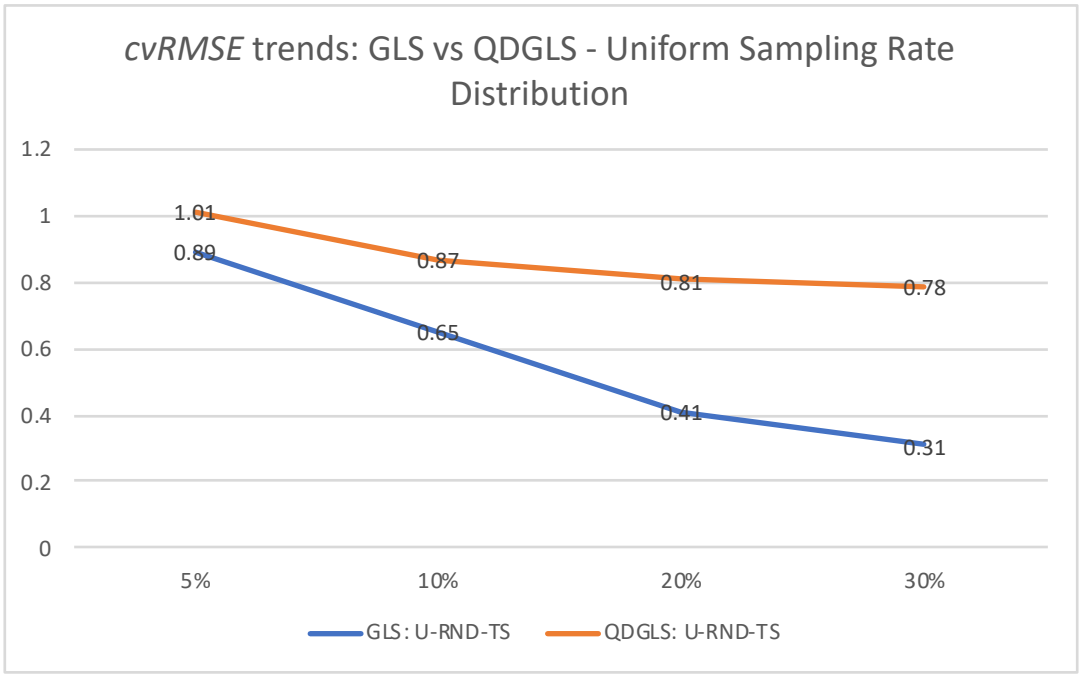


Figure 7.4 comparison between simultaneous and QD-GLS best performances for scenarios involving a uniform sampling rate distribution.

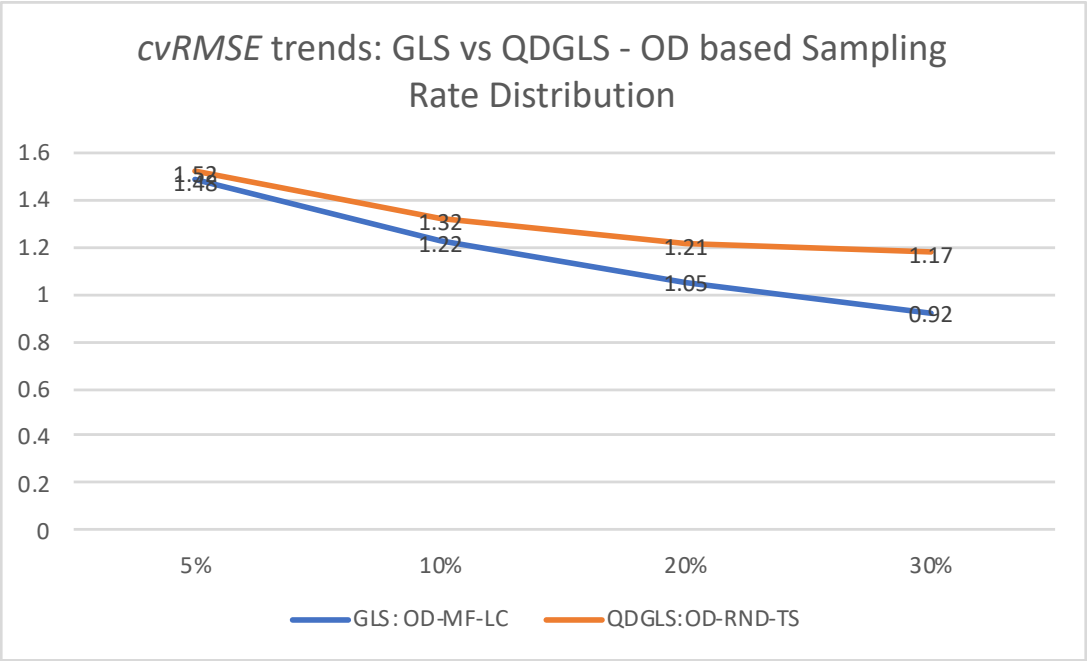


Figure 7.5 comparison between simultaneous and QD-GLS best scenarios performances for scenarios involving a o-d weighted sampling rate distribution.

7.4 Conclusions

This part of the work aims at assessing the effective improvement which various types of trajectory data can provide when informing two types of GLS estimation/updating models: the simultaneous GLS estimator and the quasi-dynamic GLS estimator. The models are enriched by introducing two fundamental adjustments in their formulations given the presence of the trajectory data sample: the a priori demand flows are obtained applying one of the direct scaling methods tested in Chapter 5 and the DTA model is substituted by introducing the direct estimates of the dynamic assignment map entries according to the procedure illustrated in Section 6.2.3.

The results of the systematic analysis confirmed some of the outcomes already present in literature. Evidence showed that regardless of the trajectory data sample, deriving a first estimate of the demand flows rescaling the observed o-d flows applying a direct scaling methods, little improvements can be achieved performing a further updating procedure by means of GLS estimators.

8 Conclusions

8.1 Research questions and main findings

The purpose of this thesis was to develop a deep understanding of the opportunities and the limitations of trajectory data to assess its potential for ameliorating the o-d flows estimation/updating problem and for conducting o-d related analysis. The proposed work involved both real trajectory data analysis and laboratory experiments based on synthetic data to investigate the implications of the trajectory data sample distinctive features (e.g. sample representativeness and bias) on demand flows accuracy. The trajectory data analysis was developed analysing a sample provided by a private company namely INRIX, one of the leading providers of mobility data. The dataset consisted of 50.933.281 GPS data points spanning over 31 days of October 2017. The corresponding 2.328.471 trajectories were collected from 101.090 mobile devices, private and fleet vehicles crossing the geographic area approximately matching the entire city of Napoli and some of the surrounding suburb areas. The analysed trajectory dataset allowed to derive a first crude estimation of time-dependent o-d matrices representing the collected trips taking place between two specific locations (traffic analysis zones) at a certain time of a day. To investigate the sample representativeness, the penetration rate and its distribution among origins, destinations and o-d pairs were evaluated. Since the true o-d flows are unknown, the estimation of the penetration level and its distribution is not straightforward, thus only an inference based on census data (e.g. population, workforce, employees by traffic analysis zone, commuter trips) can be performed. While, when census data is unavailable, the analysis could be performed using as benchmark data a reliable historical o-d matrix, usually provided from previous studies.

Evidence from the trajectory data analysis suggested a strong variability of the sampling rate defined by o-d pair for both regional and urban level, meaning that for the given sample, the penetration rate distribution is not uniform among o-d pairs, thus the probability of extracting a trip from the population (total number of actual trips) varies according to the considered o-d pair. The experimental results have been utilized as a starting point to deepen the analysis of the implications of penetration level and distribution on demand flows accuracy when applying o-d flows estimation methods and inherent scaling techniques. Upscaling the observed o-d

flows by applying direct scaling methods demonstrated to be an essential step for ameliorating o-d flows estimates derived by trajectory data. Considering this, the performances of three direct scaling methods have been analysed by means of laboratory experiments based on a synthetic data case study. The synthetic ground truth data setup consisted in developing a true testbed in terms of o-d flows, network characteristics and any other relevant parameter related to demand and supply, including an assignment model performing the interaction between demand and supply which allowed to calculate the true link flows and traffic flows characteristics and to ensure mutual consistency between o-d flows and traffic flow characteristics throughout the entire estimation process. The laboratory experiments covered a wide experimental plan accounting for the range variability and variety of trajectory data sample and traffic measurements sample characteristics leading to a total number of experiment instances equal to 11520. Indeed, the experiments were conducted considering the entire range of penetration level variability (1-100%) and distribution (uniform, o-d based) and different levels of link count sample numerosity (0.5,1,2%). Final considerations were deduced comparing upscaled/updated o-d flows and related link traffic flows to the ground-truth values, therefore results might provide useful guidelines for researchers and practitioners dealing with various types of trajectory data sample and conducting o-d related applications. Specifically, conclusions are referred to the two types of trajectory data classified according to the adopted trajectory data generation process: opportunistic data and purpose-oriented data (see Section 1.1.2 for more details).

Final considerations were deduced comparing upscaled/updated o-d flows and related link traffic flows to the ground-truth values, therefore results might provide useful guidelines for researchers and practitioners dealing with various types of trajectory data sample and conducting o-d related applications.

If a purpose oriented sampling can be adopted, thus a sample of trajectory data can be generated according to a well-structured sampling process involving a uniform distribution of the sampling rate, it is necessary to identify the level of representativeness (sampling rate) such that the desired accuracy level on demand flows can be achieved. Results deriving from scenarios assuming a uniform sampling rate distribution showed that the minimum sampling rate threshold above which the coefficient of variation of the *RMSE* calculated for o-d flows is less than 0.30 is equal to 35%. In such case, the issue of representativeness associated with trajectory data can be potentially corrected by rescaling observed o-d flows by means of direct scaling methods. Specifically, best performances are reached using a time-slice based direct scaling method as in Equation 2.47.

Conversely, if an opportunistically collected trajectory data sample is at hand, a preliminary processing of the sample should be implemented to identify the appropriate level of spatial aggregation to derive observed o-d flows and other information such as assignment map entries and path choice probabilities, such that a proper investigation on the characteristics of the sample can be developed. Subsequently, it is necessary to derive a first estimate of the level of representativeness comparing observed o-d flows with census/historical data and evaluate its distribution among the element of network (e.g. o-d pairs). In this case, the analysis of direct scaling methods demonstrated that population bias associated with trajectory data can be partially addressed by rescaling observed o-d flows with traffic counts assuming different scaling factors per o-d pairs. In absolute term, demand accuracy level remains still unsatisfactory for the entire range of sampling rate variability. This result is particularly relevant considering the range of variability of trajectory data samples available so far: concerning the range of available penetration level of such samples (e.g. 1-10%, FHWA), the coefficient of variation of RMSE calculated comparing upscaled demand flows to ground truth values is still higher than 1.5 for all the scenarios assuming an o-d based distribution. Overall, best performances in these scenarios were reached adopting a link-count based direct scaling methods, introduced in this thesis in Equation 5.16. Nevertheless, the upscaled o-d flows may be integrated as prior estimate in existing o-d updating model formulations. Indeed, the last part of this research consisted in another set of 36 laboratory experiments analysing the effective improvement which various types of trajectory data can provide when informing two types of GLS estimation/updating models: the simultaneous GLS estimator and the quasi-dynamic GLS estimator. Both models are enriched by introducing two fundamental adjustments in their formulations given the presence of the trajectory data sample: the a priori demand flows are obtained applying a direct scaling method and the DTA model is substituted by introducing the direct estimates of the dynamic assignment map entries. The results of the systematic analysis confirmed some of the outcomes already present in literature. Evidence showed that regardless of the trajectory data sample, deriving a first estimate of the demand flows by rescaling the observed o-d flows applying a direct scaling methods, little improvements can be achieved performing a further updating procedure by means of GLS estimators. Better improvements are obtained for the scenarios assuming a o-d based penetration rate distribution, in which best performance were achieved adopting a link-count based direct scaling methods to obtain the a-priori o-d matrix (see Equations 5.14 to 5.16).

Given the central role of dynamic traffic assignment in o-d estimation/updating procedures, a further analysis was required to evaluate the reliability of standard assignment procedures (e.g.

model-based) and to compare their outcomes with direct estimation performances by means of synthetic experiments based on a ground-truth population. The evaluation was developed in terms of total link flows and single assignment map entries, using as reference values the ground-truth values of the simulated population. Results show that direct estimates produce an error on link flows estimates rapidly decreasing with the sampling rate, but less rapidly when compared to the error trend of assignment map entries estimates. In addition, for relatively low sampling rates and considering optimal conditions (i.e. a synthetic population consistent with a simple choice context), model-based estimates yield comparable results with direct-based estimates, although both relatively high in absolute terms. Therefore, replacing assignment models with direct estimation of assignment map entries can avoid the introduction of behavioural assumptions, although, using such estimates can introduce a significant error into the estimation process. This could be the main explanation of the poor performances of tested methods, despite high levels of sampling rates.

To support the implementation of the quasi-dynamic estimator in presence of trajectory data and the synthetic case study setup, two preliminary studies have been conducted: the former provides insights on how to implement the quasi-dynamic o-d estimation framework when dealing with congested networks and the latter is dedicated to the assessment of quasi-dynamic evolution of the demand in urban context. Results from the former study have demonstrated that quasi-dynamic framework can still provide robust solutions of the problem when considering highly congested network especially using derivative-free algorithms to solve the optimization problem. From the latter study assessing its applicability in the urban context, experimental evidence suggests that the quasi-dynamic o-d matrix cannot be used as an adequate approximation of the trajectory o-d matrix. However, this result allowed to define the value of the intrinsic error associated to the ground-truth o-d flows of the large-scale network test site and thus to develop a realistic case study.

Future research on this topic could further validate these results generating more than one synthetic population to investigate on the implications of different demand structures and levels of the intrinsic error. Furthermore, experiment should be repeated including congestion phenomena into the analysis to evaluate its impact on the results, indeed all the experiments carried out in this work consider congestion phenomena as negligible, thus the relationship between demand flows and link flows is represented by a fixed assignment map which, in highly congested networks, cannot reproduce the actual evolution of traffic flow. In addition, regarding the direct scaling methods, improving the classification of scaling factor defined per each o-d pair may yield better results and a partition of the set of link counts can be defined, such that

two independent set of measurements data can be used separately in two phases, respectively for direct scaling and for further updating procedures.

References

1. Aerde, M.V., Hellinga, B., Yu, L., Rakha, H., n.d. VEHICLE PROBES AS REAL-TIME ATMS SOURCES OF DYNAMIC O-D AND TRAVEL TIME DATA 21.
2. Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies 58, 240–250. https://doi.org/10.1016/j.trc.2015.02.018
3. Antoniou, C., Barceló, J., Breen, M., Bullejos, M., Casas, J., Cipriani, E., Ciuffo, B., Djukic, T., Hoogendoorn, S., Marzano, V., Montero, L., Nigro, M., Perarnau, J., Punzo, V., Toledo, T., van Lint, H., 2016. Towards a generic benchmarking platform for origin–destination flows estimation/updating algorithms: Design, demonstration and validation. Transportation Research Part C: Emerging Technologies 66, 79–98. https://doi.org/10.1016/j.trc.2015.08.009
4. Antoniou, C., Ben-Akiva, M., Koutsopoulos, H.N., 2006. Dynamic traffic demand prediction using conventional and emerging data sources. IEE Proceedings - Intelligent Transport Systems 153, 97. https://doi.org/10.1049/ip-its:20055006
5. Antoniou, C., Lima Azevedo, C., Lu, L., Pereira, F., Ben-Akiva, M., 2015. W-SPSA in practice: Approximation of weight matrices and calibration of traffic simulation models. Transportation Research Part C: Emerging Technologies 59, 129–146. https://doi.org/10.1016/j.trc.2015.04.030
6. Ashok K. and Ben-Akiva M. (2002). Estimation and Prediction of Time-Dependent Origin- destination Flows with a Stochastic Mapping to Path Flows and Link Flows. *Transportation Science 36*, 184-198.
7. Ashok K. and Ben-Akiva M. (1993). Dynamic Origin-Destination Matrix Estimation and Prediction for Real-Time Traffic Management Systems. C.F. Daganzo editor (Elsevier), *International Symposium on Transportation and Traffic Theory*, 465-484.
8. Ashok, K., Ben-Akiva, M.E., 2000. Alternative Approaches for Real-Time Estimation and Prediction of Time-Dependent Origin–Destination Flows. Transportation Science 34, 21–36. https://doi.org/10.1287/trsc.34.1.21.12282
9. Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., Puchinger, J., 2019. Inferring dynamic origin-destination flows by transport mode using mobile phone data.

Transportation Research Part C: Emerging Technologies 101, 254–275. https://doi.org/10.1016/j.trc.2019.02.013

10. Bahoken, F., Raimond, A.-M.O., n.d. Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths 15.
11. Balakrishna, R., Antoniou, C., Ben-Akiva, M., Koutsopoulos, H.N. and Wen, Y., 2007. Calibration of microscopic traffic simulation models: Methods and application. Transportation Research Record, 1999(1), pp.198-207.
12. Balakrishna, R., Koutsopoulos, H. N. and M. Ben-Akiva (2005). Calibration and validation of dynamic traffic assignment systems. *Proceedings of 16th ISTTT*, 407-426.
13. Balakrishna, Ramachandran & Koutsopoulos, Haris. (2008). Incorporating Within-Day Transitions in Simultaneous Offline Estimation of Dynamic Origin-Destination Flows Without Assignment Matrices. Transportation Research Record. 2085. 31-38. 10.3141/2085-04.
14. Barceló, J., L. Montero, M. Bullejos, M. Linares, and O. Serch (2013). Robustness and computational efficiency of a Kalman Filter estimator of time dependent OD matrices exploiting ICT traffic measurements. Transportation Research Record: Journal of the Transportation Research Board, 2344 (4), pp. 31–39.
15. Barceló, J., Montero, L., Marqués, L., Marinelli, P., Carmona, C., n.d. TRAVEL TIME FORECASTING AND DYNAMIC OD ESTIMATION IN FREEWAYS BASED ON BLUETOOTH TRAFFIC MONITORING 20.
16. Bell, M.G.H. (1983). The estimation of origin-destination matrix from traffic counts. Transportation Science 10, 198-217.
17. Bell, M.G.H. (1991). The estimation of origin–destination matrices by constrained generalized least squares. Transportation Research 25B, 13–22.
18. Bianco, L., G. Confessore, and P. Reverberi. A Network Based Model for Traffic Sensor Location with Implications on O-D Matrix Estimates. Transportation Science, Vol. 35, No. 1, 2001, pp. 50–60.
19. Bierlaire M. and F. Crittin (2004). An efficient algorithm for real-time estimation and prediction of dynamic OD table. *Operations Research* 52(1).
20. Bonnel, P., Fekih, M., Smoreda, Z., 2018. Origin-Destination estimation using mobile network probe data. Transportation Research Procedia 32, 69–81. https://doi.org/10.1016/j.trpro.2018.10.013
21. Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M., Smoreda, Z., 2015. Passive Mobile Phone Dataset to Construct Origin-destination Matrix: Potentials and

Limitations. Transportation Research Procedia 11, 381–398. https://doi.org/10.1016/j.trpro.2015.12.032

22. Bracken, J. and McGill, J.T., 1973. Mathematical programs with optimization problems in the constraints. Operations Research, 21(1), pp.37-44.
23. Bricka, S.G., Sen, S., Paleti, R., Bhat, C.R., 2012. An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. Transportation Research Part C: Emerging Technologies 21, 67–88. https://doi.org/10.1016/j.trc.2011.09.005
24. Caitlin C., Pereira F., Zhao F., Dias I., Lim H., Ben-Akiva M., and Zegras P. Future Mobility Survey. Transportation Research Record: Journal of the Transportation Research Board 2354 (December 2013): 59–67.
25. Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C., 2011. Estimating Origin-Destination Flows Using Mobile Phone Location Data. IEEE Pervasive Computing 10, 36–44. https://doi.org/10.1109/MPRV.2011.41
26. Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies 26, 301–313. https://doi.org/10.1016/j.trc.2012.09.009
27. Cantelmo G, Viti F, Tampère CMJ, Cipriani E, Nigro M. Two-Step Approach for Correction of Seed Matrix in Dynamic Demand Estimation. Transportation Research Record. 2014; 2466(1):125-133. doi:10.3141/2466-14
28. Cantelmo, G., Cipriani, E., Gemma, A., Nigro, M., 2014. An Adaptive Bi-Level Gradient Procedure for the Estimation of Dynamic Traffic Demand. IEEE Transactions on Intelligent Transportation Systems 15, 1348–1361. https://doi.org/10.1109/TITS.2014.2299734
29. Cantelmo G., F. Viti, E. Cipriani and N. Marialisa, "A Two-Steps Dynamic Demand Estimation Approach Sequentially Adjusting Generations and Distributions," 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015, pp. 1477-1482, doi: 10.1109/ITSC.2015.241.
30. Cantelmo, G., Viti, F., Cipriani, E., Nigro, M., 2017. A Utility-based Dynamic Demand Estimation Model that Explicitly Accounts for Activity Scheduling and Duration. Transportation Research Procedia 23, 440–459. https://doi.org/10.1016/j.trpro.2017.05.025
31. Cantelmo, G. (2018) Dynamic Origin-Destination Matrix Estimation with Interacting Demand Patterns (PhD Dissertation).

32. Cantelmo, G. and Viti, F. (2020) A Big Data Demand Estimation Model for Urban Congested Networks. Transport and Telecommunication Journal, Vol.21 (Issue 4), pp. 245-254. https://doi.org/10.2478/ttj-2020-0019
33. Carrese, S., Cipriani, E., Mannini, L., Nigro, M., 2017. Dynamic demand estimation and prediction for traffic urban networks adopting new data sources. Transportation Research Part C: Emerging Technologies 81, 83–98. https://doi.org/10.1016/j.trc.2017.05.013
34. Cascetta E. (2001). *Transportation systems engineering: theory and methods,* Kluwer Academic Publishers.
35. Cascetta E. and Postorino M.N. (2001). Fixed point models for the estimation of O-D matrices using traffic counts on congested networks. Transportation Science 35(2).
36. Cascetta E., Inaudi D. and G. Marquis (1993). Dynamic Estimators of Origin-Destination Matrices using Traffic Counts. *Transportation Science* 27, 363-373
37. Cascetta E., Papola A. and Cartenì A. (2005). Prediction reliability of the transport simulation models: a before and after study in Naples. Proceedings *of 2005 ETC Conference*, Strasbourg.
38. Cascetta, E. (1984). Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator. Transportation Research 18B, 289–299.
39. Cascetta, E. and F. Russo (1997). Calibrating aggregate travel demand models with traffic counts: estimators and statistical performance. *Transportation* 24, 271-293.
40. Cascetta, E. and Nguyen, S. (1988). A unified framework for estimating or updating origin/destination matrices from traffic counts. Transportation Research 22B, 437–455.
41. Cascetta, E., Papola, A., Marzano, V., Simonelli, F., Vitiello, I., 2013. Quasi-dynamic estimation of o–d flows from traffic counts: Formulation, statistical validation and performance analysis on real data. Transportation Research Part B: Methodological 55, 171–187. https://doi.org/10.1016/j.trb.2013.06.007
42. Chen, A., S. Pravinvongvuth, P. Chootinan, M. Lee, and W. Recker. Strategies for Selecting Additional Traf c Counts for Improving O-D Trip Table Estimation. Transportmetrica, Vol. 3, No. 3, 2007, pp. 191–211.
43. Chen, C., Bian, L., Ma, J., 2014. From traces to trajectories: How well can we guess activity locations from mobile phone traces? Transportation Research Part C: Emerging Technologies 46, 326–337. https://doi.org/10.1016/j.trc.2014.07.001

44. Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation Research Part C: Emerging Technologies 68, 285–299. https://doi.org/10.1016/j.trc.2016.04.005
45. Chen, C., Xuegang (Jeff) Ban, Feilong Wang, Jingxing Wang, Choudhury Siddique, Fan, R., Jaehun Lee, 2017. Understanding GPS and Mobile Phone Data for Origin-Destination Analysis. https://doi.org/10.13140/RG.2.2.13294.66889
46. Cipriani, E., M. Florian, M. Mahut, and M. Nigro (2011). A gradient approximation approach for adjusting temporal origin-destination matrices. Transportation Research Part C: Emerging Technologies, Vol. 19, No. 2, 2011, pp. 270–282.
47. Cipriani, E., Nigro, M., Fusco, G., Colombaroni, C., 2014. Effectiveness of link and path information on simultaneous adjustment of dynamic O-D demand matrix. European Transport Research Review 6, 139–148. https://doi.org/10.1007/s12544-013-0115-z
48. Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiratta, S.R., González, M.C., 2015. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations, and Opportunities. Transportation Research Record 2526, 126–135. https://doi.org/10.3141/2526-14
49. Cremer, M. and H. Keller (1981). Dynamic identification of O-D flows from traffic counts at complex intersections. *Proceedings of 8th ISTTT*.
50. Cremer, M. and H. Keller (1984). A systems dynamics approach to the estimation of entry and exit O-D flows. *Proceedings of 9th ISTTT*.
51. Cremer, M. and H. Keller (1987). **A** new class of dynamic methods for the identification of origin-destination flows. *Transportation Research* 21B, 117-132.
52. Cui, Y., Meng, C., He, Q., Gao, J., 2018. Forecasting current and next trip purpose with social media data and Google Places. Transportation Research Part C: Emerging Technologies 97, 159–174. https://doi.org/10.1016/j.trc.2018.10.017
53. D. Bauer, G. Richter, J. Asamer, B. Heilmann, G. Lenz and R. Kölbl, "Quasi-Dynamic Estimation of OD Flows From Traffic Counts Without Prior OD Matrix," in IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 6, pp. 2025-2034, June 2018. doi: 10.1109/TITS.2017.2741528
54. Daganzo, C.F. and Sheffi, Y. (1977). On stochastic models of traffic assignment. Transportation Science, 11(3).
55. Dash, M., Nguyen, H.L., Hong, C., Yap, G.E., Nguyen, M.N., Li, X., Krishnaswamy, S.P., Decraene, J., Antonatos, S., Wang, Y., Anh, D.T., Shi-Nash, A., 2014. Home and

Work Place Prediction for Urban Planning Using Mobile Network Data, in: 2014 IEEE 15th International Conference on Mobile Data Management. Presented at the 2014 15th IEEE International Conference on Mobile Data Management (MDM), IEEE, Brisbane, Australia, pp. 37–42. https://doi.org/10.1109/MDM.2014.65

56. Demissie, M.G., de Almeida Correia, G.H., Bento, C., 2013. Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. Transportation Research Part C: Emerging Technologies 32, 76–88. https://doi.org/10.1016/j.trc.2013.03.010
57. Dixon, M.P., Rilett, L.R., 2005. Population Origin–Destination Estimation Using Automatic Vehicle Identification and Volume Data. J. Transp. Eng. 131, 75–82. https://doi.org/10.1061/(ASCE)0733-947X(2005)131:2(75)
58. Djukic, T., Van Lint, J.W.C., Hoogendoorn, S.P., 2012. Application of Principal Component Analysis to Predict Dynamic Origin–Destination Matrices. Transportation Research Record 2283, 81–89. https://doi.org/10.3141/2283-09
59. DUONG, T. and HAZELTON, M.L. (2005), Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation. Scandinavian Journal of Statistics, 32: 485-506. https://doi.org/10.1111/j.1467-9469.2005.00445.x
60. Ehlert, A., M.G.H. Bell, and S. Grosso. The Optimization of Traffic Count Locations in Road Networks. Transportation Research Part B, Vol. 40, No. 6, 2006, pp. 460–479.
61. Erhardt, G.D., Rizzo, L., 2018. Evaluating the biases and sample size implications of multi-day GPS-enabled household travel surveys. Transportation Research Procedia 32, 279–290. https://doi.org/10.1016/j.trpro.2018.10.051
62. FHWA-HEP—16-083 Synopsis of New Methods and Technologies to Collect Origin-Destination (O-D) Data
63. Florian, M. and Chen, Y. (1995). A coordinate descent method for the bi-level O–D matrix adjustment problem. International Transportation Operations Research 2, 165–179.
64. Frederix, R., n.d. The effect of dynamic network loading models on DTA-based OD estimation 23.
65. Gan, L., H. Yang, and S. C. Wong. Traffic Counting Location and Error Bound in Origin–Destination Matrix Estimation Problems. Journal of Transportation Engineering, Vol. 131, No. 7, 2005, pp. 524–534.

66. Ge, Q., Fukuda, D., 2016. Updating origin–destination matrices with aggregated data of GPS traces. Transportation Research Part C: Emerging Technologies 69, 291–312. https://doi.org/10.1016/j.trc.2016.06.002
67. Gps.gov (2019)
68. Gundlegard, D., Karlsson, J.M., 2009. Route classification in travel time estimation based on cellular network signaling, in: 2009 12th International IEEE Conference on Intelligent Transportation Systems. Presented at the 2009 12th International IEEE Conference on Intelligent Transportation Systems (ITSC), IEEE, St. Louis, pp. 1–6. https://doi.org/10.1109/ITSC.2009.5309692
69. Gundlegård, D., Rydergren, C., Breyer, N., Rajna, B., 2016. Travel demand estimation and network assignment based on cellular network data. Computer Communications 95, 29–42. https://doi.org/10.1016/j.comcom.2016.04.015
70. Hadachi A., Mozhgan Pourmoradnasseri, Kaveh Khoshkhah, Unveiling large-scale commuting patterns based on mobile phone cellular network data, Journal of Transport Geography, Volume 89, 2020, 102871,ISSN 0966-6923, https://doi.org/10.1016/j.jtrangeo.2020.102871.
71. Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transportation Research Part C: Emerging Technologies 44, 363–381. https://doi.org/10.1016/j.trc.2014.04.003
72. Hazelton, M. (2000). Estimation of origin–destination matrices from link flows on uncongested networks. Transportation Research 34B, 549–566.
73. Hazelton, M. (2003). Some comments on origin–destination matrix estimation. Transportation Research 37A, 811–822.
74. Huang, H., Cheng, Y., Weibel, R., 2019. Transport mode detection based on mobile phone network data: A systematic review. Transportation Research Part C: Emerging Technologies 101, 297–312. https://doi.org/10.1016/j.trc.2019.02.008
75. Iqbal, Md.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies 40, 63–74. https://doi.org/10.1016/j.trc.2014.01.002
76. J. Tang, Y. Wang, W. Hao, F. Liu, H. Huang and Y. Wang, "A Mixed Path Size Logit-Based Taxi Customer-Search Model Considering Spatio-Temporal Factors in Route Choice," in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 4, pp. 1347-1358, April 2020, doi: 10.1109/TITS.2019.2905579.

77. Janzen, M., Vanhoof, M., Smoreda, Z., Axhausen, K.W., 2018. Closer to the total? Long-distance travel of French mobile phone users. Travel Behaviour and Society 11, 31–42. https://doi.org/10.1016/j.tbs.2017.12.001
78. Jing, W., Dianhai, W., Xianmin, S., Di, S., 2011. Dynamic OD Expansion Method Based on Mobile Phone Location, in: 2011 Fourth International Conference on Intelligent Computation Technology and Automation. Presented at the 2011 International Conference on Intelligent Computation Technology and Automation (ICICTA), IEEE, Shenzhen, China, pp. 788–791. https://doi.org/10.1109/ICICTA.2011.204
79. Karmarkar, N., 1984, December. A new polynomial-time algorithm for linear programming. In Proceedings of the sixteenth annual ACM symposium on Theory of computing (pp. 302-311).
80. Kim, H., Jayakrishnan, R., 2010. The estimation of a time-dependent OD trip table with vehicle trajectory samples. Transportation Planning and Technology 33, 747–768. https://doi.org/10.1080/03081060.2010.536629
81. Krishnakumari, P., van Lint, H., Djukic, T., Cats, O., 2019. A data driven method for OD matrix estimation. Transportation Research Part C: Emerging Technologies. https://doi.org/10.1016/j.trc.2019.05.014
82. Larijani, A.N., Olteanu-Raimond, A.-M., Perret, J., Brédif, M., Ziemlicki, C., 2015. Investigating the Mobile Phone Data to Estimate the Origin Destination Flow and Analysis; Case Study: Paris Region. Transportation Research Procedia 6, 64–78. https://doi.org/10.1016/j.trpro.2015.03.006
83. Lee, R.J., Sener, I.N., Mullins, J.A., 2016. An evaluation of emerging data collection technologies for travel demand modeling: from research to practice. Transportation Letters 8, 181–193. https://doi.org/10.1080/19427867.2015.1106787
84. Leontiadis, I., Lima, A., Kwak, H., Stanojevic, R., Wetherall, D., Papagiannaki, K., 2014. From Cells to Streets: Estimating Mobile Paths with Cellular-Side Data, in: Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies - CoNEXT '14. Presented at the the 10th ACM International, ACM Press, Sydney, Australia, pp. 121–132. https://doi.org/10.1145/2674005.2674982
85. Lo, H., Zhang, N. and H. Lam (1996). Estimation of an origin-destination matrix with random link choice proportions: a statistical approach. Transportation Research 30B, 309–324.

86. Lo, H.P. and C.P. Chan. (2003). Simultaneous estimation of an origin-destination matrix and link choice proportions using traffic counts. Transportation Research 37A, 771–788.
87. Lu, C.-C., Zhou, X., Zhang, K., 2013. Dynamic origin–destination demand flow estimation under congested traffic conditions. Transportation Research Part C: Emerging Technologies 34, 16–37. https://doi.org/10.1016/j.trc.2013.05.006
88. Lu, L., Y. Xu, C. Antoniou and M. Ben-Akiva (2015). An Enhanced SPSA Algorithm for the Calibration of Dynamic Traffic Assignment Models. Transportation Research Part C, 51, pp. 149-166.
89. Ma, J., Li, H., Yuan, F., Bauer, T., 2013. Deriving Operational Origin-Destination Matrices From Large Scale Mobile Phone Data. International Journal of Transportation Science and Technology 2, 183–204. https://doi.org/10.1260/2046-0430.2.3.183
90. Maher, M. (1983). Inferences on trip matrices from observations on link volumes: a Bayesian statistical approach. Transportation Research 20B, 435–447.
91. Markovic, N., Sekula, P., Vander Laan, Z., Andrienko, G., Andrienko, N., 2019. Applications of Trajectory Data From the Perspective of a Road Transportation Agency: Literature Review and Maryland Case Study. IEEE Trans. Intell. Transport. Syst. 20, 1858–1869. https://doi.org/10.1109/TITS.2018.2843298
92. Marzano, V., Papola, A., Simonelli, F., 2009. Limits and perspectives of effective O–D matrix correction using traffic counts. Transportation Research Part C: Emerging Technologies 17, 120–132. https://doi.org/10.1016/j.trc.2008.09.001
93. Marzano, V., Papola, A., Simonelli, F., Papageorgiou, M., 2018. A Kalman Filter for Quasi-Dynamic o-d Flow Estimation/Updating. IEEE Transactions on Intelligent Transportation Systems 19, 3604–3612. https://doi.org/10.1109/TITS.2018.2865610
94. Michau, G., Pustelnik, N., Borgnat, P., Abry, P., Nantes, A., Bhaskar, A., Chung, E., 2017. A Primal-Dual Algorithm for Link Dependent Origin Destination Matrix Estimation. IEEE Trans. on Signal and Inf. Process. over Networks 3, 104–113. https://doi.org/10.1109/TSIPN.2016.2623094
95. Mitra, A., Attanasi, A., Meschini, L., Gentile, G., 2020. Methodology for O-D matrix estimation using the revealed paths of floating car data on large-scale networks. IET intell. transp. syst 14, 1704–1711. https://doi.org/10.1049/iet-its.2019.0684
96. Montero, L., Ros-Roca, X., Herranz, R., Barceló, J., 2019. Fusing mobile phone data with other data sources to generate input OD matrices for transport models.

Transportation Research Procedia 37, 417–424. https://doi.org/10.1016/j.trpro.2018.12.211

97. Nigro, M., Abdelfatah, A., Cipriani, E., Colombaroni, C., Fusco, G., Gemma, A., 2018. Dynamic O-D Demand Estimation: Application of SPSA AD-PI Method in Conjunction with Different Assignment Strategies. Journal of Advanced Transportation 2018, 2085625. https://doi.org/10.1155/2018/2085625
98. Nigro, M., Cipriani, E., del Giudice, A., 2018. Exploiting floating car data for time-dependent Origin–Destination matrices estimation. Journal of Intelligent Transportation Systems 22, 159–174. https://doi.org/10.1080/15472450.2017.1421462
99. Nihan, N. L. and G. A. Davis (1987). Recursive estimation of origin-destination matrices from input-output counts. *Transportation Research* 21B, 149-163
100. Nihan, N. L. and G. A. Davis (1989). Application of prediction-error minimization and maximum likelihood to estimate intersection O-D matrices from traffic counts. *Transportation Science* 23B, 1989, 77-90.
101. Okutani I. and Y. Stephanedes (1984). Dynamic Prediction of Traffic Volume through Kalman Filtering Theory. Transportation Research B, 18(2).
102. Ortuzar, J. And L. Willumsen (2011). Modelling transport. 4th edition. Wiley ed.
103. Park, J., Murphey, Y.L., McGee, R., Kristinsson, J.G., Kuang, M.L., Phillips, A.M., 2014. Intelligent Trip Modeling for the Prediction of an Origin–Destination Traveling Speed Profile. IEEE Transactions on Intelligent Transportation Systems 15, 1039–1053. https://doi.org/10.1109/TITS.2013.2294934
104. Parry, K., Hazelton, M.L., 2012. Estimation of origin–destination matrices from link counts and sporadic routing data. Transportation Research Part B: Methodological 46, 175–188. https://doi.org/10.1016/j.trb.2011.09.009
105. Rao, W., Wu, Y.-J., Xia, J., Ou, J., Kluger, R., 2018. Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data. Transportation Research Part C: Emerging Technologies 95, 29–46. https://doi.org/10.1016/j.trc.2018.07.002
106. Rui Yao, Shlomo Bekhor Data-driven choice set generation and estimation of route choice models, Transportation Research Part C: Emerging Technologies, Volume 121, 2020, 102832, ISSN 0968-090X, https://doi.org/10.1016/j.trc.2020.102832.

107. Sanaullah, I., Quddus, M., Enoch, M., 2016. Developing travel time estimation methods using sparse GPS data. Journal of Intelligent Transportation Systems 20, 532–544. https://doi.org/10.1080/15472450.2016.1154764

108. Shihsien L., Fricker J. D., Estimation of a trip table and the Θ parameter in a stochastic network, Transportation Research Part A: Policy and Practice, Volume 30, Issue 4, 1996, Pages 287-305, ISSN 0965-8564, https://doi.org/10.1016/0965-8564(95)00031-3.

109. Simonelli, F., Marzano V., Papola A. and Vitiello I. 2012. A network sensor location procedure accounting for o–d matrix estimate variability, Transportation Research Part B: Methodological, 46, issue 10, p. 1624-1638.

110. Simonelli F., Tinessa F., Marzano V., Papola A. and Romano A., "Laboratory experiments to assess the reliability of traffic assignment map," 2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2019, pp. 1-9, doi: 10.1109/MTITS.2019.8883390.

111. Spall, J.C., 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE transactions on automatic control, 37(3), pp.332-341.

112. Spall, J. C. (1998a). Implementation of the simultaneous perturbation algorithm for stochastic approximation, *IEEE Transactions on Aerospace and Electronic Systems* **34**: 817–823.

113. Spall, J.C. (1992). Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation, *IEEE Transactions on Automatic Control*, vol. **37**, pp. 332-341.

114. Spall, J.C. (1998b). An Overview of the Simultaneous Perturbation Method for Efficient Optimization. Johns Hopkins Apl Technical Digest, **19**(4): 482-492.

115. Tang, J., Liang, J., Zhang, S., Huang, H., Liu, F., 2018. Inferring driving trajectories based on probabilistic model from large scale taxi GPS data. Physica A: Statistical Mechanics and its Applications 506, 566–577. https://doi.org/10.1016/j.physa.2018.04.073

116. Tavana, H & Mahmassani, HS 2000, 'Estimation and application of dynamic speed-density relations by using transfer function models', Transportation Research Record, no. 1710, pp. 47-57. https://doi.org/10.3141/1710-06

117. Tavana, H. (2001). Internally-consistent estimation of dynamic network origin-destination flows from intelligent transportation systems data using bi-level optimization, PhD dissertation.

118. Toledo T., Kolechkina T., Wagner P., Ciuffo B., Azevedo C., Marzano V., Flötteröd G. (2015). Network model calibration studies. In: Daamen W., Buisson C. and S. Hoogendoorn (eds).Traffic simulation and data: validation methods and applications, CRC Press, Taylor and Francis, London, pp. 141-162

119. Toledo, T., Kolechkina, T., 2013. Estimation of Dynamic Origin–Destination Matrices Using Linear Assignment Matrix Approximations. IEEE Transactions on Intelligent Transportation Systems 14, 618–626. https://doi.org/10.1109/TITS.2012.2226211

120. Tolouei, R., Psarras, S., Prince, R., 2017. Origin-Destination Trip Matrix Development: Conventional Methods versus Mobile Phone Data. Transportation Research Procedia 26, 39–52. https://doi.org/10.1016/j.trpro.2017.07.007

121. Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: Travel demand estimation using big data resources. Transportation Research Part C: Emerging Technologies 58, 162–177. https://doi.org/10.1016/j.trc.2015.04.022

122. Van Zuylen, H.J. and L. G. Willumsen (1980). The most likely trip matrix estimated from traffic counts," Transportation Research Part B: Methodological, vol. 14, pp. 281-293.

123. Vardi, Y. (1996). Network tomography: estimating source–destination traffic intensities from link data. Journal of the American Statistical Association 91, 365–377.

124. Vij, A., Shankari, K., 2015. When is big data big enough? Implications of using GPS-based surveys for travel demand analysis. Transportation Research Part C: Emerging Technologies 56, 446–462. https://doi.org/10.1016/j.trc.2015.04.025

125. Wang, F., Chen, C., 2018. On data processing required to derive mobility patterns from passively-generated mobile phone data. Transportation Research Part C: Emerging Technologies 87, 58–74. https://doi.org/10.1016/j.trc.2017.12.003

126. Wang, H., Calabrese, F., Di Lorenzo, G., Ratti, C., 2010. Transportation mode inference from anonymized and aggregated mobile phone call detail records, in: 13th International IEEE Conference on Intelligent Transportation Systems. Presented at the 2010 13th International IEEE Conference on Intelligent Transportation Systems - (ITSC

2010), IEEE, Funchal, Madeira Island, Portugal, pp. 318–323. https://doi.org/10.1109/ITSC.2010.5625188

127. Wang, M.-H., Schrock, S.D., Vander Broek, N., Mulinazzi, T., 2013. Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data. International Journal of Intelligent Transportation Systems Research 11, 76–86. https://doi.org/10.1007/s13177-013-0058-8

128. Wu, C., Thai, J., Yadlowsky, S., Pozdnoukhov, A., Bayen, A., 2015. Cellpath: Fusion of cellular and traffic sensor data for route flow estimation via convex optimization. Transportation Research Part C: Emerging Technologies 59, 111–128. https://doi.org/10.1016/j.trc.2015.05.004

129. Wu, X., Guo, J., Xian, K., Zhou, X., 2018. Hierarchical travel demand estimation using multiple data sources: A forward and backward propagation algorithmic framework on a layered computational graph. Transportation Research Part C: Emerging Technologies 96, 321–346. https://doi.org/10.1016/j.trc.2018.09.021

130. Yang, H. (1995). Heuristic algorithms for the bi-level origin-destination matrix estimation problem. Transportation Research 29B, 231-242.

131. Yang, H., Y. Iida, T. Sasaki (1991) An analysis of the reliability of an origin-destination trip matrix estimated from traffic counts, Transportation Research Part B: Methodological, 25(5), pp. 351–363.

132. Yang, X., Lu, Y., Hao, W., 2017a. Origin-Destination Estimation Using Probe Vehicle Trajectory and Link Counts. Journal of Advanced Transportation 2017, 1–18. https://doi.org/10.1155/2017/4341532

133. Yang, X., Lu, Y., Hao, W., 2017b. Origin-Destination Estimation Using Probe Vehicle Trajectory and Link Counts. Journal of Advanced Transportation 2017, 1–18. https://doi.org/10.1155/2017/4341532

134. Yim, K.N., and H.K. Lam. Evaluation of Count Location Selection Methods for Estimation of O-D Matrices. Journal of Transportation Engineering, Vol. 124, No. 4, 1998, pp. 376–383.

135. Zandbergen, P.A. (2009), Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning. Transactions in GIS, 13: 5-25. doi:10.1111/j.1467-9671.2009.01152.

136. Zhan, X., Zheng, Y., Yi, X., Ukkusuri, S.V., 2017. Citywide Traffic Volume Estimation Using Trajectory Data. IEEE Transactions on Knowledge and Data Engineering 29, 272–285. https://doi.org/10.1109/TKDE.2016.2621104

137. Zhang, J., Wang, F.Y., Wang, K., Lin, W.H., Xu, X. and Chen, C., 2011. Data-driven intelligent transportation systems: A survey. IEEE Transactions on Intelligent Transportation Systems, 12(4), pp.1624-1639.

138. Zhao, S., Zhang, K., 2017. Observing individual dynamic choices of activity chains from location-based crowdsourced data. Transportation Research Part C: Emerging Technologies 85, 1–22. https://doi.org/10.1016/j.trc.2017.09.005

139. Zhou, X., Mahmassani, H.S., 2006. Dynamic Origin–Destination Demand Estimation Using Automatic Vehicle Identification Data. IEEE Transactions on Intelligent Transportation Systems 7, 105–114. https://doi.org/10.1109/TITS.2006.869629

140. Zhou, X., Mahmassani, H.S., 2007. A structural state space model for real-time traffic origin–destination demand estimation and prediction in a day-to-day learning framework. Transportation Research Part B: Methodological 41, 823–840. https://doi.org/10.1016/j.trb.2007.02.004

141. Zilske, M., Nagel, K., 2014. Studying the Accuracy of Demand Generation from Mobile Phone Trajectories with Synthetic Data. Procedia Computer Science 32, 802–807. https://doi.org/10.1016/j.procs.2014.05.494

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
Fondo Sociale Europeo, Azione I.1 “Dottorati Innovativi con caratterizzazione Industriale”

UNIONE EUROPEA
Fondo Sociale Europeo