universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Original-Destination matrix estimation problem – comparison of different distance measurements“

verfasst von / submitted by

Sofya Kalinushkina

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 915
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	
Betreut von / Supervisor:	Univ.-Prof. Dr. Karl F. Dörner

ACKNOWLEDGMENT

First, I would like to thank the supervisor of my thesis Univ.-Prof. Dr. Karl F. Dörner for giving me the opportunity to work on this interesting topic and for his valuable feedback during my work.

In addition, I would like to thank Mag. Andreas Krawinkler for his ongoing support and his critical and applicable feedback. He always took the time to help me through various technical issues and guided me into the right direction.

At the end, I would like to thank my parents and my friend for always staying supportive and optimistic in various ways and for being patience and understanding during my work on the thesis.

Vienna, April 16

Contents

List of Figures

List of Tables

List of Abbreviations

AVI	Automatic Vehicle Identification
AVL	Automatic Vehicle Location
BI	Bayesian Inference
CPU	Central Processing Unit
C-C	Customer-Customer
GLS	Generalized Least Squares
EA	Evolutionary Algorithm
EM	Entropy Maximization
IM	Information Minimization
OD	Original-Destination
ML	Maximum Likelihood
NN	Nearest Neighbor Heuristic
SA	Simulated Annealing
SD	Standard Deviation
SPSA	Simultaneous Perturbation Stochastic Approximation

1 Introduction

Origin-destination (OD) matrices are highly valued for various research fields as traffic management, transportation systems operation, for purpose of transportation planning, design and analysis in the longer term (Nair et al., 2013). Origin destination matrix represents traffic demand in an area which is "essentially a table containing the number of trips from each origin to each destination in the area" as Abrahamsson (1998) wrote in his survey. Hongjun et al. (2010) define that for elaboration of traffic management schemes the OD-matrices should be obtained rapidly and accurately.

According to the literature, the OD-matrix estimation models can be divided into two categories, namely static (time-independent) and dynamic (time-dependent) based on its application. The traffic flows in static models are time-independent and an average OD demand is assumed stationary, that is to say time-independent models describe the traffic situations on the average.

Peterson states that such models demand not as much input data and exist "well-established analytical models". There are many methods, which solve time-independent OD-matrix estimation models. Whereas in the last two decades various dynamic algorithms are proposed which are designed for short-term strategies like traffic control on freeways, route guidance, intersections etc. (Bera and Rao, 2011).

As an extension of the static models dynamic ones on the other hand can reproduce an impact of the special events on the traffic according to Peterson (2007). For time-dependent models, the input data should contain detailed information about the traffic situation. When time as a dimension is considered, the complexity of the estimation problem increases significantly. Until now most time-evolving models are based on simulation.

Commonly, the OD data were obtained from home interviews by conducting big scale surveys such as roadside surveys or on-board surveys. However, the collection of this kind of data is expensive in terms of money and work force and at the same time, it contains internal errors because of "sampling processes and problems in elaboration" (Perrakis et al., 2012). Moreover, by the time how the survey data will be definitely collected and processed, the obtained OD data may become quickly outdated. (Bera and Rao, 2011).

However, there is some software that can estimate OD-matrices. Nevertheless, depends on the study area it is not always flexible as much as it is required in respect of the type of available

data. Bera and Rao (2011) confirm that research is still going on to estimate reliable OD-matrices efficiently.

In this thesis, we emphasize the distance measurements instead of demand flow for static origin-destination estimation problem, that means that the thesis focuses on the methods and the concepts enabled to deliver computation results within reasonable time frame and still are reliable and accurate. In order to define an appropriate approach we study dependence between an aggregative level of input data and an accuracy of the obtained results.

Therefore, the new complex concept including several methods has been proposed and analyzed first on the testing benchmark, later on the real-world examples with different complexity. The proposed and described concept confirms his features as an efficient method in compassion to the available software and as a method suitable for application on the real-world cases.

The remainder of this master thesis has a following structure. Chapter 2 provides the reader a short and focused literature review on the static OD-matrix estimation problem as well as on basic literature sources for the dynamic version of estimation problem. Chapter 3 starts by defining the OD-matrix estimation problem and providing basic notations on complexity and further notes on the question of research. After that, the OD-matrix estimation methods explained in more details.

Chapter 4 discusses the analytical framework and thereafter examines the model on the testing benchmark and analyzes the obtained results. The fifth chapter starts by presenting the case studies and then discusses the results of computational experiments. Finally, Chapter 7 draws a conclusion and provides some ideas and comments for further research.

2 Literature Review

Methods of estimating time-independent OD-matrices.

In the past decades, the time-independent OD-matrix estimation problem has been well studied. In the literature, there are several surveys that review the estimation problem like Abrahamsson (1998), Bell and Iida (1997), Willumsen (1981) etc.

Initially the researchers tried to relate the OD-matrix as a function of models (like the gravity models) with related parameters. Several researchers like Robillard (1975), Högberg (1976) used **Gravity model** based approaches and some (Tamin and Willumsen, 1989; Tamin et al., 2003) used Gravity-Opportunity based models for estimating OD-matrices. However, Willumsen (1981) notes that because the gravity model are not able to handle with accuracy external trips it is the main drawback of those models.

After the Entropy concept has been introduced by Wilson (1970) **Information Minimization** (IM) and **Entropy Maximization** (EM) techniques are used as tools for building models in transportation, urban and regional planning context. Willumsen (1978) implemented EM to estimate OD-matrix and IM approach was used by van Zuylen (1978). Unfortunately, IM and EM based models do not consider the uncertainties in traffic counts that counts as their shortcomings.

Among other OD estimation methods, there are several **statistic approaches** such as maximum likelihood (ML), generalized least squares (GLS) and Bayesian inference (BI), which partially based on the EM method. Bayesians inference framework method has been first introduced by (Maher, 1983) for the OD-matrix estimation. GLS estimator based approach has been studied by Cascetta (1984) and Bell (1991) etc. Bell (1983), Spiess (1987), Cascetta and Nguyen (1988) studied the models based on the most likelihood estimation. These approaches were generalized and further improved by other authors (Yang et al., 1992; Lo et al., 1996; Hazelton, 2000).

Moreover, some researchers proposed **bi-level models** for estimating OD-matrices. Such models was described in the paper of Kim et al. (2001) that also discussed the problem of model developed by Yang et al. (1992) and suggested an alternative model using genetic algorithm. Lately, Codina et al. (2006) presented two alternative algorithms solving bi-level problems and examined them on small networks. Recently, Lundgren and Peterson (2008) elaborated a heuristic for bi-level problem and tested their algorithm on a large size network

of Stockholm. There are some more studies (Hazelton, 2003; van Aerde et al., 2003; Li, 2012 etc.) based on statistical approaches.

As it was mentioned in the previous chapter, traditionally survey method was considered to be used for obtaining OD data by conducting large-scale surveying studies, which require amounts of money and time. Recently, Jin, 2016) in his paper suggested new **simplified survey method** without large-scale OD surveying studies in a small- and medium-sized cities. The proposed approach was tested in the small city of Chungju, South Korea, and obtained results outperformed exiting methods from the literature due to significant cost reduction as well as suitability for simplification that enable to speed up the process of collecting data and therefore gives an opportunity to reproduce and update the data more frequently.

Methods of estimating time-dependent OD-matrices.

Beside static estimations, dynamic estimations have developed gradually. Cascetta et al. (1993) and Sherali and Park (2001) solved the dynamic OD estimation problem using **standard gradient methods**. Considerable amount of the works in OD-matrix estimation so far have focused on the more complexed algorithms that improve estimation and forecasting of OD-matrices and at the same time requires less time for processing. To those works belongs the papers of Bierlaire and Crittin (2004), Kattan and Abdulhai (2006) Zhou and Mahmassani (2007).

Others suggested to add available **automatic identification data** or data from **traffic counts** (Asakura et al., 2000; Antoniou et al., 2004; Dixon and Rilett, 2005) to name a few. **Automatic vehicle location (AVL) data** are part of automatic identification data, which can provide a large fragment of OD flow data. AVL data have a big potential, therefore they received attention from the researchers. Those data can be observed from in-vehicle traffic sensors like GPS or GSM. The following authors like van Aerde et al. (1993), Ashok and Ben-Akiva (2000), Caceres et al. (2007) used AVL data to develop various models for the estimation of OD flows .

Another OD flow data source is represented by **automatic vehicle identification (AVI) data** that are also important for estimating demand flows of dynamic OD-matrices. Such kind of data can be observed from AVI sensors such as electronic-toll collection devices, infrared cameras, Bluetooth, Wi-Fi, etc. (Djukic et al., 2015). Several models that used AVI data in

order to estimate OD flows have been developed by Asakura et al. (2000), Dixon and Rilett (2002) and Antoniou et al. (2004).

The data from portable gadgets such as smartphones, GPS navigational devices (Moreira-Matias et al., 2016), Bluetooth devices (Barceló et al., 2010) are nowadays a sources of a real-time information which provide an opportunity to improve an accuracy of OD estimation. Lou and Yin (2010) and Frederix et al. (2011) proposed a method, which decompose network into small zones in order to deal with high dimensional OD estimation problem (Djukic et al., 2012).

Other method that frequently used for estimating and predicting OD-matrices is the **Kalman filter** algorithm (Djukic et al., 2012). The standard linear Kalman filter theory was first time mentioned by Okutani (1987) who used this theory to get "optimal estimates of the state vector in each time interval". Ashok and Ben-Akiva (1993) proposed a Kalman filtering approach in order to be able to update an OD-matrix constantly. Kachroo et al. (1997) studied the application of Kalman filtering approaches for network OD-matrix estimation from link traffic counts. Sherali et al. (1997) enhanced a constrained optimization algorithm but with high computational cost.

However, Kalman filtering formulation has disadvantages because it requires sufficient data and intensive matrix operations. Dan et al., 2011) proposed adaptive filtering process based on the simplified and improved **Sage-Husa filtering algorithm** that increases the filter accuracy.

In the literature the dynamic origin–destination estimation problem has also been solved applying the **simultaneous perturbation stochastic approximation** (SPSA) algorithm (Toledo and Kolechkina, 2013). Different researchers (Balakrishna and Koutsopoulos, 2008; Cipriani et al., 2011; Spall, 1998; Spall, 1999; Spall, 2005) have used this algorithm for the problem solution. Tympakianaki et al. (2015) underline the main advantage of SPSA is that it allows formulating the problem quite general.

Moreover, several authors suggested **meta-heuristic** approaches. These methods includes evolutionary algorithm (EA) (Kattan and Abdulhai, 2006) and simulated annealing (SA) (Stathopoulos and Tsekeris, 2004). Generally meta-heuristics by definition are able to detect not only local optima but also global optima which is a major benefit of those methods However, normally they demanding a frequently evaluation of objective function that usually

computationally expensive in the context of time-evolving OD estimation (Toledo and Kolechkina, 2013).

Later on, Kattan and Abdulhai (2010) in their paper compared basic EA with a hybrid EA and parallel EA. They conclude that the parallel EA model outperforms the basic EA as well as hybrid EA algorithms in terms of computational time and quality of the solution. Nevertheless, the results of the hybrid and parallel EA runs yield savings in computation resources and find an improvement in the solution quality compared to the basic EA model, which still outperforms the local-search algorithm.

Until now both static and dynamic OD-matrix estimation problems have been studied by many researchers, which developed various models using different solution algorithms. The researchers have mostly adopted the statistical approaches like ML, GLS or BI for solving the time-independent problem for instance.

The review shows that most of the algorithms developed for static OD-matrices have its own advantages and drawbacks and was implemented on small networks. However, the most important consideration required its application for real world networks. Nevertheless, there are a few approaches that could be found in the literature, which focused on the application on the large size networks.

The time-independent OD-matrices are easier to estimate as compared to dynamic ones because real-time traffic information for all the OD pairs is not always available. In spite of that, the OD estimation problem stays resource and CPU intensive because the above-mentioned methods have to concern both high dimensional OD-matrices and the methods' complexity.

3 Original-Destination Matrix Model

This chapter is organized as following. In general, in the first part the model and its mathematical formulation are described and the aim of the research is specified. The second part is devoted to approaches from the recent researches used for estimating the OD-matrices and overview the under study methods.

3.1. Model description

In this subchapter will be presented the formulation of the problem in a general way and then discussed the complexity of the problem itself as well as the goal of the research.

3.1.1. General formulation

A distance matrix to wide extend in mathematics, information science and particularly in graph theory, is a matrix or two-dimensional array that includes the length between each pair of the elements in a given set. Rows corresponds to the origins and columns to the destinations of the users (Guihaire and Hao, 2008). In the applications cases of the graph theory the elements are frequently referred to as nodes, points or vertices.

Let us introduce some notations: let g stays for a OD matrix and Ω indicates the set of feasible OD-matrices as it proposed by Peterson (2007). Capital N represents a set of nodes and A represents a set of links in the traffic network accordingly. Each link $a \in A$ is a certain segment of a street or of a transit connector, e.g. a "fake" connection between a living area and the main streets or motorways. We assume that the network is a directed graph so the two-way streets are designed separately, i.e. such streets will be represented by two links in the opposite directions.

In the problem of estimating OD-matrix our goal is to determine a feasible OD-matrix $g \in \Omega$, with $g = \{g_i\}$, $i \in$ I. After that, the OD-matrix assigned onto the links in the network. The assignment follows an assignment proportion matrix denoted by $P \in \{p_{ai}\}$, $i \in I, a \in A$, where p_{ai} stays for the OD demand's proportion of g_i that takes link a. Peterson (2007) in his work emphasized that the notation $P = P(g)$ used because these proportions generally depend on the volumes of the traffic, i.e. on the OD-matrix.

When the assignment of the OD-matrix to the network is complete, the OD-matrix induces a volumes of the link flow $v = \{v_a\}$, $a \in A$, on the links in the network. It is assumed that

detected flows $\tilde{v} = \{\tilde{v}_a\}$, are accessible for the links' subset, $a \in \tilde{A} \subseteq A$, and a target matrix $\hat{g} \in \Omega$ also is achievable.

The OD-matrix estimation problem proposed by (Peterson, 2007) can be formulated as follows:

$$\min_{g,v} F(g,v) = \gamma_1 F_1(g,\hat{g}) + \gamma_2 F_2(v,\tilde{v}),$$
$$s.t. \ \sum_{i \in I} p_{ai}(g) g_i = v_a, \ \forall a \in \tilde{A}, \quad (1)$$
$$g \in \Omega$$

The function $F_1(g,\hat{g})$ in the (1) equality are generalized measures of the distance between the estimated OD-matrix g and the target matrix $\hat{g}$. The function $F_2(v,\tilde{v})$ also in the (1) equality are generalized measures among the estimated link flows v and the detected link flows $\tilde{v}$. We assume that these functions are convex and they can be modeled regarding the quality variations of the given data.

The non-negative parameters γ_1, γ_2 indicate the uncertainty in the information supplied by $\hat{g}$ and $\tilde{v}$. Thus the problem thus can be interpreted as a two-objective problem, where F_1 and F_2 express the objectives, and γ_1 and γ_2 correspond to the weighting factors. If an extreme case when $\gamma_1 = 0$ is considered, then the target matrix will have no impact, and on the other hand when $\gamma_2 = 0$ is under consideration, the target matrix will be copied and the link flows will have no impact.

3.1.2. Complexity and question of research

The complexity of a problem refers to the efficiency with which it can be solved. When the problem becomes more complex, than it is harder to exactly solve it within reasonable time. As Brucker et al. state, the algorithm efficiency can be estimated by calculating the running time needed for solving the problem with a certain input size. In this context so called "easy" and "hard" problems can be distinguished. An algorithm can solve "easy" problem in polynomial time. A "hard" problem, usually noted as "NP-hard", on the other hand describes problems that are generally hard to solve to optimality.

The OD- matrix estimation problems in general are not belongs to the NP-hard problems but the time-dependent cases become more complex and require new faster methods that can reflect challenges of reality.

Despite the fast, that estimating the shortest distance between given pairs of nodes considered as an elementary operation in a broad diversity of applications, such operations should be solved in a few milliseconds because the applications require a quick responses on their order. Nowadays, algorithms which can exactly evaluate the shortest path distance (Dijkstra, 1959) fail on real-world networks, because of their inability to solve the problem within a reasonable timeframe. Also totally precalculating the shortest paths and then saving them is usually infeasible by reason of lack of memory space that can store an extremely large matrices (Maruhashi et al., 2012).

Moreover, there are elements that complicate the OD-matrix problem. Namely, when the geographic data (geodata) are considered certain limitations should be taken into account. For instance, geographical elements as mountains, forests, rivers that e.g. can cross high populated or industrial areas are definitely restrict a possibility to determine the shortest path.

Hence, the question of our research is to analyze a correlation between aggregation levels and exactness of real world distances, i.e. how an increase in aggregation level reduces a precision of real world distances. Under the aggregation level is to understand an input data about customer positioning either geodata or raster data.

3.2. Estimation methods

This chapter presents in the first part successful achievements of other authors that dialed with similar methods and in the second part explains the estimation approaches, which are examined and analyzed on the small instance in the third part.

3.2.1. Previous accomplishments

In the literature the OD-matrix estimation problem are very common but the problem of distance measurements was not stressed enough. Unfortunately, we could not find enough works that use similar concept for the OD-matrix problem. However from recent researches the most similarly principle are met in the book "Effiziente Entfernungsberechnung durch Graphenreduktion bei Transportplanungen" written by Prestifilippo (2003).

The author studies various graph reduction approaches that are able to efficiency estimate distances by transportation planning, i.e. the methods, which lead to speed up processing of optimal routes. Because presently an available digital data about the structure of the transportation system are anomaly high and very detailed, such approaches aim to reduce data in order to evaluate an optimal route without loss of relevant information.

Prestifilippo analyzed european transportation network as well as the transportation system of the european states. He concluded that an existing simple selection method is not able to deliver an optimal solution within reasonable timeframes. In addition, profitability of the procedure is generally an exception than a rule and frequently an implementation leads to the boundaries of memory resources.

Therefore Prestifilippo proposed new methods that are independent but at the same time compatible. The efficiency of those approaches is demonstrated on the cases and on the specific practical examples for various logistic concepts. Interested reader refer to the origin book of Prestifilippo, 2003).

The first method called Bubble is an intelligent data selection procedure. The core idea includes a construction of prior defined bubbles around each node, a sorting nodes and edges in hierarchical order within a subset and a selection of relevant nodes and edges from the top level. Irrelevant data are excluded from further consideration. The method by the way is highly influenced by an input amount of nodes and a size of the bubbles that are varied.

Other methods proposed by Prestifilippo are variations of graph reduction, which eliminate an irrelevant nodes and edges that lead to distance stretching. The approaches denoted as Sackgasseneliminierung (blind alleys elimination), Streckendesegmentierung (distance desegmentation) and Teilgrapheneliminierung (subgraphs elimination). All those methods compare to the Bubble approach are lossless regarding information.

Finally, the author examines method that combined Bubble and graph reduction approaches. Such a method uses advantages of both approaches that improve processing time as well as computation results. Prestifilippo concludes that proposed him methods could be applied either separately as a distance measurements procedure for various cases or can be integrated through sequential inclusion in a process. The construction of the reduction process allows combining them in any way.

3.2.2. Raster concept

Before we start to explain the raster concept and its main features, a couple of questions should be answered, e.g. what is the raster in itself and what is its background?

In a spoken language, words as grid, framework, matrix or network are used as a synonym for raster. Collins Dictionary defines a word grid as "something, which is in a pattern of straight lines that cross over each other, forming squares". Oxford Dictionaries by the way describes

grid such "a network of lines that cross each other to form a series of squares or rectangles". Generally, the grids are used on maps to find a particular location.

Thus, raster data or as it is normally called gridded data are nothing but geodata. In this master thesis the raster data were received from Statistic Austria. Due to statistical requirements, Statistic Austria produces new geodata and updates older ones on an ongoing basis.

The results are available either for download free of charge or for a fee on the website. In this thesis, various grid sizes are explored such as 250m, 500m and 1000m. The provided data were later reworked and reduced in such a manner that the points are only distributed over observed area excluding uninhabited places or places without any industrial activities.

For better understanding, we will briefly discuss a concept of raster points which was applied for solving OD-matrix problem. The core idea of a concept is to divide explored space into a matrix of equally sized cells and afterwards replace customer locations by raster points. Obviously, the idea has its advantages and drawbacks.

As pro arguments can be mentioned an ability to simplify the model and to speed up processing time. On the other hand, the main disadvantage of the grid is generalization and loss of uniqueness. Therefore, some trade-off has to be determined. The presented below Fig. 1 shows visually a part of grid data with grid size 250m. As it mentioned in Chapter 2, there are no grids in the forest areas or for instance in the Danube.

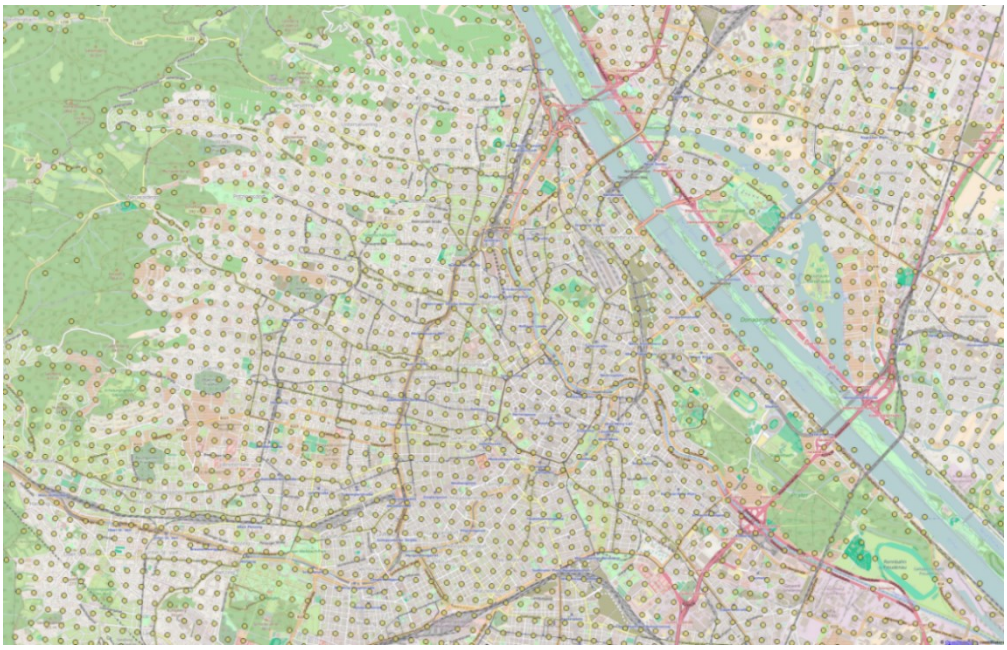

Figure 1. Distribution of raster points with size 250m

On the Fig. 2 can be seen the simplified scheme of the concept. Green spots 1-8 here mean raster points and orange triangles C1, C2, C3 and C4 – customers. Orange dotted curves between C1 and C2, C3 and C4 are feasible direct routes. Green curves that connect customers C1 and C2 and customers C3 and C4 via raster points 2 and 8 are feasible alternative routes.

The main principle is to determine the nearest raster point for each customer using NN algorithm and build so-called three-parts-route between chosen points. Length of latter will be then compared with length of the direct route and if a deviation is less than preselected threshold, then an origin route can be replaced by three-parts-route.

The reasons that motivated us to introduce a raster concept were several. Firstly, currently software detecting direct optimal routes is time consuming and fails for large instances. Secondly, raster matrix enables to calculate all distances between existing raster points during a pre-processing stage.

Consequently, we will always have an access to the obtained results and despite of requirement of sufficient memory resources, an increase in processing time during an evaluation procedure for further individual distance matrices is significant. In other words, introduced grids permit to avoid each time path calculations from the street graph because the data will be read off the pre-computing results data.

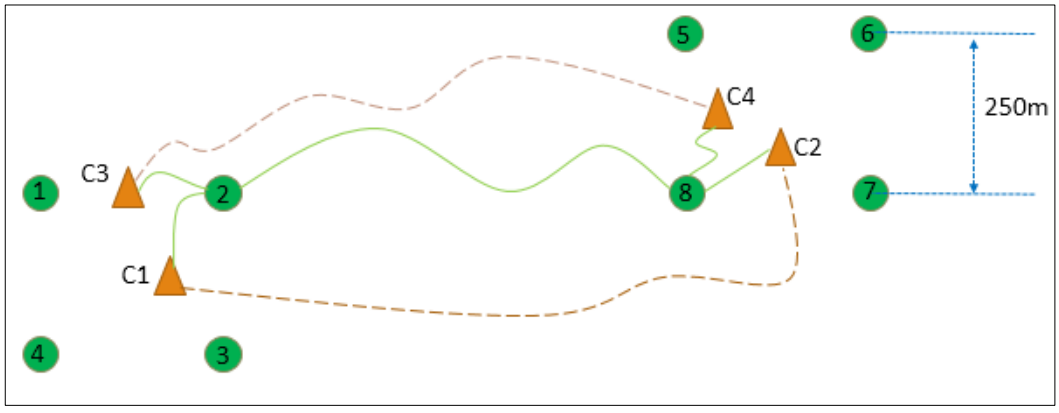

Figure 2. Scheme of raster point concept

Thus, our goal is to determine method to estimate distance between raster corner points and customers locations in such a manner that deviation from original route is kept within reasonable bounds.

3.2.1. Study methods

Current subchapter will focus on and describe in more details methods as well as software that are implemented and analyzed in the Chapter 6. Please note that all approaches are examined on randomly chosen customers and distances expressed in kilometers.

The distance measurements methods are based on the well-known categories such as Shortest Path, Euclidean distance, Manhattan distance etc. We allow ourselves to remind the main features of the above-mentioned terms. An interested reader are referred to the paper of Ahuja et al. (1990) and the book of Deza and Deza (2009). Here is only briefly overview provided.

The **Shortest path problem** is the problem in the graph theory, which determines a path between two nodes in a graph. Ahuja et al. in their paper discuss the fastest algorithm that solves the shortest path problem efficiency. Among that the most important and reliable algorithms to solve the problem are considered Dijkstra (Dijkstra, 1959) and Bellman-Ford (Cormen et al., 2001) algorithms.

Euclidean and **Manhattan** distances in mathematics are measures of distance between two points in a metric space. Euclidean distance is ordinary called straight line and can be calculated using general formula if *x* and *y* are coordinates of two points (Deza and Deza, 2009):

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

Manhattan distance is "the distance between two points in a grid based on a strictly horizontal and/or vertical path that is along the grid lines" (Wikipedia). The formula that is used for computation are the following but only valid if *x* and *y* are coordinates of two points (Deza and Deza, 2009):

$$d(x\,,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (3)$$

Now we will discuss the approaches and supported software. As a supporting tool in this thesis Microsoft tool MapPoint Europe 2013 are applied. The tool was implemented in order to calculate real distance between two objects. The derived results used as a standard measure.

Available software is able to evaluate the shortest path that minimized total travel distance and the shortest path where objective is to minimized total travel time. Due to avoid likely common misunderstanding between those two terms the second one that minimized time will be further preferred as the fastest path.

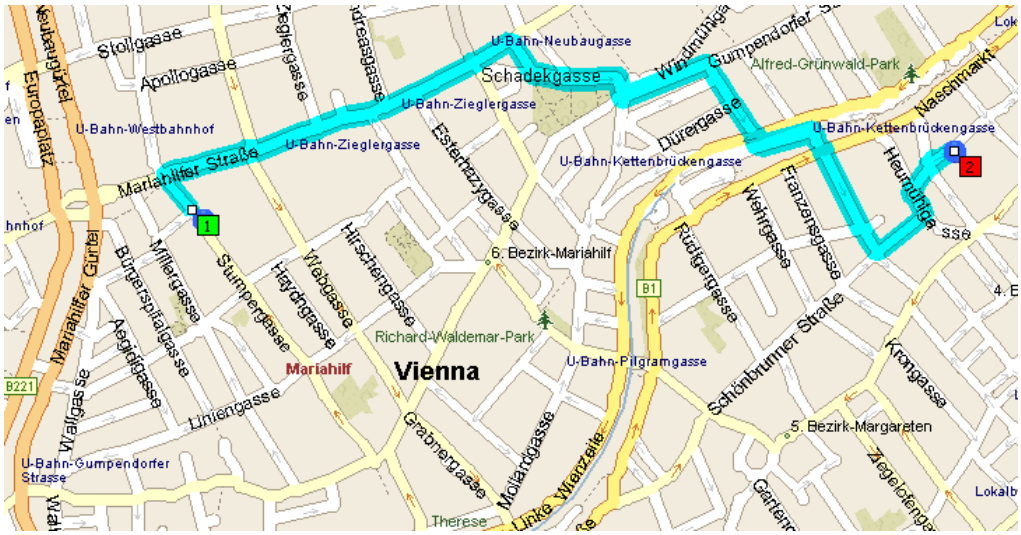

Figure 3. An example of the shortest path between two points

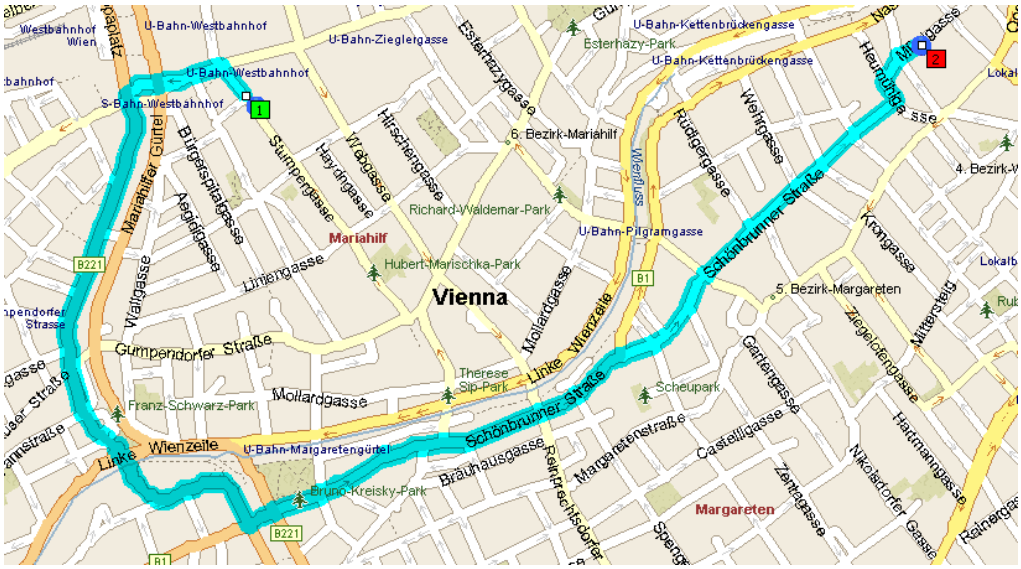

Figure 4. An example of the fastest path between two points

Fig. 3 and Fig. 4 demonstrate the shortest and the fastest route between two randomly chosen locations with total distance 2,2km and 3,5km and total travel time 14 minutes and 11 minutes respectively. Apparently, that the shortest route is preferable over the fastest because this thesis stress on the static model and time minimization objective is less important for the aims of the current research.

Raster - Raster method

The next approach is a part of the raster point concept described above in 4.2.2. The Raster-Raster method evaluates only the length between grid's corner points as the shortest path. Figure 5 provides the result of implementation the NN algorithm for raster size 250m. Yellow dots on the Fig. 5 are raster units, green dots are the nearest points and blue points represent positioning of customers.

It is important to notice that Raster-Raster method has some limitations. Firstly, despite the fact that the NN algorithm are very quickly method and yields a near optimal solution, the selected points could lead however to the significant tour stretching due to "greedy" nature of the algorithm. Secondly, there is likelihood that not between all pairs of raster points is a connection because raster data are aggregated data.

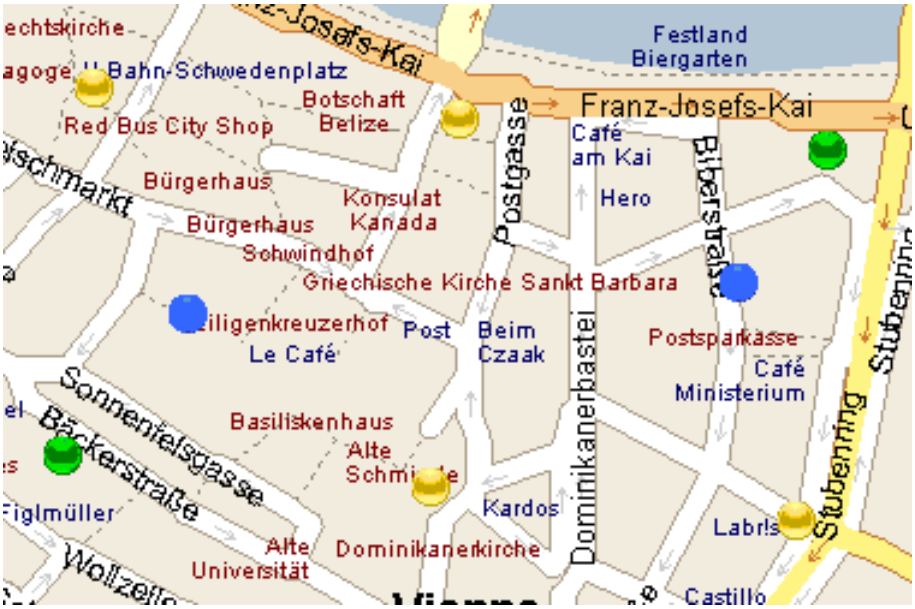

Figure 5. Illustration of the nearest to the customers 250m raster units

Fig. 6 represents visually the shortest route between raster points. In order to ensure that chosen raster unit size is the most suitable for a current problem, several experiments with different raster size, namely 500m and 1000m, will be run and the results presented in the Chapter 5.

Figure 6. The shortest route between selected raster points

C-C Euclid, C-C Manhattan.

From the theoretical point of view, distance could be measured as a Euclidean or Manhattan distances and known as beelines. The distance measured between two customers using Euclidean methods is denoted as C-C Euclid and using Manhattan method – C-C Manhattan. Beelines in our case are likely to be not applicable for every cases due to the reason that existing transportation network in the most cases is not able to offer such connections.

However, both approaches are examined for the comparison reasons. On the Fig. 7 the results of the both approaches are presented. Grey and purple lines draw beelines computed following either Euclidean or Manhattan methods respectively. Obviously that the distance measured by the Manhattan distance method will be longer, than the Euclidean by definition but in real-world cases the latter could better reflect the reality.

Figure 7. Example of the C-C Euclid and C-C Manhattan approaches

Euclid-Real-Euclid, Manhattan-Real-Manhattan

As it was explained in the previous subchapter, the raster concept builds a so-called "three-parts-tour". Interested reader referred again to the Chapter 4.2.1. Therefore to be able to apply those concept and increase it accuracy different distance measurements methods such as either Euclidean or Manhattan distances have been applied to estimate the length between customer position and grid's corner points. Consequently Euclid-Real-Euclid as well as Manhattan-Real-Manhattan approaches denote the routes consisting of three parts in such a manner that the length between customers and corner points are evaluated using the formulas either for Euclidean or for Manhattan distances respectively. Schematic the "three-parts-routes" are presented on the Fig. 8 and 9.

All the method mentioned in the current chapter will be implemented on the practical example and the analysis is provided below in the subchapter 5.2

Figure 8. Example of Euclid-Real-Euclid method

Figure 9. Example of Manhattan-Real-Manhattan method

.

4 Model Analysis

In order to evaluate estimation methods that are interest of our study and presented in the last section, this chapter is dedicated to the preliminary implementation. For a better understanding of the process, the first part will be discussing the analytical framework of the analysis, including the computational characteristics as well as the statistical indicators. The second part will present the results of implementation on a real instance and discuss their meaning.

4.1. Analytical framework

This subchapter provides a briefly review of the indicators and indexes that are used in analysis of the method's outcomes.

4.1.1. Computational time (CPU)

There are many different definitions of CPU but here will be provided only one of them.

"CPU time is the amount of time for which a central processing unit (CPU) was used for processing instructions of a computer program or operating system." (Wikipedia) In other words, it is the measurement of the length of time that is used as an indicator of how much processing is required for a process.

Normally programs and applications do not use the processor 100% of the time that they are running. The CPU time is actual time when the program uses the CPU to perform tasks such as mathematic or statistic operations. It is also known as processing time or computational time. Todd as well as some authors calls it running time.

The computational time is commonly "measured in clock ticks or seconds". (Technopedia) In the current work seconds is used as a unit of measure. All in all the processing time is an important indicator for estimation a performance speed of the methods that allows to conclude if testing algorithms are enough efficient.

4.1.2. Quartile

Quantile is statistic and a general way of describing a point in a set of numbers where a certain proportion of numbers is less than that reference point. It is simply the value that corresponds to a specified proportion of a sample. Quartiles gives us an information about how the data are spread between the breaking points also called quarters (Laerd Statistics).

The second quartile is also named the median (Wikibooks). Summary statistics such as the median, first quartile and third quartile are measurements of position, i.e. these numbers indicate where a specified proportion of the distribution of data lies. For instance, the median is the middle position of the data under study, i.e. 50% of the data have values less than the median. Similarly, 25% of the data have values less than the 1st quartile and ¾ of the data have values less than the third quartile (About education). Although not universally accepted, but there is also exist the fourth quartile that is the maximum value of the set (Wikibooks). In statistics, median is regarded as a better measure of centrality and spread versus average.

4.1.3. Standard deviation

Bland and Altman (1996) defined the standard as "a measure that is used to quantify the amount of variation or dispersion of a set of data values". Standard deviation, be referred to as SD, also denoted by the Greek letter sigma σ or s in the literature. There is definitely more definitions of standard deviation. For instance, it can be defined as "a measure of the spread of scores within a set of data" (Laerd Statistics). A measure of spread, sometimes also called a measure of dispersion, which is used to describe the variability in a sample.

For better understanding, the standard deviation is a statistic that tells how tightly all the various examples are clustered around the mean in a set of data. In case when the examples are tightly bunched together, the standard deviation is small. When the examples are spread apart, it tells that standard deviation is a relatively large (Robert Niles).

Calculation of standard deviation is important for correctly interpretation the data and it is frequently applied for comparing a model with real-world data. Its main advantage is that it has the units of the original data. In our analysis of computational results the standards deviation shows an average deviation from the mean value.

4.1.4. Extremum points

Extremum (pl. extrema) are in mathematical analysis the maximum (pl. maxima) and minimum (pl. minima) points of a function, i.e. the largest and smallest values which the function reaches within a given range. For the set of data the maximum and minimum of a set indicate the greatest and the least values in a set, respectively (Wikipedia). Extremum is the simplest order statistic and by computing the difference between maximum and minimum values, the spread of the data can be estimated (Siddharth Kalla).

In our thesis we consider absolute values of maximum or minimum. In other word, for instance, the obtained maximum value means that this value of the function is larger than any other values in the interval of interest or in the data set. The extrema points of our computational results enable us to estimate the performance of the provided methods as well as to value the worth cases in comparison to the results of other methods from the literature.

4.2. Preliminary test

In order to be able to demonstrate all abovementioned methods, a benchmark (testing set) of ten customers that are located in Vienna has been generated.

Numbering of testing sets has to follow some rules. The first number (before slash) means number of customers that are included in the benchmark and second number (after slash) means raster size, for instance 10/250 means that the set consists of 10 customers and uses raster size 250m. Here is the map illustrating customer locations.

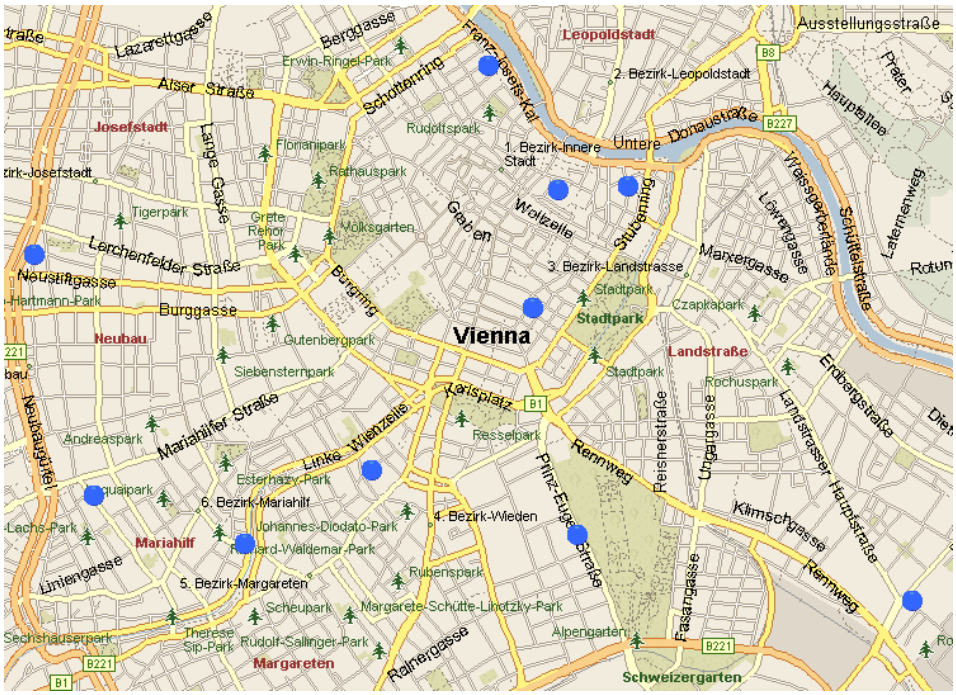

Figure 10. Positioning of ten customers from testing set 10/250

At the first step, the real distances for a generated set are evaluated by using Microsoft software MapPoint. Please note that distances as well as further deviations measured in kilometers. These distances were measured as Shortest Path and as Fastest Path. The resulted distance matrices can be found in Appendix as Table 16 and Table 17. As can be seen the both matrices are asymmetric due to the structure of current transportation system, i.e. the

measured distance is highly depended on the direction and the transportation network, for instance, existence of one-way streets, blind alleys etc.

At the next step, the available data have been analyzed and the results are summarized in Table 1.

Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
Shortest path	0:32	-2,28	-0,59	-0,25	-0,05	0,00	0,46
Fastest path	0:32	0,00	0,05	0,25	0,59	2,28	0,46

Table 1. Computational results of real distance for case 10/250.

Table 1 sums up the results of comparison the shortest path with the fastest path and vise verse. Obviously, the shortest path method determines the routes more efficient. The third quantile shows that 75% of the determined routes are shorter and the forth quantile also proofs that the routes are no longer as the fastest paths. The deviation between fastest and shortest paths confirms the same. All values are positive that means that distance matrix determined with the fastest path method finds slightly longer connections. Therefore, the shortest path method is preferable because it delivers better results within the same time. Thus, all further results of above-mentioned methods will be compared with the results of the shortest path method.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
10/250	C-C Euclid	0:45	-2,15	-0,94	-0,34	-0,06	0,95	0,59
10/250	C-C Manhattan	0:45	-1,49	-0,27	0,0	0,5	2,4	0,69
10/250	Raster-Raster	0:44	-1,49	-0,59	-0,17	0,0	0,39	0,41
10/250	Euclid-Real-Euclid	0:45	-1,27	-0,38	0,0	0,19	0,6	0,41
10/250	Manhattan-Real-Manhattan	0:45	-1,2	-0,31	0,0	0,22	0,7	0,41

Table 2. Results of the comparison for the case 10/250.

All the methods mentioned in the Chapter 4 were examined on the generated testing set of 10 customers. Than the results have been compared with the results of the shortest path method and finally sum up in Table 2. However, Table 2 includes exclusively the raster system with raster unit 250m and as we can see, has some negative values. These values can be interpreted as following: the methods considered in the table are able to determine the more efficient routes. For instance 50% of all routes generated by the each method from the table are shorter approximately 0,5km or as long as origin ones.

The experiments continue with the raster system with raster units 500m and 1km. Due to extension of the raster sizes, firstly from 250m to 500m and secondly from 500m to 1000m, the results tables differ from the previous one. Table 3 and Table 4 that present the results for further cases 10/500 as well as 10/1000 for instance in comparison with Table 2 contain slightly less negative values.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
10/500	C-C Euclid	0:44	-2,15	-0,94	-0,34	-0,06	0,95	0,59
10/500	C-C Manhattan	0:44	-1,49	-0,27	0,0	0,5	2,4	0,69
10/500	Raster-Raster	0:44	-1,53	-0,56	-0,16	0,0	1,16	0,45
10/500	Euclid-Real-Euclid	0:43	-1,27	-0,08	0,13	0,35	1,69	0,45
10/500	Manhattan-Real-Manhattan	0:43	-1,2	0,0	0,29	0,57	1,84	0,47

Table 3. Results of the comparison for the case 10/500.

These can be explained that by implementing wider raster matrix corner points are definitely farther from the customer that leads to increasing the total route length. However, the extension of the raster units diminishes an amount of raster points in the raster system in size. In addition, such reduction can speed up the computation process that can be seen in the column CPU of the Table 3 as well as Table 4. For the provided case the reduction is couple

of seconds that likely insignificant for such a small case but can has a crucial effect by examine on the large benchmarks.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
10/1000	C-C Euclid	0:43	-2,15	-0,94	-0,34	-0,06	0,95	0,59
10/1000	C-C Manhattan	0:43	-1,49	-0,27	0,0	0,5	2,4	0,69
10/1000	Raster-Raster	0:43	-2,03	-0,96	-0,47	0,0	1,03	0,65
10/1000	Euclid-Real-Euclid	0:44	-1,2	-0,14	0,1	0,73	2,27	0,64
10/1000	Manhattan-Real-Manhattan	0:44	-1,02	0,0	0,36	0,91	2,5	0,66

Table 4. Results of the comparison for the case 10/1000.

Besides that, C-C Euclid and C-C Manhattan methods demonstrate absolutely the same performance due to the fact that grids are not involved in computation process. Therefore, we will run those approaches only ones for the new amount of customers but not repeat for different raster sizes.

Moreover, last two methods listed in the Tables 3 and 4 demonstrate appropriate performance. For instance for case 10/250 the third quantile shows that 75% of the "three-parts-routes" are around 200m longer than the shortest paths. Obviously, by increasing raster size the deviation also increase but if raster size double the deviation rise by ca. 70%. We assess such performance very high.

In conclusion, it can be added that the results of the methods described in the Chapter 4 on the randomly generated testing set of 10 customers, proof that raster size directly relates to the tour length. Moreover, the analyzed deviation lies under the reasonable threshold which denotes that the raster concept enable to evaluate OD distance matrix without losing accuracy of the real distances.

5 Computational Analysis

This chapter is dedicated to the actual implementation of the methods mentioned in the Chapter 3. For a better understanding of the process, the chapter is starting by presenting predefined parameter. Here is given an overview of the input data, then the characteristics of computer, where all tests as well as computational analysis performed, are provided and after that the thresholds are discussed that enable us to estimate the performance of the proposed concepts on the various cases. The second part presents the computation results on the various cases, which summed up in the tables and graphically illustrated on the boxplots.

5.1. Parameter setting

This thesis is focused on the different distance measurements for OD-matrix problem. The problem is to be solved in urban area of Austria, namely in Vienna and in Lower Austria. The input data were provided by the end supplier of foodstuffs. Name of the company will be keeping a secret. Figure 11 depicts the area under the study. Obviously, clients of the company are not equally distributed over the region. However, it can be noted that they are concentrated in Vienna and spread over the cities in Lower Austria.

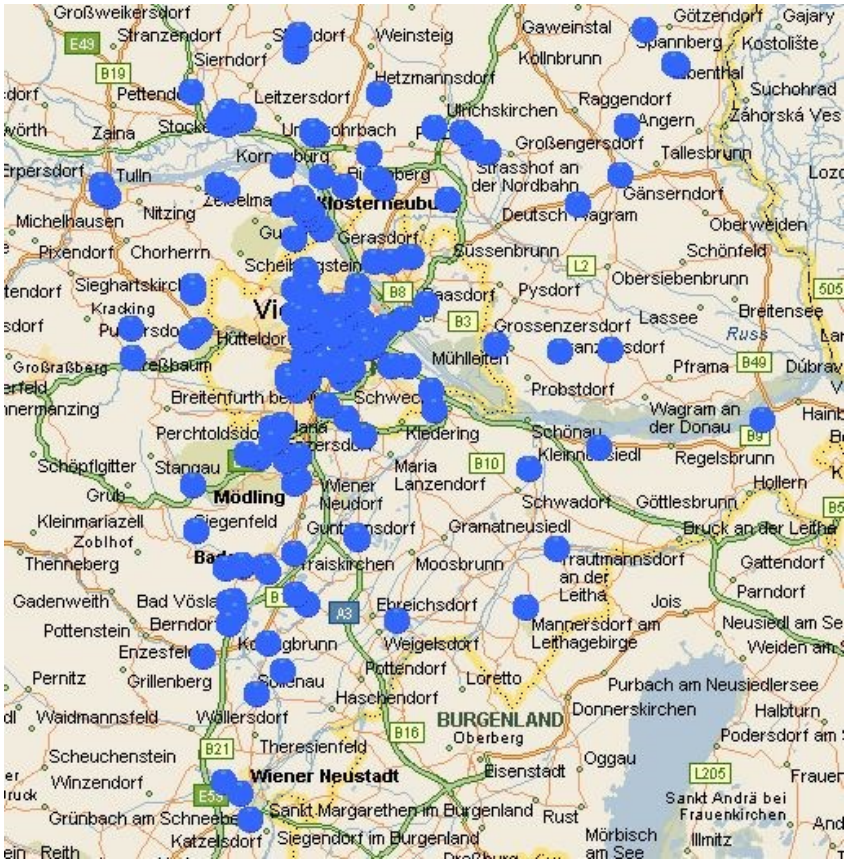

Figure 11. Overview of all customers' locations

The testing was performed on an SONY Vaio® Notebook PC with a 2.67 GHz processor and 4 GB RAM, running on Windows 7. The models were coded in Python 3.5 using the Eclipse as an integrated development environment. For each case we define some thresholds in a way that a threshold is doubled grid size, e.g. for grid size 250m the threshold is 500m. According to those thresholds, the obtained results are analyzed and consequentially the qualities of the concepts are estimated.

5.2. Results analysis

In following subchapter the various cases are generated and developed code is run with case specific modifications. Here is presented five cases numbered in increase order. First four cases (Case 1 – Case 4) contain 100 customers each, however, the customer's positioning vary. The cases from one to three include only clients within Vienna and Case 4 considers also customers in Lower Austria. Last but not least, case number five combines the customers from Vienna and Lower Austria in a ratio of 1:1 that is why the total number of the clients under consideration is 200.

5.2.1. Case 1 – Vienna inner city

As it was already mentioned, the case includes 100 customers within Vienna. The Case 1, shortly C1, considers especially the customers in the city center and in the 3rd district. Fig. 12 shows how the customers distributed in the chosen districts.

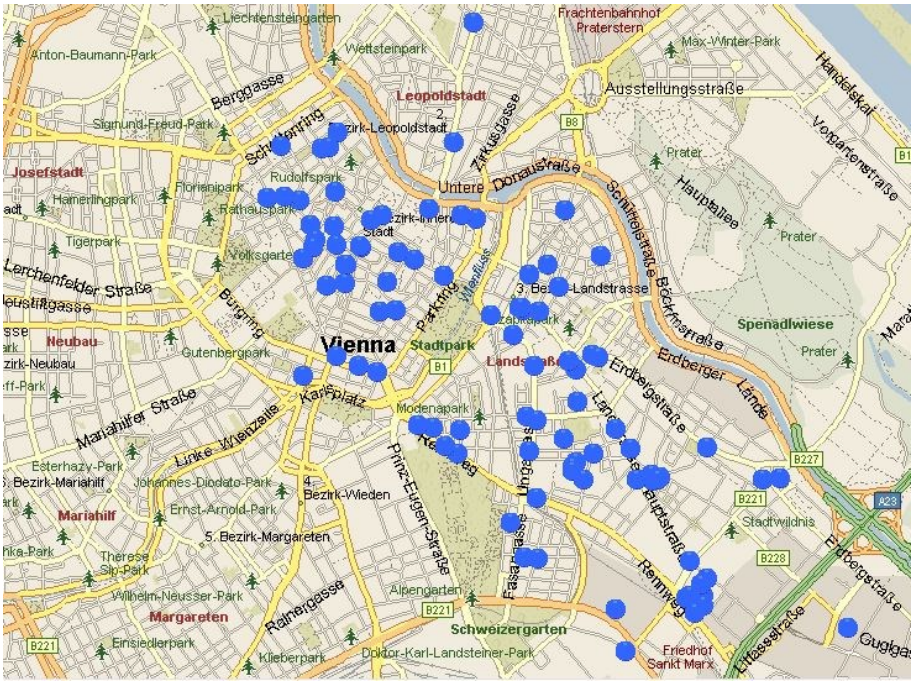

Figure 12. Overview of customer's locations for Case 1

Due to closeness location of the customers, it was decided to examine the Case 1 only on the grid size of 250m. Table 5 summarizes the comparison results for the current case using raster matrix of size 250m.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
C1/250	C-C Euclid	1:10	-4,17	-0,96	-0,55	-0,23	0,49	0,59
C1/250	C-C Manhattan	0:52	-4,09	-0,47	-0,03	0,32	1,73	0,7
C1/250	Raster-Raster	0:52	-4,12	-0,59	-0,21	0,01	1,25	0,57
C1/250	Euclid-Real-Euclid	1:10	-3,86	-0,37	0,0	0,24	1,53	0,58
C1/250	Manhattan-Real-Manhattan	0:52	-3,75	-0,32	0,06	0,3	1,62	0,59

Table 5. Results of the comparison for the Case 1

It is evident that the best performance demonstrates the Euclidean distance method, denoted as C-C Euclid in the Table 5, however, the measured distance is nothing but beeline. In other words, it is simply a linear distance that unfortunately could not be reliable in areas with geographical elements. In the current case there is no geographical elements like mountains or rivers but the under study area is a part of the city with high concentrated transportation network.

The C-C Euclid method determined better results than the shortest path method in 75% of outcomes. In addition, even 100% of all results are under predefined threshold of 500m. Thus, we can conclude that this approach can be implemented in areas with high concentration of the customers and it is very important to note that the under study area should be small and urban one.

The C-C Manhattan method is on the opposite not suitable for this case because its performance the worst one compare to the others listed methods in the Table 5. Therefore, for cases where nodes (customers) locate very close we do not advice to choose Manhattan distance as a distance measure.

The second best outcome from the Table 5 delivers the Raster-Raster method that is more realistic because the distance between the points of the grids is measured as the shortest path. Despite that, the Raster-Raster method is incomplete in the sense of measured distance, i.e. this method estimates simply the length between chose points of the raster matrix and not the whole length between customers.

The last two approaches, the Euclid-Real-Euclid and the Manhattan-Real-Manhattan methods are of most interest because their complex approach to estimate the length is closer to reality. As it was explained in Chapter 3, the route consists of three parts that enable to evaluate the distance more precisely. The numbers from the Table 5 confirm that less than 75% of the total numbers of routes obtained by the Euclid-Real-Euclid or the Manhattan-Real-Manhattan methods lie under the threshold respectively. We briefly reminder that the threshold for the raster size 250m is determined as a doubled value of the size, namely 250m *2 = 500m.

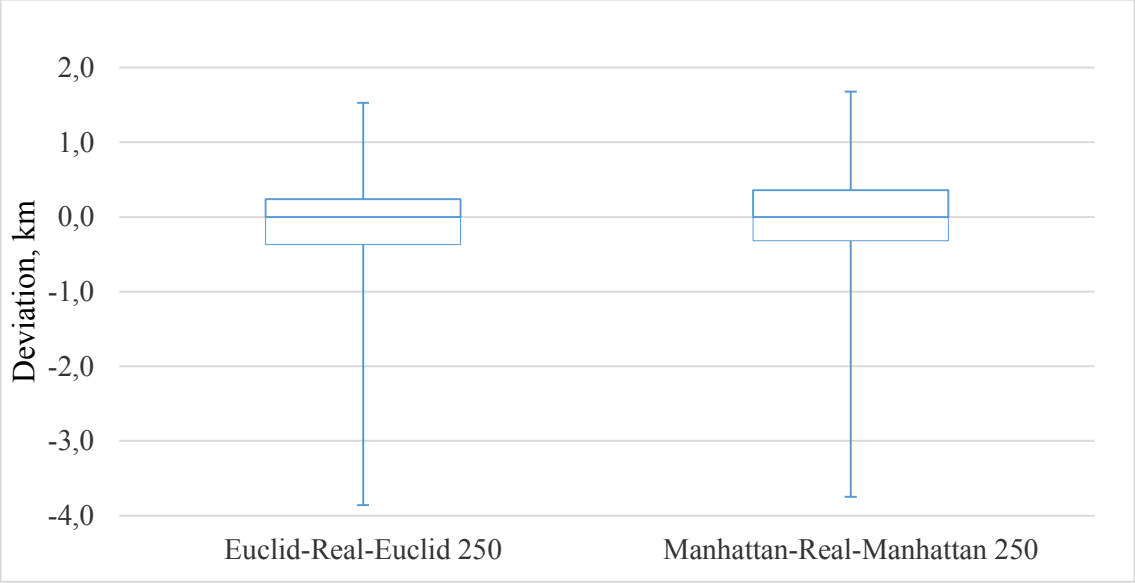

Figure 13. Boxplots of the three-parts-route for Case 1

Figure 13 graphically represents the last two methods from the Table 5. It could be seen that the results derived by the Euclid-Real-Euclid and the Manhattan-Real-Manhattan methods look very similar, they are more or less ident. Such situation can be explained by near positioning of the customers. As it was already said, the Case 1 includes the clients situated in the city center. Thus, it can be concluded, that for the cases where nodes concentrate in one area, either the Euclid-Real-Euclid or the Manhattan-Real-Manhattan methods could be implemented for the evaluating complete distance between end nodes.

5.2.2. Case 2 – Vienna split

The analysis of the Case 2, refers as C2, has absolutely the same structure as the analysis of previous case. At the beginning, Figure 14 depicts the positioning of the customers. Note that second case like the Case 1 also includes 100 customers within Vienna. However, the Case 2 considers the customers in the city center and in the districts on the other side of the Danube, i.e. the set looks as if it is split up two parts. Figure 14 shows the distribution of the customers in the chosen districts.

Figure 14. Overview of customer's locations for Case 2

The Case 2 as well as the Case 1 was tested only on the grid size of 250m because despite the fact that customers are situated on the opposite sides of the Danube, they still lie within Vienna and easily accessible due to developed transportation network. Table 6 sums up the comparison results for the current case using raster matrix of size 250m.

Again, the smallest deviation in length can be observed in the results of the C-C Euclid method. However, the results of the Euclidean distance mislead because in the current case there is geographical elements, namely the Danube, that disturbs the evaluation process and limits the options for connections. The complexity of the case is that that an influence of the geographical object is difficult to evaluate.

The C-C Manhattan method one more time demonstrates that Manhattan distance is not suitable for measuring direct distance between end nodes. The quality of the computational results is low, namely the standard deviation 1,76km is the highest value compare to the C-C Euclid method 1,14km for instance. Therefore, the C-C Manhattan method is also not appropriate approach for the cases with any geographical obstacles.

As it was mentioned in analysis of the Case 1, the Raster-Raster method is incomplete therefore we will not take into consideration the driven results, which will be mostly used for comparison purpose.

Although, the Euclid-Real-Euclid and the Manhattan-Real-Manhattan methods obtained slightly worse results that in the Case 1, the most of routes are within the predefined threshold that is 500m. For instance the standard deviation in the Case 1 is ca. 0,58km and in the Case 2 – 0,71km. Or in the Case 1 deviation of the results in 75% of events is less than 300m and in the Case 2 in less than 75% of events– under 600m. All those allow us to conclude that the proposed methods are robust.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
C2/250	C-C Euclid	1:45	-5,69	-1,9	-0,8	-0,23	1,53	1,14
C2/250	C-C Manhattan	1:38	-4,79	-0,29	0,45	1,94	6,97	1,76
C2/250	Raster-Raster	1:38	-2,89	-0,24	0,02	0,33	5,41	0,71
C2/250	Euclid-Real-Euclid	1:45	-2,47	-0,05	0,21	0,53	5,61	0,71
C2/250	Manhattan-Real-Manhattan	1:38	-2,4	0,0	0,27	0,6	5,64	0,71

Table 6. Results of the comparison for the Case 2

Figure 15 graphically represents the last two methods from the Table 6. Obviously, the Euclid-Real-Euclid and the Manhattan-Real-Manhattan methods deliver slightly differ from each other and significantly deviate from the outcomes of the Case 1. As it reflected on the Fig. 15 the maximum deviation in the Case 2 is around 5,6km and in the Case 1 on the other hand is an only ca. 1,6km. Cause of such a huge gap, which comes in a quarter of the results, could be the positioning of the nodes in the second case.

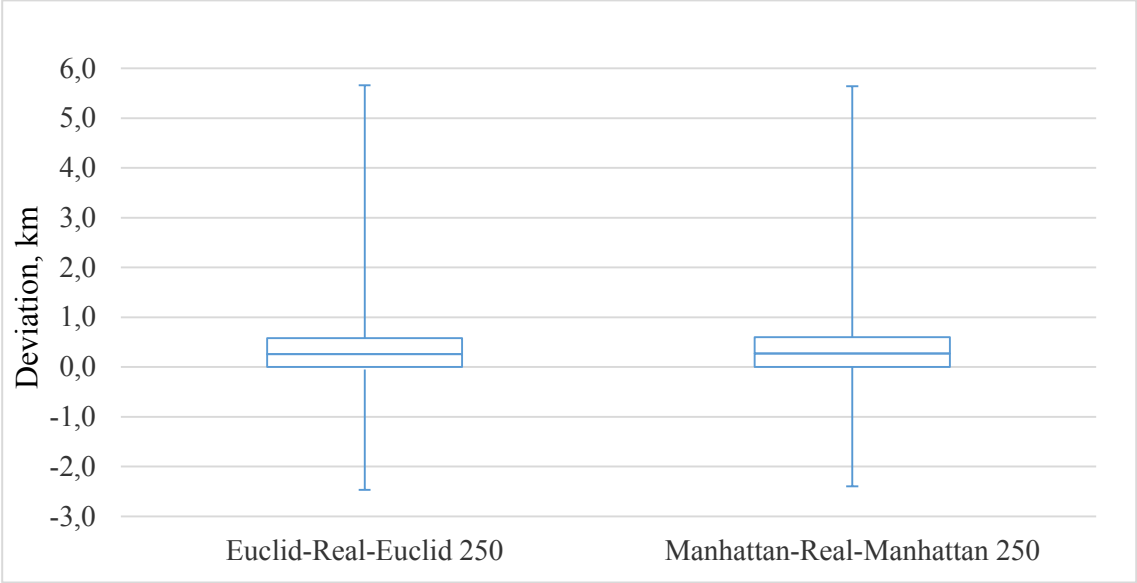

Figure 15. Boxplots of the three-parts-route for Case 2

Despite of all those deviations from the Case 1, the Euclid-Real-Euclid method as well as the Manhattan-Real-Manhattan method demonstrates an appropriate performance in the Case 2. Thus, we come to conclusion that for the cases where data are split to subgroups the both methods reflected on the Figure 15 could be implemented for the estimation OD-matrix.

5.2.3. Case 3 – Vienna total

The Case 3, shortly C3, will be analyzed following the similarly structure of analyses as the both previous case. Note that the Case 3 like the Case 1 and the Case 2 also contains 100 customers within Vienna. However, the Case 3 includes the customers from all districts of Vienna, i.e. on the both sides of the Danube. During the generation procedure, the customers were selected in a way that the Case 3 contains from four to five clients in each district. So the positions are distributed more or less uniform. Figure 16 reflects the customers' positions within Vienna's districts.

Figure 16. Overview of customer's locations for Case 3

It should be noted that the Case 3 as well as the following Cases 4 and 5 will be tested on the various grid sizes, namely 250m, 500m, and 1km due to increase distance between each pair of nodes. Table 7 presents the computational results of the comparison the obtained length with the shortest path for the raster size 250m and Figure 17 shows graphically the deviation of the three-parts-routes evaluated using the Euclid-Real-Euclid or the Manhattan-Real-Manhattan methods respectively.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
C3/250	C-C Euclid	1:53	-6,39	-1,29	-0,58	0,0	3,34	1,08
C3/250	C-C Manhattan	1:43	-6,39	-0,09	0,68	1,64	7,89	1,61
C3/250	Raster-Raster	1:43	-2,91	-0,34	-0,08	0,17	6,07	0,66
C3/250	Euclid-Real-Euclid	1:53	-2,71	-0,13	0,12	0,39	6,31	0,66
C3/250	Manhattan-Real-Manhattan	1:43	-2,64	-0,06	0,2	0,46	6,4	0,66

Table 7. Results of the comparison for the Case 3/250

As like in the previous cases the C-C Euclid method demonstrate the best performance also in the Case 3, however it becomes less reliable, and the customers more distant from each other. It is obvious from the Table 7 that the Euclid-Real-Euclid approach slightly outperforms similar approach based on the Manhattan distance. In whole, the raster concept delivers a reliable outcome, e.g. ¾ of all the results lies under the upper bound of 500m.

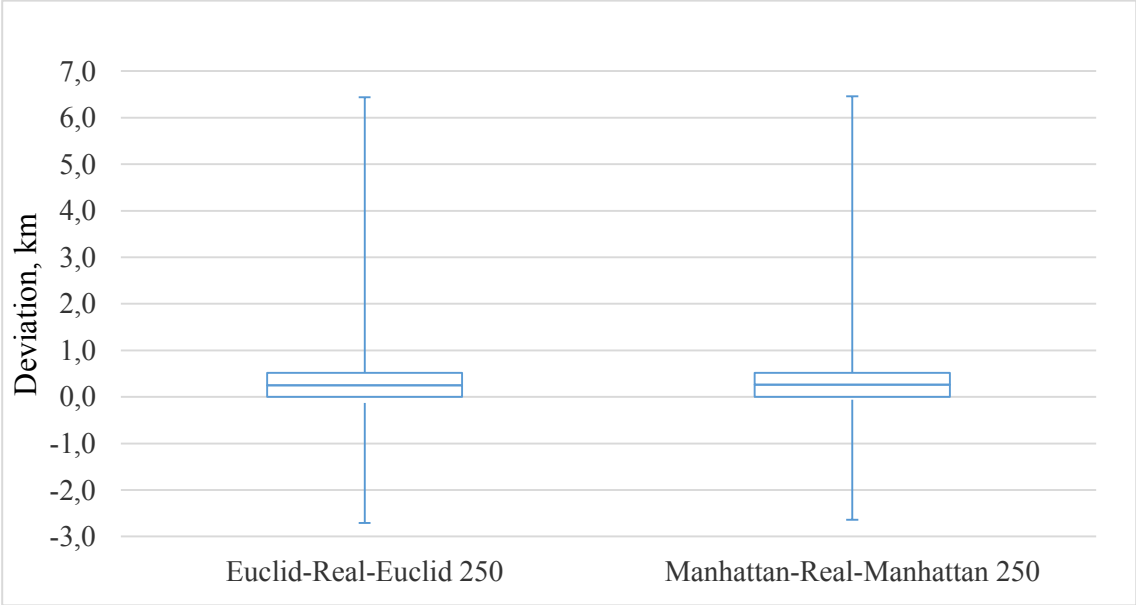

Figure 17. Boxplots of the three-parts-route for Case 3/250

Table 8 presents the computational results for the wider raster size, namely 500m. Here the upper threshold determined as 1000m. We would like to notice that the results from the methods C-C Euclid as well as C-C Manhattan stay the same because those approaches are independent form the proposed raster concept.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
C3/500	C-C Euclid	1:32	-6,39	-1,29	-0,58	0,0	3,34	1,08
C3/500	C-C Manhattan	1:07	-6,39	-0,09	0,68	1,64	7,89	1,61
C3/500	Raster-Raster	1:07	-2,65	-0,43	-0,11	0,18	5,83	0,67
C3/500	Euclid-Real-Euclid	1:32	-2,04	0,0	0,31	0,64	6,26	0,68
C3/500	Manhattan-Real-Manhattan	1:07	-1,82	0,11	0,45	0,81	6,38	0,7

Table 8. Results of the comparison for the Case 3/500

On the contrast the Raster-Raster, the Euclid-Real-Euclid and the Manhattan-Real-Manhattan methods are correlate with the raster size, i.e. incrementing the grid size leads to increase of deviation. Again, we would like to notice that the constructed three-parts-route methods deliver high quality results. For example the standard deviation by the Manhattan-Real-Manhattan method growths from 0,66km to 0,7km.

Figure 18 shows that the differences in performance of the last two methods listed in the Table 8 are clearer. For instance, the level of median is higher by the Manhattan-Real-Manhattan method than by the Euclid-Real-Euclid method. The boxplot represented the Manhattan-Real-Manhattan method shifted a bit up.

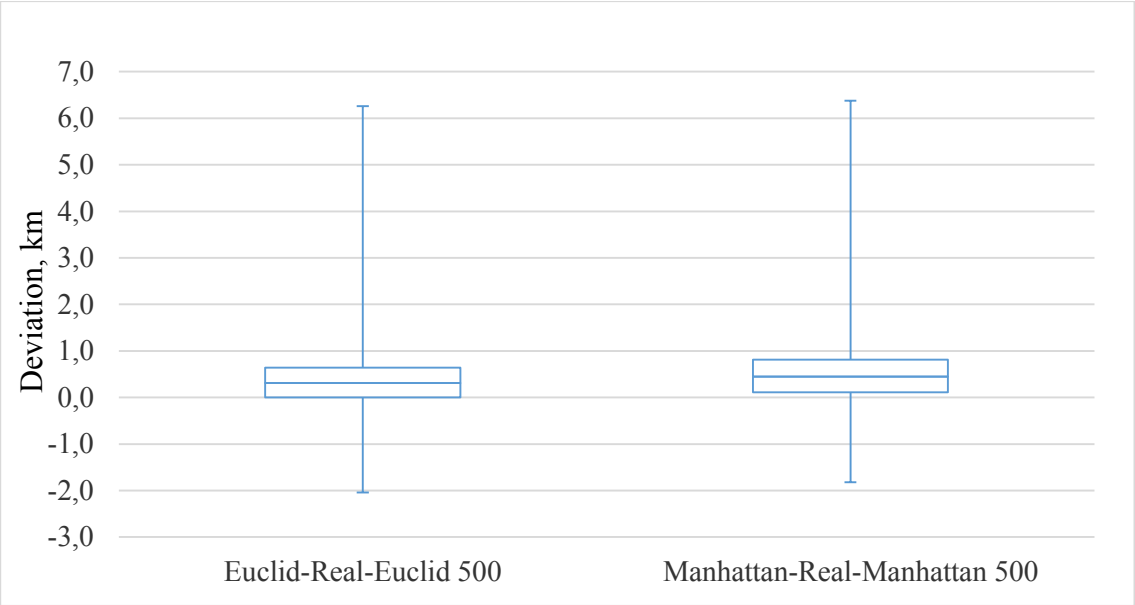

Figure 18. Boxplots of the three-parts-route for Case 3/500

Last test run on the most wide grid size in our study 1 km. Table 9 and Figure 19 present the performance results. Similar observations mentioned by the analysis of the results for the 250m grid size as well as for the 500m raster matrix are valid for the wider grid size. The wider raster distance or grid size, the higher the maximum deviation as well as the standard deviation from the shortest path routes for the Euclid-Real-Euclid and the Manhattan-Real-Manhattan.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
C3/1000	C-C Euclid	1:55	-6,39	-1,29	-0,58	0,0	3,34	1,08
C3/1000	C-C Manhattan	1:36	-6,39	-0,09	0,68	1,64	7,89	1,61
C3/1000	Raster-Raster	1:36	-4,75	-0,56	-0,05	0,42	6,41	0,86
C3/1000	Euclid-Real-Euclid	1:55	-3,44	0,31	0,82	1,38	7,15	0,89
C3/1000	Manhattan-Real-Manhattan	1:36	-2,98	0,53	1,06	1,67	7,45	0,93

Table 9. Results of the comparison for the Case 3/1000

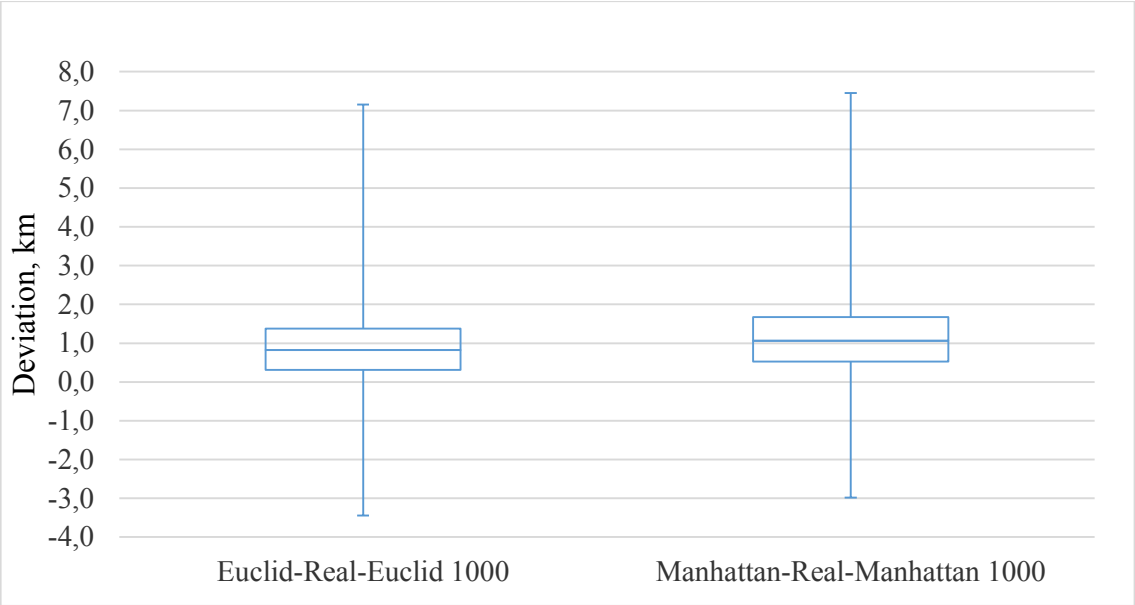

Figure 19. Boxplots of the three-parts-route for Case 3/1000

The difference between the Manhattan-Real-Manhattan method and the Euclid-Real-Euclid method that are of interest in our study is even clearer (see Figure 19).

After the analysis of all performed tests, we come to the conclusion that, firstly, the complex methods we focused on, namely the Manhattan-Real-Manhattan and the Euclid-Real-Euclid, prove their robustness and reliability. Secondly, for the case where vertices situated in one particular area the wide raster matrix is not needed because doubling raster size does not lead to an equivalent increase in the remaining figures such as a value of standard deviation. For instance if we track the standard deviation for the Case 3 it can be seen that it raises form 0,66km to 0,68km and finally rich 0,89km when the grid size doubled from 250m to 500m and then from 500m to 1000m.

5.2.4. Case 4 – Lower Austria

The test runs for the Case 4 will follow absolutely the same structure that the analysis of the Case 3. Note that Case 4 includes also 100 customers and will be performed using three different raster sizes. In the tables provided computational results Case 4 refers as C4 and after slash placed the raster size.

Case 4 is another case and differs from the abovementioned benchmarks in a way that it considers the customers in Lower Austria and not within Vienna. As it depicted on the Figure 20, the customers spread over the region of Lower Austria. As it could be seen that the distances between customers are larger that can influence the results of deviation.

Figure 20. Overview of customer's locations for Case 4

Table 10 demonstrates computational results for the Case 4 with the smallest raster size. The figures are varying unexpected. For instance, the standard deviation is four times higher as in the case where the customers located within urban area, e.g. 0,66km for the Euclid-Real-Euclid method in the Case 3 versus 4,96km for the equivalent method in the Case 4.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
C4/250	C-C Euclid	1:51	-34,48	-9,58	-3,82	0,06	13,43	6,98
C4/250	C-C Manhattan	1:38	-30,86	-2,89	2,42	8,54	29,18	9,54
C4/250	Raster-Raster	1:38	-11,38	0,86	3,21	6,89	29,41	5,27
C4/250	Euclid-Real-Euclid	1:51	-1,52	1,45	3,88	7,42	29,7	4,96
C4/250	Manhattan-Real-Manhattan	1:38	-1,22	1,51	3,98	7,53	29,79	4,98

Table 10. Results of the comparison for the Case 4/250

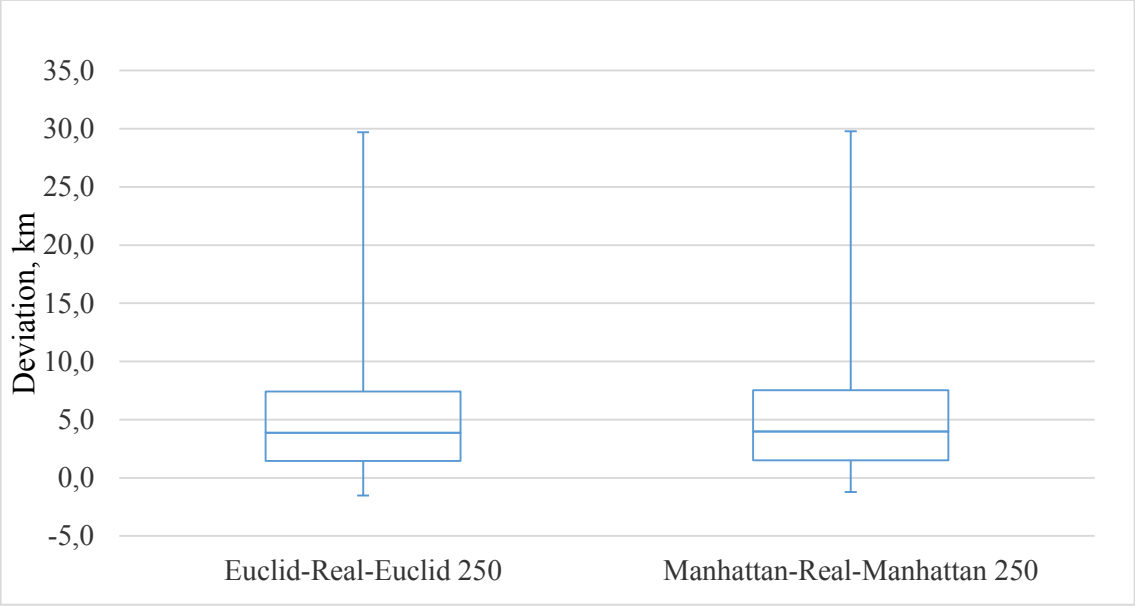

Figure 21. Boxplots of the three-parts-route for Case 4/250

Figure 21 proves the worst performance of the Euclid-Real-Euclid and the Manhattan-Real-Manhattan approaches on the generated Case 4, e.g. the deviation exceed predefined threshold. Even the first quarter of deviation the Raster-Raster, the Euclid-Real-Euclid and the Manhattan-Real-Manhattan methods from the shortest path does not satisfied the upper bound for the 250m raster size that is 500m.

Table 11 shows how the performance of the proposed methods changes with changing grid size. The Euclid-Real-Euclid and the Manhattan-Real-Manhattan have insignificantly changed their performance while customer-customer methods remain the same that because that latter methods are not influenced by the grid data at all. Still the methods that are of the most interest in our study deliver infeasible solution. Figure 22 clearly illustrates the huge deviation of the three-parts-routes from the shortest paths.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
C4/500	C-C Euclid	1:43	-34,48	-9,58	-3,82	0,06	13,43	6,98
C4/500	C-C Manhattan	1:37	-30,86	-2,89	2,42	8,54	29,18	9,54
C4/500	Raster-Raster	1:37	-13,26	0,78	3,21	6,94	29,72	5,29
C4/500	Euclid-Real-Euclid	1:43	-2,57	1,72	4,21	7,79	30,25	5,0
C4/500	Manhattan-Real-Manhattan	1:37	-1,96	1,81	4,28	7,93	31,53	5,03

Table 11. Results of the comparison for the Case 4/500

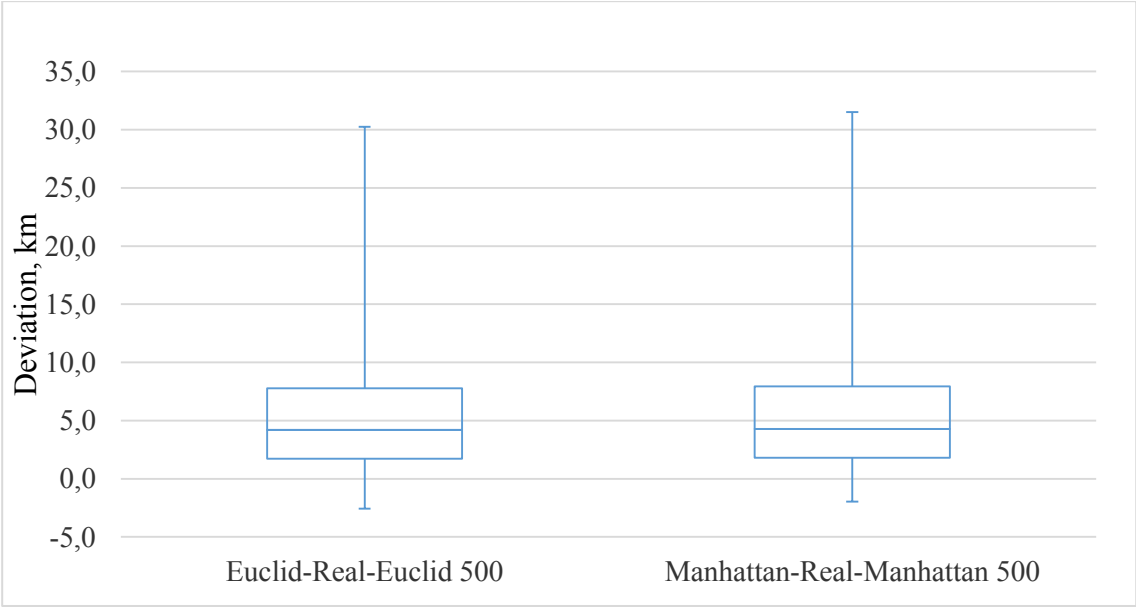

Figure 22. Boxplots of the three-parts-route for Case 4/500

Further increase of the raster size from 500 m to 1km is only slightly changing the results. For instance, the size was doubled when the standard deviation for the Manhattan-Real-Manhattan method is increased from 5,03km to 5,27km. The deviation for the Manhattan-Real-Manhattan method is generally higher that proves also the boxplot that is shifted up in comparison to the boxplot of another method represented on the Figure 23. Thus, the proposed method is inapplicable for the Case 4 due to high amount of unfeasible routes.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
C4/1000	C-C Euclid	1:43	-34,48	-9,58	-3,82	0,06	13,43	6,98
C4/1000	C-C Manhattan	1:52	-30,86	-2,89	2,42	8,54	29,18	9,54
C4/1000	Raster-Raster	1:43	-13,6	0,89	3,36	6,93	28,14	5,21
C4/1000	Euclid-Real-Euclid	1:43	-2,39	2,41	4,98	8,37	28,66	5,02
C4/1000	Manhattan-Real-Manhattan	1:52	-1,41	2,8	5,43	9,31	33,53	5,27

Table 12. Results of the comparison for the Case 4/1000

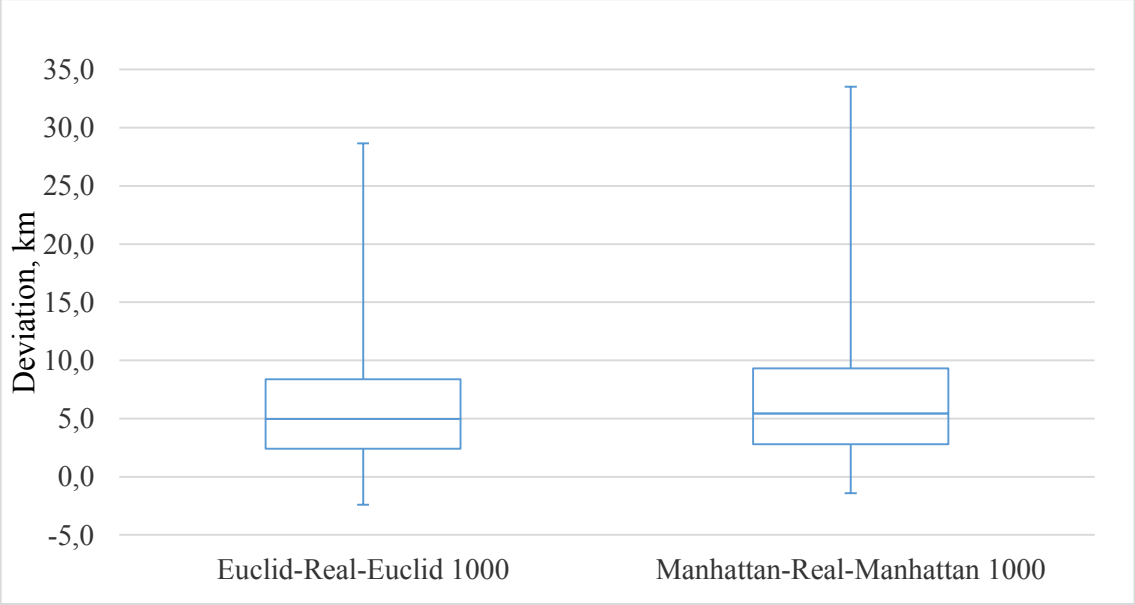

Figure 23. Boxplots of the three-parts-route for Case 4/1000

5.2.5. Case 5 – Vienna and Lower Austria

Finally, we will have a look at the last Case 5 that contains all customers selected for the testing purpose. In comparison with the all cases mentioned above this particular case is a combination of the customers from Vienna and Lower Austria in ratio one to one. Thus, the Case 5, shortly C5, includes doubled amount of the customers that is 200, i.e. the customers' proportion is equal 100 clients placed within Vienna and 100 - within Lower Austria. Figure 24 reflect the location positioning.

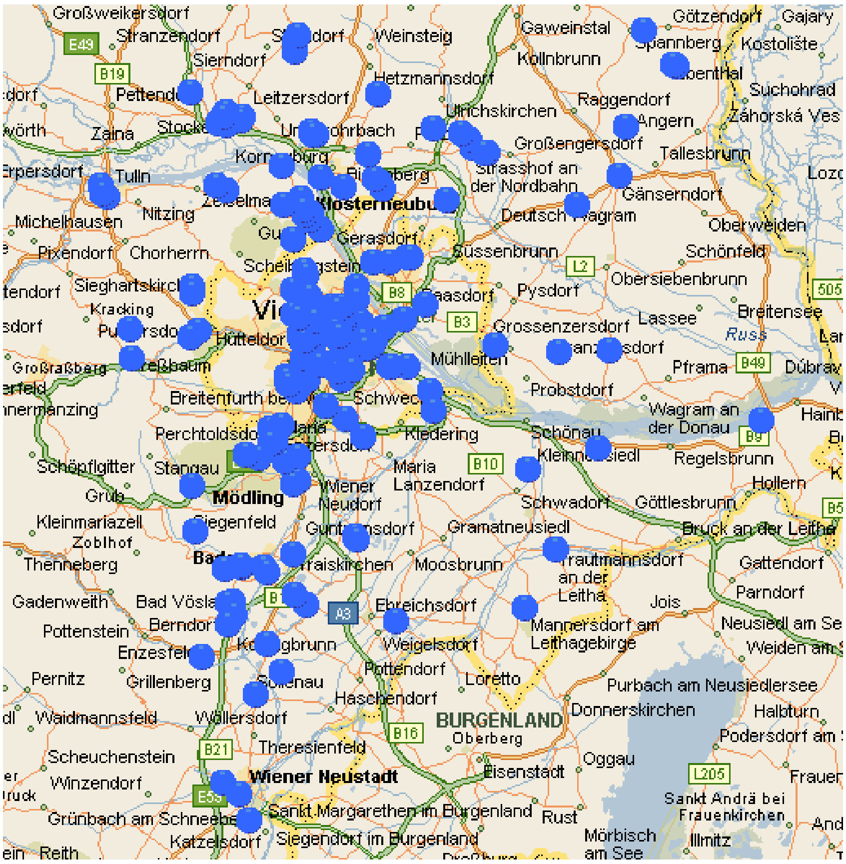

Figure 24. Overview of customer's locations for Case 5

Moreover, we continue to test the provided case on the various raster sizes. The obtained results are of interest for research purpose due to higher complexity of the chosen case. Firstly, the Case 5 is two times bigger than the cases tested before. Secondly, it covers areas with difference transportation structures and population concentration. Thirdly, the distances between customers within the benchmark vary very high.

Case 5 are the mix case so to say, it is includes urban as well as rural areas, geographical elements, areas with high concentration of the customers and part of the region with low concentration. Therefore, computational results are very complex and mixed.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
C5/250	C-C Euclid	2:06	-34,48	-5,17	-1,97	-0,09	13,43	4,73
C5/250	C-C Manhattan	2:02	-30,86	-0,38	1,82	5,76	29,18	6,53
C5/250	Raster-Raster	2:02	-11,38	0,0	0,89	3,2	29,41	3,73
C5/250	Euclid-Real-Euclid	2:06	-2,71	0,28	1,29	3,61	29,7	3,57
C5/250	Manhattan-Real-Manhattan	2:02	-2,64	0,34	1,35	3,7	29,79	3,58

Table 13. Results of the comparison for the Case 5/250

Table 13 presents the results for the smallest raster size. If we look at the standard deviation, the new proposed complex method demonstrates an effective performance. However, Figure 25 obviously depicts very high level of the maximum deviation due to present of the customers from rural areas. Nevertheless, 25% of the results satisfied the predefined threshold compare to the results for the Case 4.

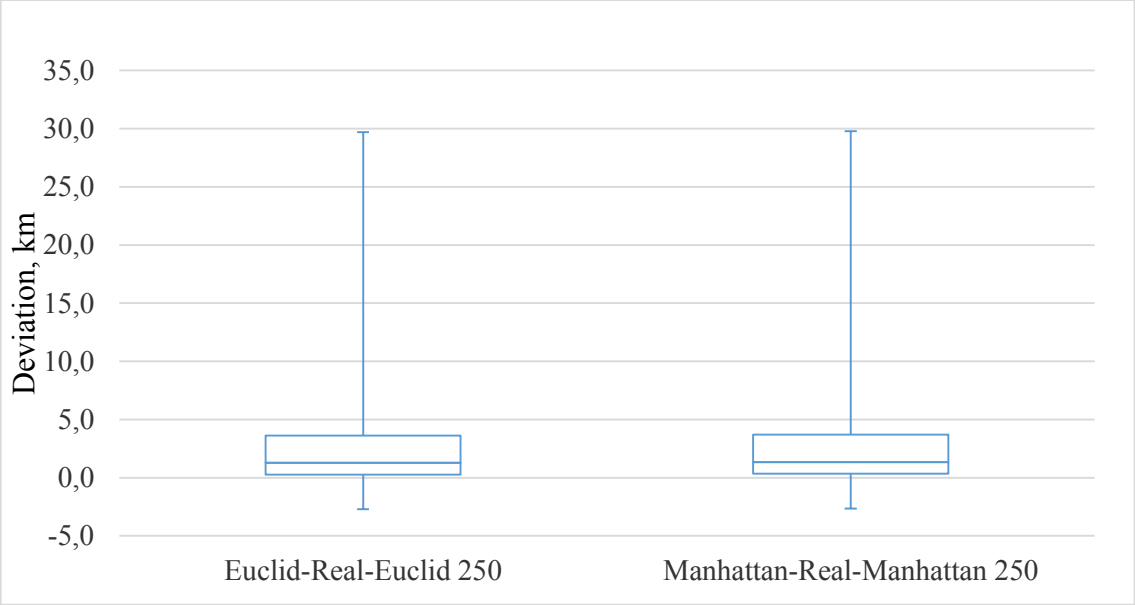

Figure 25. Boxplots of the three-parts-route for Case 5/250

Actually, from the Table 14 and the Figure 26 can be noted that the same trend is kept, deviation increases with spreading raster size. The standard deviation is insignificant increased but the value of the fourth quartile is also increased. Figure 26 proves that statement as well.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
C5/500	C-C Euclid	1:59	-34,48	-5,17	-1,97	-0,09	13,43	4,73
C5/500	C-C Manhattan	2:02	-30,86	-0,38	1,82	5,76	29,18	6,53
C5/500	Raster-Raster	1:59	-13,26	-0,03	0,89	3,2	29,72	3,76
C5/500	Euclid-Real-Euclid	1:59	-2,57	0,49	1,57	3,91	30,25	3,61
C5/500	Manhattan-Real-Manhattan	2:02	-1,96	0,62	1,69	3,99	31,53	3,62

Table 14. Results of the comparison for the Case 5/500

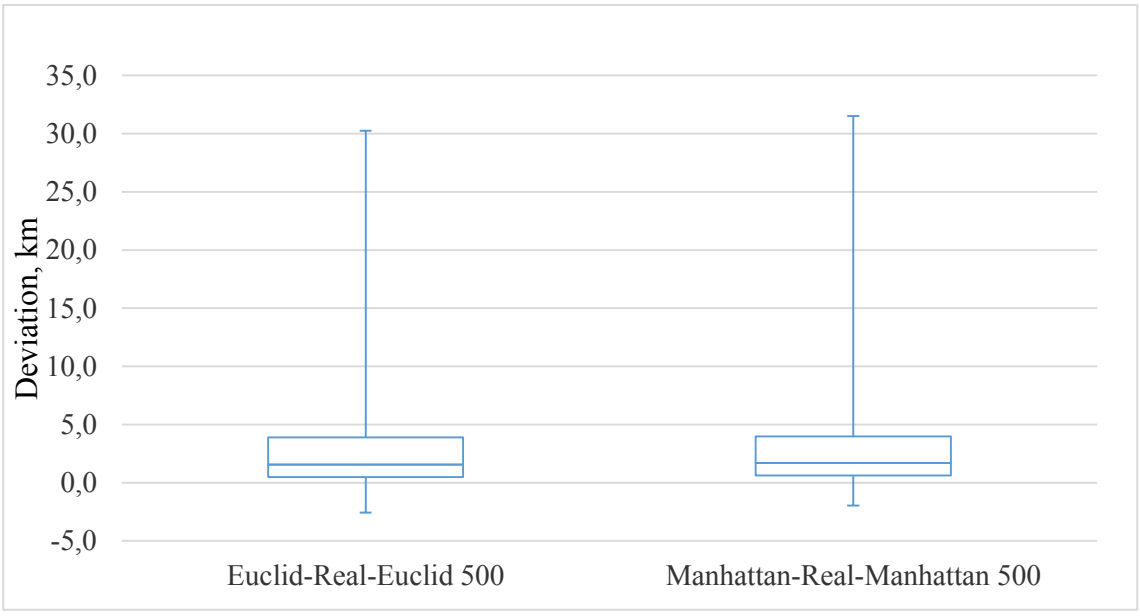

Figure 26. Boxplots of the three-parts-route for Case 5/500

The last test run on the 1000m raster matrix, the results are summarized in the Table 15, and the deviation of the Manhattan-Real-Manhattan method as well as the Euclid-Real-Euclid method is depicted on the Figure 27. It can be seen that the wider raster size brings no improvement in the obtained results.

Case	Method	CPU	Min.	1.Quartile	2.Quartile / Median	3.Quartile	4.Quartile / Max.	SD
C5/1000	C-C Euclid	1:58	-34,48	-5,17	-1,97	-0,09	13,43	4,73
C5/1000	C-C Manhattan	1:55	-30,86	-0,38	1,82	5,76	29,18	6,53
C5/1000	Raster-Raster	1:55	-13,6	-0,0	1,05	3,34	28,14	3,77
C5/1000	Euclid-Real-Euclid	1:58	-3,44	0,99	2,23	4,72	28,66	3,72
C5/1000	Manhattan-Real-Manhattan	1:55	-2,98	1,24	2,57	5,21	33,53	3,95

Table 15. Results of the comparison for the Case 5/1000

Despite of that the C-C Euclid method deliver the results that are under upper bound, we assume that they are mostly infeasible because the beelines are definitely inapplicable in the mixed and complex case like the Case 5. The C-C Manhattan method shows the worst performance in every case from the easiest Case 1 to the complicated Case 5. Therefore, implementation of the Manhattan distance as a distance measure is not advisable for the OD-matrix estimation.

Moreover, the final test with 1km raster matrix shows that the Euclid-Real-Euclid method and the Manhattan-Real-Manhattan method are inapplicable for such kind of mixed cases because only one quarter of routes obtained by those methods lies under the threshold of 2 km. Nevertheless, it should be noticed that the performance of the mentioned methods is efficient in terms of computational time, e.g. Case 5 needs ca.2min for the processing 200 customers on the smallest grid size, and the wider the raster size the less time needed for the processing.

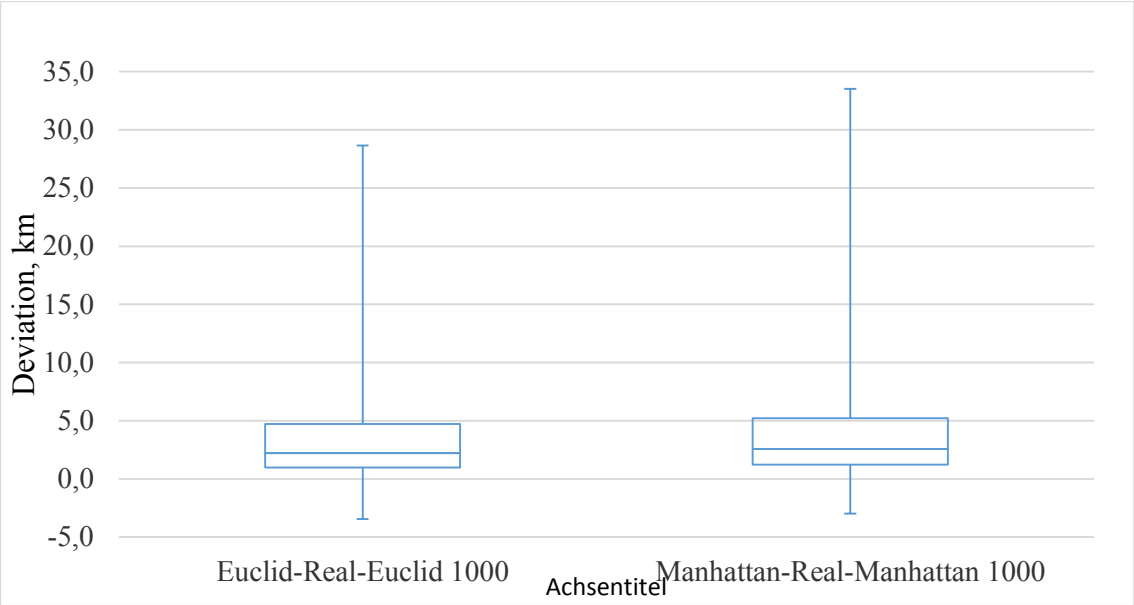

Figure 27. Boxplots of the three-parts-route for Case 5/1000

Thus, we come to the conclusion that proposed raster concept and the methods based on that concept demonstrate an appropriate performance on the cases where the urban areas are under study like in the Case 1, Case 2 and Case 3. In addition, a Euclidean distance is the proper distance measurement than the Manhattan distance. Also the tests with various grid sizes show that the wider grid size is not required for the urban areas.

6 Conclusion

Origin-destination (OD) matrices are one of the key elements for the traffic management development and organization. In the world with fast technical progress, the requirements for the evaluation time are increased significantly. For new gadgets that uses origin-destination matrix as an input data, the computational time should take less than milliseconds. Therefore the methods for the fast obtainment of the OD-matrices are discussed by the researchers constantly.

This thesis is focused on the static model of the OD estimation problem. It discusses the problem of different distance measurements methods. One new concept is proposed and examined on the various cases. For the preliminary test one artificial benchmark is generated and for the further examinations several realistic cases are chosen. The cases is designed in such a way that they cover different areas or combine various types of areas in one case like it was done in the Case 5. The results demonstrate that new suggested method based on the grid data is rapid for the OD-matrix estimation and reliable at the same time.

The findings of the computational analysis is that the complex method using raster concept and Euclidean distance approach works in the urban areas very successfully as it can be seen for instance in the Case 1 with the customers positioning in the city center of Vienna. And on the other hand, the method fails in the rural areas or in the cases including rural areas as it was presented in the Case 4 where only rural area of Lower Austria studied and in the Case 5 which combined rural areas of Lower Austria with urban area of Vienna.

Moreover, the geographical elements could also lead to a strong deviation like in the Case 2 where under study area includes the Danube. The implementation results confirm that rural area is not suitable for the developed concept and methods. Thus, the concept is more suitable and robust for the application in the urban areas.

In addition, the grid size of 250m demonstrates the best behavior in comparison to the other sizes. The increase of the grid size does not lead to any improvement except the decrease in processing time. Consequently, the wider raster sizes like the suggested in this thesis 500m as well as 1km are not necessary if the urban areas are considered.

References

About education *What is a quantile* [Online]. Available at http://statistics.about.com/od/Descriptive-Statistics/a/What-Is-A-Quantile.htm (Accessed 25 February 2016).

Abrahamsson, T. (1998) 'Estimation of origin-destination matrices using traffic counts–a literature survey', *IIASA Interim Report IR-98-021/May*, vol. 27.

Ahuja, R. K., Mehlhorn, K., Orlin, J. and Tarjan, R. E. (1990) 'Faster algorithms for the shortest path problem', *Journal of the ACM (JACM)*, vol. 37, no. 2, pp. 213–223.

Antoniou, C., Ben-Akiva, M. E. and Koutsopoulos, H. (2004) 'Incorporating automated vehicle identification data into origin-destination estimation', *Transportation Research Record: Journal of the Transportation Research Board*, no. 1882, pp. 37–44.

Asakura, Y., Hato, E. and Kashiwadani, M. (2000) 'Origin-destination matrices estimation model using automatic vehicle identification data and its application to the Han-Shin expressway network', *Transportation*, vol. 27, no. 4, pp. 419–438.

Ashok, K. and Ben-Akiva, M. E., eds. (1993) *Dynamic origin-destination matrix estimation and prediction for real-time traffic management systems*.

Ashok, K. and Ben-Akiva, M. E. (2000) 'Alternative approaches for real-time estimation and prediction of time-dependent origin–destination flows', *Transportation Science*, vol. 34, no. 1, pp. 21–36.

Balakrishna, R. and Koutsopoulos, H. (2008) 'Incorporating Within-Day Transitions in Simultaneous Offline Estimation of Dynamic Origin-Destination Flows Without Assignment Matrices', *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2085, pp. 31–38.

Barceló, J., Montero, L., Marqués, L. and Carmona, C. (2010) 'Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring', *Transportation Research Record: Journal of the Transportation Research Board*, no. 2175, pp. 19–27.

Bell, M. G. H. (1983) 'The estimation of an origin-destination matrix from traffic counts', *Transportation Science*, vol. 17, no. 2, pp. 198–217.

Bell, M. G. H. (1991) ‘The estimation of origin-destination matrices by constrained generalised least squares’, *Transportation Research Part B: Methodological*, vol. 25, no. 1, pp. 13–22.

Bell, M. G. H. and Iida, Y. (1997) *Transportation network analysis*.

Bera, S. and Rao, K. K. V. (2011) ‘Estimation of origin-destination matrix from traffic counts: the state of the art’, *European Transport / Trasporti Europei*, vol. 49, pp. 3–23.

Bierlaire, M. and Crittin, F. (2004) ‘An efficient algorithm for real-time estimation and prediction of dynamic OD tables’, *Operations Research*, vol. 52, no. 1, pp. 116–127.

Bland, J. M. and Altman, D. G. (1996) ‘Statistics notes: Measurement error’, *BMJ*, vol. 312, no. 7047, p. 1654.

Brucker, P., Qu, R. and Burke, E. (2011) ‘Personnel scheduling: Models and complexity’, *European Journal of Operational Research*, vol. 210, no. 3, pp. 467–473.

Caceres, N., Wideberg, J. P. and Benitez, F. G. (2007) ‘Deriving origin destination data from a mobile phone network’, *Intelligent Transport Systems, IET*, vol. 1, no. 1, pp. 15–26.

Cascetta, E. (1984) ‘Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator’, *Transportation Research Part B: Methodological*, vol. 18, no. 4, pp. 289–299.

Cascetta, E., Inaudi, D. and Marquis, G. (1993) ‘Dynamic estimators of origin-destination matrices using traffic counts’, *Transportation Science*, vol. 27, no. 4, pp. 363–373.

Cascetta, E. and Nguyen, S. (1988) ‘A unified framework for estimating or updating origin/destination matrices from traffic counts’, *Transportation Research Part B: Methodological*, vol. 22, no. 6, pp. 437–455.

Cipriani, E., Florian, M., Mahut, M. and Nigro, M. (2011) ‘A gradient approximation approach for adjusting temporal origin-destination matrices’, *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 2, pp. 270–282.

Codina, E., García, R. and Marín, A. (2006) ‘New algorithmic alternatives for the O–D matrix adjustment problem on traffic networks’, *European Journal of Operational Research*, vol. 175, no. 3, pp. 1484–1500.

Collins Dictionary [Online]. Available at http://www.collinsdictionary.com/ (Accessed 3 March 2016).

Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. (2001) 'Introduction to algorithms second edition', *The Knuth-Morris-Pratt Algorithm*.

CPU Time [Online]. Available at https://en.wikipedia.org/wiki/CPU_time (Accessed 25 February 2016).

Dan, M., Zhicai, J. and Hongfei, J., eds. (2011) *Adaptive-filtering based dynamic OD matrix estimation*.

Deza, M. M. and Deza, E. (2009) *Encyclopedia of distances*, Springer.

Dijkstra, E. W. (1959) 'A note on two problems in connexion with graphs', *Numerische mathematik*, vol. 1, no. 1, pp. 269–271.

Dixon, M. P. and Rilett, L. R. (2002) 'Real-Time OD Estimation Using Automatic Vehicle Identification and Traffic Count Data', *Computer-Aided Civil and Infrastructure Engineering*, vol. 17, no. 1, pp. 7–21.

Dixon, M. P. and Rilett, L. R. (2005) 'Population origin–destination estimation using automatic vehicle identification and volume data', *Journal of Transportation Engineering*, vol. 131, no. 2, pp. 75–82.

Djukic, T., Barceló, J., Bullejos, M., Montero M., L., Cipriani, E., van Lint, H. and Hoogendoorn, S., eds. (2015) *Advanced traffic data for dynamic OD demand estimation: The state of the art and benchmark study*.

Djukic, T., Flötteröd, G., van Lint, H. and Hoogendoorn, S., eds. (2012) *Efficient real time OD matrix estimation based on Principal Component Analysis*, IEEE.

Frederix, R., Viti, F. and Tampère, C., eds. (2011) *A hierarchical approach for dynamic origin-destination matrix estimation on large-scale congested networks*.

Guihaire, V. and Hao, J.-K. (2008) 'Transit network design and scheduling: A global review', *Transportation Research Part A: Policy and Practice*, vol. 42, no. 10, pp. 1251–1273.

Hazelton, M. L. (2000) 'Estimation of origin–destination matrices from link flows on uncongested networks', *Transportation Research Part B: Methodological*, vol. 34, no. 7, pp. 549–566.

Hazelton, M. L. (2003) 'Some comments on origin–destination matrix estimation', *Transportation Research Part A: Policy and Practice*, vol. 37, no. 10, pp. 811–822.

Högberg, P. (1976) ‘Estimation of parameters in models for traffic prediction: a non-linear regression approach’, *Transportation Research*, vol. 10, no. 4, pp. 263–265.

Hongjun, C., Chongfang, L., Ling, W. and Hua, H., eds. (2010) *Regional OD matrix estimation based on the ratio of intersection turning*, IEEE.

Jin, J. (2016) ‘Estimating a Transit Passenger Trip Origin–Destination Matrix Using Simplified Survey Method’, in *Wireless Communications, Networking and Applications,* Springer, pp. 95–103.

Kachroo, P., Ozbay, K. and Narayanan, A., eds. (1997) *Investigating the use of Kalman filtering approaches for dynamic origin-destination trip table estimation*, Proceedings of Southeastcon '97: Engineering the New Century, pp. 138-142.

Kattan, L. and Abdulhai, B. (2006) ‘Noniterative approach to dynamic traffic origin-destination estimation with parallel evolutionary algorithms’, *Transportation Research Record: Journal of the Transportation Research Board*, no. 1964, pp. 201–210.

Kattan, L. and Abdulhai, B. (2010) ‘Comparative analysis of evolutionary, local search, and hybrid approaches to O/D traffic estimation’, *Journal of Transportation Engineering*, vol. 137, no. 1, pp. 46–56.

Kim, H., Baek, S. and Lim, Y. (2001) ‘Origin-destination matrices estimated with a genetic algorithm from link traffic counts’, *Transportation Research Record: Journal of the Transportation Research Board*, no. 1771, pp. 156–163.

Laerd Statistics *Measures of spread, range, quartiles* [Online]. Available at https://statistics.laerd.com/statistical-guides/measures-of-spread-range-quartiles.php (Accessed 25 February 2016).

Laerd Statistics *Measures of spread, standard deviation* [Online]. Available at https://statistics.laerd.com/statistical-guides/measures-of-spread-standard-deviation.php (Accessed 25 February 2016).

Li, B. (2012) ‘Bayesian inference for origin-destination matrices of transport networks using the EM algorithm’, *Technometrics*.

Lo, H. P., Zhang, N. and Lam, W. H. K. (1996) ‘Estimation of an origin-destination matrix with random link choice proportions: a statistical approach’, *Transportation Research Part B: Methodological*, vol. 30, no. 4, pp. 309–324.

Lou, Y. and Yin, Y. (2010) ‘A decomposition scheme for estimating dynamic origin–destination flows on actuation-controlled signalized arterials’, *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 5, pp. 643–655.

Lundgren, J. T. and Peterson, A. (2008) ‘A heuristic for the bilevel origin–destination-matrix estimation problem’, *Transportation Research Part B: Methodological*, vol. 42, no. 4, pp. 339–354.

Maher, M. J. (1983) ‘Inferences on trip matrices from observations on link volumes: a Bayesian statistical approach’, *Transportation Research Part B: Methodological*, vol. 17, no. 6, pp. 435–447.

Maruhashi, K., Shigezumi, J., Yugami, N. and Faloutsos, C., eds. (2012) *EigenSP: A More Accurate Shortest Path Distance Estimation on Large-Scale Networks*.

Maxima and Minima [Online]. Available at https://en.wikipedia.org/wiki/Maxima_and_minima (Accessed 26 February 2016).

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J. and Damas, L. (2016) ‘Time-evolving O-D matrix estimation using high-speed GPS data streams’, *Expert Systems with Applications*, vol. 44, pp. 275–288.

Nair, S., Radhakrishnan, N. and Mathew, S. (2013) ‘Trip matrix estimation from traffic counts using CUBE’, *IJRET: International Journal of Research in Engineering and Technology*, pp. 142–148.

Okutani, I. (1987) ‘The Kalman filtering approaches in some transportation and traffic problems’, *Transportation and traffic theory*.

Oxford Dictionaries [Online]. Available at http://www.oxforddictionaries.com/de/ (Accessed 3 March 2016).

Perrakis, K., Karlis, D., Cools, M., Janssens, D., Vanhoof, K. and Wets, G. (2012) ‘A Bayesian approach for modeling origin–destination matrices’, *Transportation Research Part A: Policy and Practice*, vol. 46, no. 1, pp. 200–212.

Peterson, A. (2007) *The origin-destination matrix estimation problem: Analysis and computations*, Dissertation, Norrköping, Sweden, Deptment of Science and Technology, Linköpings universitet.

Prestifilippo, G. (2003) *Effiziente Entfernungsberechnung durch Graphenreduktion bei Transportplanungen*, Verlag Praxiswissen.

Robert Niles *Standard deviation* [Online]. Available at http://www.robertniles.com/stats/stdev.shtml (Accessed 25 February 2016).

Robillard, P. (1975) 'Estimating the OD matrix from observed link volumes', *Transportation Research*, vol. 9, no. 2, pp. 123–128.

Sherali, H. D., Arora, N. and Hobeika, A. G. (1997) 'Parameter optimization methods for estimating dynamic origin-destination trip-tables', *Transportation Research Part B: Methodological*, vol. 31, no. 2, pp. 141–157.

Sherali, H. D. and Park, T. (2001) 'Estimation of dynamic origin–destination trip tables for a general network', *Transportation Research Part B: Methodological*, vol. 35, no. 3, pp. 217–235.

Siddharth Kalla *Range (Statistics)* [Online]. Available at https://explorable.com/range-in-statistics (Accessed 26 February 2016).

Spall, J. C. (1998) 'Implementation of the simultaneous perturbation algorithm for stochastic optimization', *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 3, pp. 817–823.

Spall, J. C. (1999) 'Stochastic optimization, stochastic approximation and simulated annealing', *Wiley Encyclopedia of Electrical and Electronics Engineering*.

Spall, J. C. (2005) *Introduction to stochastic search and optimization: estimation, simulation, and control*, John Wiley & Sons.

Spiess, H. (1987) 'A maximum likelihood model for estimating origin-destination matrices', *Transportation Research Part B: Methodological*, vol. 21, no. 5, pp. 395–412.

Stathopoulos, A. and Tsekeris, T. (2004) 'Hybrid meta-heuristic algorithm for the simultaneous optimization of the O–D trip matrix estimation', *Computer-Aided Civil and Infrastructure Engineering*, vol. 19, no. 6, pp. 421–435.

Statistic Austria *Geodata* [Online]. Available at http://www.statistik.at/web_en/publications_services/geodata/index.html (Accessed 3 March 2016).

Tamin, O. Z., Hidayat, H. and Indriastuti, A. K., eds. (2003) *The development of maximum-entropy (ME) and bayesian-inference (BI) estimation methods for calibrating transport demand models based on link volume information*.

Tamin, O. Z. and Willumsen, L. G. (1989) 'Transport demand model estimation from traffic counts', *Transportation*, vol. 16, no. 1, pp. 3–26.

Technopedia *CPU-time* [Online]. Available at https://www.techopedia.com/definition/2858/cpu-time (Accessed 25 February 2016).

Todd, R. *Computation Time* [Online]. Available at http://mathworld.wolfram.com/ComputationTime.html (Accessed 25 February 2016).

Toledo, T. and Kolechkina, T. (2013) 'Estimation of dynamic origin–destination matrices using linear assignment matrix approximations', *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 618–626.

Tympakianaki, A., Koutsopoulos, H. N. and Jenelius, E. (2015) 'c-SPSA: Cluster-wise simultaneous perturbation stochastic approximation algorithm and its application to dynamic origin–destination matrix estimation', *Engineering and Applied Sciences Optimization (OPT-i) - Professor Matthew G. Karlaftis Memorial Issue*, vol. 55, pp. 231–245.

van Aerde, M., Hellinga, B., Yu, L. and Rakha, H., eds. (1993) *Vehicle probes as real-time ATMS sources of dynamic OD and travel time data*.

van Aerde, M., Rakha, H. and Paramahamsan, H. (2003) 'Estimation of origin-destination matrices: Relationship between practical and theoretical considerations', *Transportation Research Record: Journal of the Transportation Research Board*, no. 1831, pp. 122–130.

van Zuylen, H. (1978) 'The information minimising method: validity and applicability to transport planning', *New developments in modelling travel demand and urban systems*.

Wikibooks *Quantiles* [Online]. Available at https://en.wikibooks.org/wiki/Statistics/Summary/Quantiles (Accessed 25 February 2016).

Willumsen, L. G. (1978) 'Estimation of an OD Matrix from Traffic Counts–A Review'.

Willumsen, L. G. (1981) 'Simplified transport models based on traffic counts', *Transportation*, vol. 10, no. 3, pp. 257–278.

Wilson, A. G. (1970) *Entropy in urban and regional modelling*, Pion Ltd.

Yang, H., Sasaki, T., Iida, Y. and Asakura, Y. (1992) ‘Estimation of origin-destination matrices from link traffic counts on congested networks’, *Transportation Research Part B: Methodological*, vol. 26, no. 6, pp. 417–434.

Zhou, X. and 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*, vol. 41, no. 8, pp. 823–840.

Abstract

In the field of transportation system management and similar fields of research an estimation of the origin-destination matrices are one of the important task that should be solved promptly and accuracy. Moreover, the OD-matrices are used as input data in a wide variety of applications. This thesis concerns the estimation of the OD-matrices for the static cases where time is neglected.

The goal of this thesis is to examine well-known distance measurements methods and designing new methods for the time-independent OD-matrix. The proposed methods aggregate the data and use new raster concept in order to evaluate the OD distance matrix. By doing this we analyze the dependency between the aggregation level and the accuracy of the results. All the methods are tested on the artificial instance as well as on the real-world cases with different level of complexity. The outcomes demonstrate an efficient performance.

Zusammenfassung

In den Anwendungs- und Forschungsgebieten von Transportmanagement sowie Transportoptimierungssystemen ist die möglichste exakte Abschätzung der Origin-Destination Matrizen eine der wichtigsten Aufgaben, die einerseits schnell und andererseits möglichst exakt gelöst werden sollen. Darüber hinaus werden die Daten der OD-Matrizen als Inputdaten in einer Vielzahl von Anwendungen eingesetzt. Diese Master Arbeit versucht durch Aggregation der Daten eine möglichst exakte Annährung der OD-Matrizen durch einen Rasterdatenansatz zu erzielen und erforscht den Tradeoff zwischen Genauigkeit und Rechenzeit.

Das Ziel dieser Arbeit ist es bekannte Abstandsmessungen und Methoden zu untersuchen und neue Methoden für das zeitunabhängige OD-Matrix Problem zu entwerfen. Die vorgeschlagenen Verfahren aggregieren die Daten und verwenden neue Raster Konzepte, um die OD-Distanzmatrix zu bewerten. Auf diese Weise analysieren wir die Abhängigkeit zwischen der Aggregationsebene und die Genauigkeit der Ergebnisse. Alle Methoden sind auf dem künstlichen Beispiel getestet, sowie auf realen Anwendungsfälle mit unterschiedlichen Komplexitätsgrad angewendet worden.

Curriculum Vitae

Personal Data:

Name:	Sofya Kalinushkina
Date of birth:	06.10.1990
Nationality:	Russian Federation

Education:

2013 – present	Master Degree Programme Business Administration,
	University of Vienna, Vienna, Austria
	Specializations: Supply Chain Management, Transportation Logistics
2007 – 2012	Master Degree Programme International Business Administration
	Samara State Economic University, Samara, Russian Federation
1997 – 2007	School Nr.16, Zhigulevsk, Russian Federation

Working experience:

01/2015 – present	Amex Export-Import GmbH, Vienna, Austria
	Assistant in logistic department
06/2011 – 08/2011	"Techopttorg" GmbH, Samara, Russian Federation
	Assistant in sales department
06/2009 – 08/2009	Bank "Prioritet", Samara, Russian Federation
	Training

Addition skills:

Language skills	Russian (native language)
	German (fluent)
	English (fluent)
	Spanish (basic skills)
IT skills:	Proficient knowledge in C++, Xpress, Python and Adonis
	Intermediate skills in AnyLogic
	Basic knowledge in VBA, SAP

Appendix

	278365	401211	402973	407563	407642	410850	411266	411961	416078	428386
278365	0,00	4,19	1,72	3,23	2,64	3,55	1,48	3,85	3,85	2,21
401211	4,27	0,00	3,61	1,54	5,23	1,34	4,60	3,50	4,46	2,60
402973	1,84	3,79	0,00	2,56	3,74	2,88	2,58	3,88	2,92	1,54
407563	2,82	2,02	2,16	0,00	4,33	1,53	3,29	4,21	2,92	1,17
407642	2,17	4,10	3,67	4,88	0,00	4,94	1,48	1,37	5,55	3,60
410850	3,46	1,72	2,80	0,64	4,97	0,00	3,93	4,85	3,56	1,81
411266	0,96	4,04	2,53	3,99	1,16	4,05	0,00	2,37	4,66	2,71
411961	3,30	3,51	4,04	4,51	2,34	4,46	2,65	0,00	5,92	3,97
416078	4,02	4,76	2,58	3,17	5,61	4,09	4,56	5,60	0,00	3,08
428386	2,00	2,45	1,36	2,06	3,51	1,54	2,47	3,39	3,24	0,00

Table 16. Shortest path distance matrix for the case 10/250

	278365	401211	402973	407563	407642	410850	411266	411961	416078	428386
278365	0,	4,2	2,53	3,74	2,9	4,66	1,48	4,1	4,3	2,58
401211	4,89	0,	4,16	1,91	6,77	1,34	5,29	4,24	4,92	3,17
402973	2,44	4,06	0,	2,56	4,33	3,48	2,85	4,09	3,32	1,55
407563	2,89	4,3	2,16	0,	4,77	1,53	3,29	4,43	2,92	1,17
407642	3,53	4,47	4,7	5,46	0,	5,55	2,83	1,37	6,72	4,75
410850	3,62	3,64	2,89	0,64	5,5	0,	4,02	5,16	3,65	1,9
411266	0,96	4,45	2,78	3,99	1,42	4,91	0,	3,47	4,8	2,83
411961	3,52	3,57	4,05	4,56	3,17	4,65	3,77	0,	6,07	4,1
416078	4,16	5,78	2,58	3,17	6,05	4,09	4,57	5,81	0,	3,27
428386	2,09	3,5	1,36	2,23	3,97	1,54	2,49	3,63	3,38	0,

Table 17. Fastest path distance matrix for the case 10/250