Faculty of Mathematics and Statistics
Polytechnic University of Catalonia (UPC)

Master thesis

Route analysis and modeling between transport zones based on GPS data

Moisés Ortega Collado

Director: Lídia Montero
Department of Statistics and Operations Research

September 30, 2019
Abstract

Most of the traditional traffic models rely on fixed sensors and traditional surveys. However, Global Positioning System (GPS) provides a better way of observing spatial and temporal data, therefore it can be used for investigation travel and routing behaviors. Nonetheless, raw data is not easy to be treated since it must be attached to a given network.

Data collected on the whole month of March 2017, on regular days, contains information about vehicle’s itinerary tracking and its corresponding trips. Trips are relative to Primary Crown of Barcelona network. The aim of this work is to analyze data and by treating it, to be able to define route choice models between origin and destination areas based on discrete models.
# Table of contents

Abstract .................................................................................................................. 2  
1  Introduction........................................................................................................ 4  
  1.1. The structure of the transport model......................................................... 6  
  1.2. Route choice models................................................................................ 7  
  1.3. Objectives.................................................................................................. 8  
2  Data Description................................................................................................ 9  
3  State of the art..................................................................................................12  
  3.1  Map-Matching............................................................................................12  
  3.2  Principal Component Analysis.................................................................12  
  3.3  Trip Distribution Modelling.....................................................................13  
  3.4  Discrete choice models.............................................................................16  
  3.5  The Multinomial Logit Model (MNL)........................................................22  
4  Methodology.....................................................................................................37  
5  Results..............................................................................................................56  
  5.1  Pre-processing............................................................................................57  
  5.2  Origin-Destination pattern stability.........................................................70  
  5.3  Principal Component Analysis.................................................................74  
  5.4  Clustering..................................................................................................78  
  5.5  Routes choice set algorithm.....................................................................86  
6  Conclusions.......................................................................................................96  
7  BIBLIOGRAPHY..............................................................................................98  
Annex...................................................................................................................101
1 Introduction

Mobility and transport activities are a cornerstone in the boost of society. The migration trend from rural to urban areas has existed forever; however, urban growth has accelerated during the referenced period up to a point that many experts have highlighted: the year 2008, when humanity crossed the threshold of having over 50% of the world population living in cities—a growing trend that is forecast to surpass 70% in 2050. These urban sprawl phenomena have generally occurred in an unplanned, anarchic way due to the combined results of various factors: the relative affluence drifting from rural to city populations, changes in life styles and, in particular, progress in individual mobility in the context of motor-driven transportation. This last factor implies a separation between dwelling and working areas, which is made possible by the development of transportation systems, which in turn are accompanied by the well-known consequences that we call traffic congestion. The current situation, of course, has had strong impacts on energy consumption and emissions (currently around 75% of greenhouse gases of anthropogenic origin are produced by cities) and, consequently, on the quality of life. This growing trend toward urbanization has prompted the phenomenon of “megacities”, regional conurbations that result from the growth and expansion of metropolitan areas, from the merging of two or more, or from both. The United Nations predicts that 2030 will see more than 41 megacities of more than 10 million inhabitants. Some of them, such as Tokyo and Jakarta, already have nearly 40 and 30 million, respectively. This phenomenon is having a relevant impact on spatial reorganization and, therefore, on the configuration of transport systems. This is a consequence of the mutual reciprocity between space and transport; space configures transport just as transport shapes geography.

In order to properly and righteously study a transport system, the most common way is throughout transport modelling tools, formalized and set up in what is known in the literature by the four-stage model (Ortúzar and Willumsen, 2011). This procedure is summarized in the structure of
the classic transport model compressed of four stages: data collection, modelling, scenario development and evaluation. The modelling stage comprises four steps: trip generation, distribution trip modelling, modal split and assignment.

Traffic data can be recollected by traditional information sources (such as census, interviews) or current data sources that create massive amount of data such as localization devices, as GPS, mobile phones, Bluetooth, Big Data (Ortúzar and Willumsen, 2011).

Transportation modelling issues has been studied extensively along decades by many transportation researchers (Ortúzar and Willumsen, 2011). The material and data on which commonly those studies are based on, are usually collected for its posterior analysis by a wide range of methods; the most relevant are done by active solicitation, e.g. travel surveys where respondent self-report a set of questions and activities that are ad-hoc demanded by means of paper, web or phone interviews.

Additionally, in other surveys those questionnaires are backed-up carryings GPS, or, in contrast, there are simply surveys in which subjects are barely asked to carry GPS loggers. In this work, however, data has been passively collected i.e. respondents or subjects which data is sourced without beforehand solicitation, therefore data generated is automatically extracted and can be used for investigation’s purposes of diverse nature.

The provider of the data used for all the subsequent analytic methods developed throughout this work has been INRIX. It recollects instant data from different sources of low-latency GPS tracking data (Cerqueria et al.,2018; Montero and Ros-Roca, 2019) facilitating the automatic obtention of mobility information by means of several data mining procedures. Specifically, in this study it has recollected GPS tracking data from the Primary Crown of the Metropolitan Area of Barcelona, where nearly 50% of this data provides from private consumer cars.
1.1. The structure of the transport model

Transport modelling involves 4 different stages: Data collection, modelling, scenario modelling and evaluation.

Likewise, within modelling stage, this could be divided into four steps:

The first one is the trip generation, and its objective is to model the number of trips generated for each origin and attracted by its respective destination for each study zone selected, known as Transport Analysis Zone or simply TAZ.

The second step is the trip distribution modelling, where the mobility pattern is represented in a matrix form, called usually as (OD matrix). In this step, the growth factor and gravity models are the most distinguished methods (for more information, check pg. 182-184 Ortúzar and Willumsen, 2011)

The third, or modal choice, is addressed to model the election of the transportation mode. The most representative models are: synthetic models (aggregated and disaggregated models), discrete choice models and for instance multinomial models, amongst many others.

The last of the steps is the assignment step, modelled as equilibration of supply and demand, from vehicles or passengers to the transport network. In this last step intercedes the kind of models on which this work is based on. A more concise explanation is given in the next section.
1.2. Route choice models

In the scope of the transportation research, it is often needed to describe, explain and predict choices between two or more discrete alternatives, e.g. the choice of path between a finite set of alternatives to take given an origin and a known destination. In discrete modal choice context, as stated in (Ortúzar and Willumsen, 2011), the probability of individuals choosing a given option is a function of their socioeconomic characteristics and the relative attractiveness of the option. Attractiveness of the alternatives are often explained and modelled under the scope of random utility maximization theory (RUM). These models cope with the purpose of describing how characteristics for each feasible alternative and those that define the individuals affects into how appealing the alternatives are, enclosed in the utility concept.

Utility, however, is not easy to define since there is not a general agreement in the measure of benefit for the individuals.

In the concerning case, route’s choice utility is cumbersome to know when extracted uniquely from GPS data. Since no other data available is obtained about individuals apart from their positioning, defining variables are obtained through the driving behaviour and spatial analysis.

Utility is enclosed in the commonality concept, which results from the overlapping parts amongst different routes. This physical overlapping, amongst some other factors, is the cause of correlation between alternatives. Hence, a definition of a specific term is introduced in the utility function accounting to reduce the attractiveness of the route depending on the level of overlap with other routes in the choice set. The higher is the route’s correlation relative to other, the lower the choice probability of the alternative.
1.3. Objectives

The aim of this study is to expand the knowledge of the applicability of GPS data in transportation modelling. In order to do so, travel behaviours must be analysed under the scope of random utility models and its hypothesis.

Random utility models are the theoretical background that models the discrete choice route paradigm. Hence, an analysis and verification of those assumptions on what these models are based such as homogeneity of the sample, or stability, must be analysed and assess whether they are preserved. Techniques for finding outliers and pre-processing data in order to get a truthful sample are applied.

Multinomial Logit Models (MLN) is one of the most used models to model discrete route choice. However, the capacity of this models to predict correctly path based upon the individual behaviour depends on the size of the choice set, it cannot be excessively large.

In order to get a desirable route choice, trips defining path between zones might be aggregated in a fancy way while dealing with correlation between routes, in order to reduce trips to a less extensive set of routes while the essence of the trip is preserved. Hence, a methodology or possible ways to find a method of aggregation of trips in routes is developed based on the commonality concept, (see Section 4.7.6) on which trips with high degree of overlap are united together.
2 Data Description

Founded in 2005, INRIX pioneered the practice of managing traffic by analyzing data not just from road sensors, but also from vehicles. INRIX delivers data concerning automotive and transportation industries such as real-time parking and traffic information and solutions that facilitate the safe testing and deployment of autonomous vehicles.

Spatial aggregation provided a structure that corresponds to each one of the points collected from the GPS receiver from the sample. This basic structure is referred as waypoints, since there are subsets of points that are sequentially connected and related in order to be able to trace the whole path that a given vehicle has performed in a certain instant.

The data used for the computational experiments corresponds to the first crown of the Barcelona metropolitan area, that has the top population concentration in the area with 2,837,000 inhabitants (IDESCAT) with a public transport system that includes 200 bus lines with over 4,000 stops, 15 railways lines, 11 metro lines and 6 tram lines running on almost 30km of rail network (Metropolitan Area of Barcelona).

GPS tools do not identify travel ends or address trip purposes, nevertheless, they can report travel time and coordinates of locations with small latency, therefore recording the start and end time, the routes of trips and the velocity can be computed. INRIX GPS data sources are fleets of commercial vehicles and private cars. While the travel mode is baseline obtained, trips’ aim for private users are not known, and the identification of end travels is partly arbitrary due to the usage of general fixed rules.

Fleets do no use the concept of trips, and only in monitoring traffic conditions are delivery circuits useful. Trip concept does not apply to fleets, and delivery circuits are only convenient for monitoring traffic circumstances. Consequently, those circuits are not adequate in addressing the origin-destination (OD) spatial distribution of trips nor validating OD path route choice models.
INRIX GPS tracking datasets for a month is used. The raw data consists of 846,295 trips (internal for origin/destination in the study area) and 61,925,267 waypoints in .csv files of, respectively, 236 MB (247,534,055 bytes) and 1.78 GB (1,912,468,534 bytes).

XD Segment average length is around 212 m (see Table 1) and the maximum length with 1.6 km, and generally goes behind main streets in the traffic system. The length of the segments, that is given in miles, has been converted to metres. This measure is given by default in each segment. However, it is not known which is the projection, nor the way to be calculated. Then, as each segment contains (x,y) latitude and longitude points, and taking those as the starting and the ending, this measure is calculated throughout haversine formula, approaching the earth geometry as if it would be a perfect sphere. Although this is not true, the distance between points is not significantly great in comparison with the earth, e.g. the median distance of the segments is around 112 m, while the distance from the earth’s center to the centroid of the set of points defining segments is calculated to be 6378.145·10³ m, although it might differ slightly depending if the latitude changes. Consequently, this has been calculated using (2) taking as latitude the mid-point. The mean square error respect to the given measure in Miles is calculated and by observing quantiles of errors, it seems adequate to use this approach.

<table>
<thead>
<tr>
<th>Table 1. Segments distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
</tr>
<tr>
<td>1.25</td>
</tr>
</tbody>
</table>

The formulation used to calculate the distance between two points is the following:

\[
\begin{align*}
\text{longitud} &= \text{lon}_1 - \text{lon}_2 \\ 
\text{latitude} &= \text{lat}_1 - \text{lat}_2 \\
\end{align*}
\]

\[
\begin{align*}
a &= \sin^2 \frac{\Delta\text{lat}}{2} = \cos(\text{lat}_1) \cos(\text{lat}_2) \sin^2 \frac{\Delta\text{lon}}{2} \\
\end{align*}
\]
\[ c = 2\text{atan2} (\sqrt{a}, \sqrt{1-a}) \]

\[ d = Rc \]

Where the radius of the Earth was calculated with the subsequent formula, which depends on the latitude point on the earth surface:

\[ R \varphi = \sqrt{\frac{(a^2 \cos \varphi)^2 + (b^2 + \sin \varphi)^2}{(a \cos \varphi)^2 + (b \cos \varphi)^2}} \]  

(2)

This study targets drivers of internal trips in the Barcelona metropolitan area. Before undertaking the actual data analysis, raw data from flat .csv files had to be filtered, transformed and grouped together. After this preliminary process of restructuring data, considered as basic methods of data science, other more complex actions should be applied, such as splitting, applying and joining data, computing table margins, and casting/merging data (see section 4.1 for an extended discussion). Regardless of the extent use in research on big data, its application in transportation systems, especially in the context of needs and opportunities (Chodrow et al. 2016; Antoniou et al., 2019) remains unknown.
3 State of the art

3.1 Map-Matching

GPS devices provide vehicle’s location determined by latitude and longitude values, which can be converted into two-dimensional surface. Those points, however, need to be attached to a certain network in order to infer vehicle trajectories. Map matching algorithm is design to map the observed vehicle position to a link in a real traffic network. According to (Greenfeld, 2002), giving the purpose of the study, there are different complexity levels creating a different design of the algorithm. Basically, those levels are defined by the assumptions and the necessity of precision of the map-matching. In the concerning case, there is only information about the network structure in form of segments, defined as lines in the space, with an initial and ending point in the two-dimensional space and the corresponding position of the waypoints.

The goal of the map matching algorithm is to map GPS observation corresponding to waypoints, defined as a sequence of points in the space and time, into the most probable segment linked in the network. Jagadessh et al. (2004) demonstrated a one-by-one matching point algorithm that apparently works fine for bulk size datasets.

3.2 Principal Component Analysis

Principal component analysis was introduced in 1901 by Pearson, and developed by Hotelling in the early 30’s (Jolliffe, 2002) with the aim of describing the variation of a set of uncorrelated variables in a multivariate set.

The main idea behind this technique is to reduce data set dimensionality while retaining as much as possible variation in the data set. This dimensionality reduction is attained by transforming it to a new set of variables, called the principal components, PC, which are imposed to be uncorrelated between each other. Principal components are computed as the
solution of an eigenvalue-eigenvector problem for a positive-semidefinite symmetric matrix. Since each eigenvalue is proportional to the sum of squared distances of the points from their multidimensional mean, which is correlated with the eigenvector, variance retained in the principal components is related to the larger eigenvalues.

This technique has found a wide different application in many fields, such image analysis, pattern recognition or time series forecasting. In traffic and transport modeling has also found its application. For instance, since usually O-D matrices are high-dimensional multivariate data structures, a reduction of dimensionally preserving patterns can significantly improve computational behavior without the loss of accuracy as shown in (Djukic et al., 2012).

In the scope of this master thesis, principal component analysis is performed within-cluster individuals to be the most homogeneous as possible. In order to do so, data is projected into a lower dimension subspace and then k-means is applied in the subspace. As proved in K-means Clustering via Principal Component Analysis (Ding and He, 2004), between-cluster distance remains constant while within-cluster distance shrinks. Hence, clusters tend to be more compact and k-means clustering is more effective.

### 3.3 Trip Distribution Modelling

Trip generation models are used to analyze and predict the total amount of trips generated from a specific zone named as origin, or generator, and its respectively trips attracted to each destination, also called attractions.

This information, although indispensable, is seldom enough for modeling purposes. Despite this, a pattern trip is needed i.e. zone distribution, modes of transport and routes taken.

Trip pattern can be represented in several ways (as stated in Ortúzar and Willumsen, 2011). The one chosen in this work has been in a trip ma-
Matrix known by the literature as Origin Destination matrix, or (O-D) matrix, trips made from an Origin to a Destination in a specific time period that are stored in a matrix. Moreover, (O-D) matrices might be disaggregated by several characteristics as mode of transport, or other characteristics such as trespassing zones, segments, and so on. This disaggregation is useful in order to guarantee homogeneity, although it might compromise the generalization of the results.

Formally, trip matrix is essentially a two-dimensional array where both rows and columns represent the zones study areas. In the study case, zone areas are aggregated by District and Quarter. Consider then i-th row as origin, and the j-th destination column. For diagonal cells, any i = j, are named intra-zonal trips. T_{ij} are the number of trips between zone i and j.

The total sum of trips in a certain row is equal to the total number of trips generated from that origin zone. Alternatively, the sum of trips by a certain column yields the total number of trips attracted to that zone.

\[ \sum_i T_{ij} = D_j \]  \hspace{1cm} (3)

\[ \sum_j T_{ij} = O_i \]  \hspace{1cm} (4)

<table>
<thead>
<tr>
<th>Origins</th>
<th>Destinations</th>
<th>1</th>
<th>2</th>
<th>\ldots</th>
<th>I</th>
<th>\sum_j T_{ij}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T_{11}</td>
<td>T_{12}</td>
<td>\ldots</td>
<td>T_{1i}</td>
<td></td>
<td>O_1</td>
</tr>
<tr>
<td>2</td>
<td>T_{21}</td>
<td>T_{22}</td>
<td>\ldots</td>
<td>T_{2i}</td>
<td></td>
<td>O_2</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\ldots</td>
<td>\vdots</td>
<td></td>
<td>\vdots</td>
</tr>
<tr>
<td>j</td>
<td>T_{j1}</td>
<td>T_{j2}</td>
<td>\ldots</td>
<td>T_{ji}</td>
<td></td>
<td>O_j</td>
</tr>
<tr>
<td>\sum_j T_{ij}</td>
<td>D_1</td>
<td>D_2</td>
<td>\ldots</td>
<td>D_i</td>
<td>\sum_i j T_{ij} = T</td>
<td></td>
</tr>
</tbody>
</table>

From this definition it is possible to define this matrix based on different levels of aggregation both in spatial and time dimension. In the spatial dimension, trips might be aggregated by TAZ zones defined. Moreover,
other groups might be defined. For instance, a set of frontiers TAZ zones in order to enlarge zones and hence, aggregate more trips. In this case, it has been taken Municipality and District in order to gather several TAZ zones. In the following plot, OD Matrix aggregated by Municipality and District are shown, with more trips on it is displayed according to TAZ zones. Notice that this aggregation involves several TAZ zones.

Matrix resulting from the sample is constrained to be placed in a certain period of time. In order to predict further periods with data not available, growth factor methods and gravity models might be used to estimate the future trips. In this case, only data available for 1 month has been gathered. Since is not possible to foresee whether there is a clear stability pattern in the next month periods and since these methods are highly sensitive to errors in the base sample, is not reasonable to forecast next periods.

Figure 10. TAZ-zones corresponding to the OD matrix, Municipality-District aggregation, with the major number of trips generated. On the right, the corresponding number of trips generated per each TAZ-Zone.
Figure 11. AZ-zones corresponding to the OD matrix, Municipality-District aggregation, with the major number of trips attracted. On the right, the corresponding number of trips generated per each TAZ-Zone.

3.4 Discrete choice models

Discrete choice models are well-known and largely used by the literature in those cases when individuals have to select an option from a finite set of alternatives. These types of models are commonly used in transport to describe, explain, and predict choices between two or more discrete alternatives. Since a wide variety of transport problems tackles and cope with these sorts of scenarios and, even though there are other variety of models and methods used for this problems, discrete choice models became to be since the early 1980s seriously worked on it and a corner-stone in the transport modelling (e.g. the route selection between an origin and a destination).
Despite other sort of problems as travelling salesman problem (TSP), that tackles solving the question of finding which is the shortest possible route that passes through each city, and returns to the origin, given a list of cities, or nodes, and the distances between each pair of nodes. On this way, TSP gives as a result a set of discrete paths that according to some criteria satisfies the combinatorial optimization conditions.

On the other way around, discrete choice models statistically relate the choice made by each individual to its attributes (e.g. mean speed, origin, income, ...) and the attributes of the alternatives available for choice set (e.g. routes, set of paths, modes of transport).

As a consequence, discrete choice models specify the probability that an individual chooses an option among a finite set of choices. As stated in Ortúzar and Willumsen, 2011) where is postulated that:

“the probability of individuals choosing a given option is a function of their socioeconomic characteristics and the relative attractiveness of the option.”

The attractiveness of the options is defined in a theoretical construction to represent what the individuals seek to maximize, known as utility. The alternatives, or options from where individuals must choose, do not produce utility on itself. Rather, these are derived from their characteristics (Lanchaster 1966 Lancaster, K.J. (1966) A new approach to consumer theory. Journal of Political Economy 14, 132–157) and those of the individual.

The impossibility to know all possible factors affecting individual choice decisions together with the errors measured in the sample yields the necessity to use a probabilistic model, where the relative influence or effect of each attribute is given by a coefficient (or parameter) related to the alternatives’ attractiveness.

In general, parameters cannot be calibrated using standard curve-fitting techniques, such as least squares, because dependent variables are unobserved probability, defined between 0 and 1, and the observations are
the individual choices either 0 or 1. The only exception to this fact occurs for homogeneous groups of individuals, or when the behavior of every individual is gathered on several occasions, due to observed frequencies of choice are also variables between 0 and 1 (see pag. 228 Ortúzar and Willumsen, 2011).

Disaggregate demand models (DM) are based on theories of individual behavior. Because of this, disaggregate models are more likely than aggregate to be more stable in time and space.

The estimation for this type of models works with individual data and thus probably increases their efficiency in terms of information usage, because if each single point serves as an observation, less data points are needed. Furthermore, working with individual data allows the usage of all the inherent variability present in the information. In contrast, aggregate model-ling works with observations that combine hundreds of other individual observations.

Another characteristic from DM models is that they are less prone to bias due to correlation between aggregate units. In the application of aggregate information ecological correlation can happen: individual behavior may be masked by non-identified characteristics associated with the areas. An example of this can be explained in a correlation between the number of vehicle trips and income. If we consider just the number of trips, this number will decrease with the income (negative correlation). On the contrary, if we consider the data of the vehicle trips as a household level, this number will increase with income (positive correlation).

The use of disaggregate models implies basic probability concepts because it is a probabilistic model, they generate different alternatives but do not show which one has been selected.

Another useful property of this model is that is possible to have an estimation of the coefficients for the variables included. As opposed to aggregate models, where the cost function is limited and has established parameters, in disaggregate models the utility function permits different numbers and designations of the interpretative variables. This characteris-
tic implies that DM models give a more flexible representation of the variables and also that the coefficients of the variables have a peripheral utility interpretation.

On the theoretical paradigm, random utilities models assume and verify the following statements:

- All individuals belong to the same homogeneous population. Hence, they share seemingly behavior upon the same attributes. Moreover, they act rationally and possess perfect information. i.e. Individuals are constrained to the logic exert by the *Homo oeconomicus* (Ortúzar and Willumsen, 2011).

- Given a certain finite set of available alternatives, collectively exhaustive, \( A = \{A_1, \ldots, A_N\} \), i.e. an alternative must be chosen from the set, mutually exclusive, i.e. only one alternative must be chosen from a set \( X \) of vectors of measured attributes of the individuals and their alternatives: Given an individual, \( q \), with attributes \( x \in X \) a set choice of \( A(q) \in A \) (see Ortúzar and Willumsen, 2011, chapter 8).

As long as the model cannot cope with perfect information, the associated net utility for a given individual is represented by two components; a) a measurable part which is function of the attributes and b) a random part that contains the particularities of each individual and observational errors made by the model set. This is formalised as alternatives \( A_j \in A \) has associated a net utility \( U_{jq} \) for individual \( q \).

\( V_{jq} \), which is a function of the measured attributes \( X \) and measurable.

\( \varepsilon_{jq} \), is the random part which reflect the particularities of each individual, together with any measurement or observational errors.

\[
U_{jq} = V_{jq} + \varepsilon_{jq} \tag{5}
\]

It must be point out that with this structure, two individuals with the same attributes and facing the same choice set may select different options. Furthermore, some individuals may not always select what appears to be the best deterministic alternative, fact that actually occurs.
So, without loss of generality it can be assumed that residuals $\varepsilon$ are random variables with mean 0 and a certain probability distribution to be specified. A popular and simple expression for $V$ is:

$$V_{jq} = \sum_k \theta_{kj} x_{jkq} \quad (6)$$

Where parameters $\theta_{kj}$ are assumed to be constant for all individuals in the homogeneous set but may vary across alternatives. Other possible forms, such as functional forms where (4) transforms it into:

$$V_{jq} = \sum_k \theta_{kj} f_{kj}(x_{jkq}) \quad (7)$$

Where the expression (7) is still linear combination, but the explicit functional form of the $x$ variables is somewhat arbitrary. A transformation such as Box-Tukey, may be a better option.

It is important to emphasize the existence of two points of view in the formulation of the above problem: firstly, that of the individual who calmly weighs all the elements of interest (with no randomness) and selects the most convenient option; secondly, that of the modeller who by observing only some of the above elements needs the residuals $\varepsilon$ to explain what otherwise would amount to non-rational behaviour.

The individual $q$ selects the maximum-utility alternative, that is, the individual chooses $A_i$ if and only if:

$$U_{iq} \geq U_{jq} \quad \forall A_j \in A(q) \quad (8)$$

$$V_{jq} - V_{iq} \geq \varepsilon_{iq} - \varepsilon_{jq} \quad (9)$$

However, as $\varepsilon_{iq} - \varepsilon_{jq}$ it is not possible to determine with certitude if (8) holds. Thus, the probability of choosing $A_j$ is given by:

$$P_{jq} = \Pr\{\varepsilon_{iq} \leq \varepsilon_{jq} + (V_{jq} - V_{iq}) \quad \forall A_i \in A(q)\} \quad (10)$$
And as the joint distribution of the residuals $\varepsilon$ is not known, it is not possible at this stage to derive an analytical expression for the model. What it is known, however, is that the residuals are random variables with a certain distribution which we can denote by $f(\varepsilon = f(\varepsilon_1, \ldots, \varepsilon_N)$. Notice $U, f(U)$, is the same but with different mean (i.e. $V$ rather than 0).

Therefore (6) may be written as:

$$P_{jq} = \int_{R_N} f(\varepsilon) d\varepsilon$$  \hspace{1cm} (11)

Where:

$$R_N = \left\{ \varepsilon_{iq} \leq \varepsilon_{jq} + (V_{jq} - V_{iq}), \forall A_i \in A \right\}$$  \hspace{1cm} (12)

And different model forms may be generated depending on the distribution of the residuals $\varepsilon$.

An important class of random utility models is that are generated by utility functions with independent and identically distributed (IID) residuals are multinomial logit model. In this case $f(\varepsilon)$ can be decomposed into:

$$f(\varepsilon_1, \ldots, \varepsilon_N) = \prod_n g(\varepsilon_n)$$  \hspace{1cm} (13)

where $g(\varepsilon_n)$ is the utility distribution associated with option $A_n$, and

$$P_{jq} = \int_{-\infty}^{\infty} g(\varepsilon_j) d\varepsilon_j \prod_{i \neq j} \int_{-\infty}^{V_j - V_i - \varepsilon_j} g(\varepsilon_i) d\varepsilon_i$$  \hspace{1cm} (14)

the general expression (7) reduces to:

However, this expression might be summarized and expressed as:

$$P_{jq} = \int_{-\infty}^{\infty} g(\varepsilon_j) d\varepsilon_j \prod_{i \neq j} G(\varepsilon_j + V_j - V_i)$$  \hspace{1cm} (15)
with $G$ and $f$ respectively the cumulative and the density of the standard Gumbel distribution (i.e. with position and scale parameters equal to 0 and 1).

Because of all these assumptions previously exposed, it is necessary to ensure that all individuals share the set of alternatives and they face the same behavior upon the attributes represented by the model as it has been explained in section 3.2.

### 3.5 The Multinomial Logit Model (MNL)

As mentioned before, because some determinants of the utility are unobserved, either because of latent variables, or due to errors in the measures, utility model is not a determinant model but probabilistic. Upon this condition, Multinomial Logit model assumes that random residuals in the probability of choosing an alternative are distributed IID Gumbel or also called Weibull. In the next section multinomial logit model’s properties are presented:

#### 3.5.1 Properties of Multinomial Logit Model

Multinomial Logit model relies on the following hypothesis:

- Hypothesis 1: Independence of errors

Independence of errors permits to use univariate distribution and then, probability might be computed throughout one-dimension integral as (12).

- Hypothesis 2: Gumbel distribution

\[
\forall \varepsilon \sim f \varepsilon = \frac{1}{\theta} e^{-\frac{\varepsilon - \mu}{\theta}} e^{-e^{-\frac{\varepsilon - \mu}{\theta}}} \tag{16}
\]
Being $\mu$ is the location parameter and $\theta$ the scale parameter.
The probability density function is given by:

$$P \varepsilon < t = \int_{-\infty}^{t} \frac{1}{\theta} e^{-\frac{\varepsilon - \mu}{\theta}} e^{-e^{-\frac{\varepsilon - \mu}{\theta}}} d\varepsilon = e^{-e^{-\frac{\varepsilon - \mu}{\theta}}}$$ (17)

The first and second moments are:

$$E \varepsilon = \mu + \theta \gamma$$

$$Var \varepsilon = \frac{\pi}{6} \theta^2$$ (18)

Where $\gamma \approx 0.577$ known as Euler-Mascheroni constant. It comes out from the derivation by the method of moments.

The mean of $\varepsilon_j$ is not identified if $V_j$ contains an intercept. We can generality suppose that $\mu_j = 0 \forall j$. Moreover, the overall scale of utility is not identified. Therefore, only $J - 1$ scale parameter may be identified, and a natural choice of normalization is to impose that one of the $\theta_j$ is equal to 1.

- Hypothesis 3: Identically distributed errors

As the location parameter is not identified for any error term, this hypothesis is essentially homoscedasticity hypothesis, which means that the scale parameter of Gumbel distribution is the same for all the alternatives. As one of them has been previously fixed to 1, it can therefore suppose that, without loss of generality $\theta_j = 1, \forall j \ldots J$.

3.5.2 Logit form derivation

The probability that the alternative $k$ is better than one other alternative $j$. With hypothesis 2 and 3, it can be written:
\[ P(\varepsilon_j < V_k - V_j + \varepsilon_k) = e^{-e^{-(V_k-V_j+\varepsilon_k)}} \quad (19) \]

As long as hypothesis 1 holds, (13) the product of the expression derived with (20) yields:

\[ (P_j \mid \varepsilon_j) = \prod_{i \neq j} e^{-e^{-(V_j-V_i+\varepsilon_j)}} \quad (20) \]

\[ P_j = \int_{-\infty}^{\infty} (P_j \mid \varepsilon_j)e^{-\varepsilon_j}e^{-e^{\varepsilon_j}} d\varepsilon_j \quad (21) \]

\[ P_j = \int_{-\infty}^{\infty} \left( \prod_{i \neq j} e^{-e^{-(V_j-V_i+\varepsilon_j)}} \right) e^{-\varepsilon_j}e^{-e^{\varepsilon_j}} d\varepsilon_j \quad (22) \]

And thus rewriting (20) for all the alternatives, considering also \( j \)

\[ P_j = \int_{-\infty}^{\infty} \left( \prod_i e^{-e^{-(V_j-V_i+\varepsilon_j)}} \right) e^{-\varepsilon_j} d\varepsilon_j \quad (23) \]

\[ P_j = \int_{-\infty}^{\infty} e^{-\sum_i e^{-(V_j-V_i+\varepsilon_j)}} e^{-\varepsilon_j} d\varepsilon_j = \]

\[ P_j = \int_{-\infty}^{\infty} e^{-e^{-(V_j-V_i)}} e^{-\varepsilon_j} d\varepsilon_j = \]

And, with the proper replacement of variables and the change of integration limits yields the closed form:

\[ P_j = \left[ -\frac{e^{-e^{-(V_j-V_i)}}}{\sum_i e^{-(V_j-V_i)}} \right]_0^{+\infty} = \frac{\exp(\beta V_{iq})}{\sum_{A_j \in A(q)} \exp(\beta V_{iq})} \quad (24) \]

where the utility functions \( V_{iq} \) often parameters are in linear form and those are normalised to 1 and cannot be estimated separately from \( \theta \).
3.5.3 Model specification and functional form

Once the data sample is collected, pre-processed and fitted for purpose in a certain way, the specification and estimation method to use in the model structure has to be specified. Model specification refers in which form the variables enter in the structure of the utility function.

The structure of the model, i.e. what variables enters and their form just like the utility function form, are a matter of testing and experimentation. Hence, it depends strongly in the context and data available.

The satisfaction of the alternatives is usually defined as a linear combination of variables. Although in some context such as destination choice modelling, linear utility functions are not valid or the most convenient (see, for example, Meyer et al., 1978; Foerster, 1979a; Daly, 1997).

For any alternative j, an expression for the satisfaction might be as the form of:

\[ V_{ij} = \alpha_j + \beta x_{ij} + \gamma_j z_i + \delta_j w_{ij} \]  \hspace{1cm} (25)

Where:
- \( x_{ij} \) alternative variables with a generic coefficient \( \beta \).
- \( z_i \) individual specific variables with its alternative coefficients \( \gamma_j \)
- \( w_{ij} \) represents the alternative specific variables, with an alternative specific coefficient \( \delta_j \).

In this way, variables may be one of two kinds:

- Generic, whenever coefficients in the alternatives in the utility function are identical. i.e. \( \theta_{jk} = \theta_k \)
- Specific, if the assumption of equal coefficient \( \theta_k \) is not sustainable, the kth variable only appears in \( V_j \)
Since satisfaction is ordinal, difference between the satisfaction index of two different alternatives j and k is calculated in order to successfully model the choice for the alternatives given by:

\[ V_{ij} - V_{ik} = (\alpha_j - \alpha_k) + \beta(x_{ij} - x_{ik}) + z_i(\gamma_j - \gamma_k) + (\delta_jw_{ij} - \delta_kw_{ik}) \]  \hspace{1cm} (26)

The previous expression yields to deduce the following:

For any \( j \neq k \), \( (\alpha_j - \alpha_k) \), and alternative specific coefficient must be \( \alpha_j \neq \alpha_k \), as well as \( (\gamma_j - \gamma_k) \) otherwise would disappear. Furthermore, since \( (\gamma_j - \gamma_k) \) are associated to individual specific variables, only a linear combination of them a normalization procedure must be carried out. The simplest and most common is to set one of the variables to 0 and the rest of it to be estimated relatively. So, coefficients for individual specific variables should be alternative specific, otherwise it would disappear.

The problem with individual attributes is that not always is simple to decide in which alternative utility(ies) the variable should be used. In the case of route choice, more concretely by the use of GPS data, only spatial, temporal and pattern behavior is known.

The inconvenient of entering a variable in different ways derives in different estimation results and those hampers the selection of the optimum variable in small group of options. In the absence of sufficient knowledge and theoretical premises, the sole way to choose the right variable might be trial and error.

In the concerning case, from waypoints, could be extracted some other variables, as
3.5.4 Choice set determination

Given the data set available, it must be decided which alternatives are available to each individual in the sample. A suitable trade-off between relevance and model complexity is crucial in the decision-making.

Unless individuals are asked in surveys is not possible to know beforehand what choice set could contribute to the goodness of the explanatory model. In the case of modal choice, the number of alternatives is usually small and the problem should not be severe.

Nevertheless, in route choice model’s identification of the alternatives in the choice set is a decisive and critical matter, since the total number of alternatives depends on the network and can be computationally unaffordable in function of this one. Additionally, a problem arises in to know how to measure the attractiveness of the alternatives. (see section 3.5.7)

Although they are not unique, Ortúzar and Willumsen 2011 (see pg. 270) suggests four different ways to manage or reduce the choice set size.

Consider only subsets chosen in the sample (i.e. although the area study might be an extensive network, only consider those edges visited by the trips in the sample and get rid of those not used.

All individuals have available all alternatives available. Hence, choice probabilities are decided by the model and unrealistic alternatives are low or zero.

In the latter case, too many alternatives may perturb the discriminatory capacity of the model not describing adequately the choices amongst realistic options. In the first one, however, is possible to miss realistic alternatives contained in the sample.

Aggregation across options that are distributed over space. It is customarily done by dividing the area of study into zones that are taken as the relevant spatial alternatives and individuals are homogeneous segments.
Spatial Aggregation Using Continuous Functions expressed in terms of two-dimensional coordinates to represent as geographic distributions of the spatial alternatives as individuals, and the attributes. (Ben-Akiva and Watanatada, 1980).

Another issue that comes up, if the model takes into account choices which actually would be ignored by the individual, those alternatives will be considered as a probability to be occurred even though have no chance of ever being.

Ways to handle this problem include:

Heuristic or deterministic choice-set generation rules which permit to erase some alternatives (e.g. setting some threshold distance or eliminating links which are in the opposite direction in the network).

The collection of choice-set information directly from the sample,

The use of random choice sets, in which first the probability distribution function over all possible choice sets is defined; and after, conditional on a specific choice set, a probability of choice for each alternative is defined (see the discussions by Lerman, 1984 and Richardson, 1982).

In this master thesis, the option taken is to design a heuristic model with capacity to create the choice set. Trips are distributed in the network such that might be considered as a collection of segments. However, those segments are not share with all trips between a given TAZ zone. In particular, it might be cases where segments are only used by 1 trip, or even not used at all. Thus, all segments in the path, or the geographic extension between the origin and destination might not be used as the choice set. The quantity of choices to be generated would be enormous and it would not be easy to characterize. That’s the reason why trips must be grouped into more general sets, called routes, which must preserve the information of trips, while at the same time by representative for several trips. The problem with commonalities and the path choice is addressed in section next sections.
3.5.5 Model structure and variable selection

In general, the utility function as seen in (23), must decide which variables \( x_k \in x \) should be included in the model as choice set. In the vast majority of the models is difficult beforehand to know which variables to be considered. Accordingly, a search process, described below in section 4.7.3, is usually employed to check whether they add explanatory power.

It is necessary to decide what variables should enter in the utility function although there is conviction that a given model is appropriate and that linear-in-the-parameters utility functions show no complications, especially in socio-economic studies.

In accordance with the hypotheses that any trade-off mechanisms involving time and costs are the same for all individuals, the most usual technique used until the mid-1970’s in disaggregate modelling studies was to add the variables as additional linear terms.

Usually, the selection of what variables to use in the utility function is done by testing if those variables add additional explanatory power to the model, and this is connected to model credibility and policy sensitivity in the succeeding context. Sometimes it can happen that an important variable is excluded by a stem’s selection procedure. The direction in this instance has been to override the automatic selection methodology (Gunn and Bates, 1982).

3.5.6 Route choice models

The route choice dilemma tackles the selection of route among an origin-destination pair in a given transportation network. The problem is critical in different contexts, such as GPS navigation systems, transportation planning and in intelligent transport systems (GPS). In different re-
search studies they adopted the assumption of the shortest path algorithm (meaning that travelers use the shortest route among all). Despite its efficiency, it fails in terms of realism towards behavioral context. This lack of accuracy aims towards using more sophisticated models such as route choice models (Frejinger and Bierlaire, 2007).

When predicting individual behavior in a choice situation, route choice models are generally used (Frejinger and Bierlaire, 2007). The two principal difficulties in the route choice context are the delineation of the choice set and the correlation between alternatives (Ben-Akiva and Bierlaire, 2003; Frejinger and Bierlaire, 2007).

The modelling of route choice behaviour is crucial in order to predict traffic conditions, to anticipate traveller’s behaviour under hypothetical situations, in the evaluation of traveller’s perceptions of a route and to obtain information of traveller’s reaction and adjustment to sources of information (Prato, 2009).

There are several obstacles regarding the modelling route choice behaviour, for example the challenge of predicting human behaviour and their perceptions of the route traits, the absence of knowledge about the network composition among travellers and the lack of solid data about travellers’ preferences. Therefore, complications in the collection of data might be the principal reason explaining why just few preference studies in the route choice context exist (Prato et al., 2009).

3.5.7 Path size

Route choice modeling has a problem with correlation among alternative routes in the choice set because of their partial overlap. On one hand, this situation precludes adoption of the simple multinomial logit (MNL) model because of its inability to handle correlation among alternatives. On the other hand, models such as paired combinatorial logit (PCL), cross-nested logit (CNL), logit kernel, and probit can handle correlation to a
certain extent but at the cost of much greater complexity and computational burden.

For practical applications, approximate procedures have been proposed such as C-logit (1) and path size logit (PSL) (2). These models try to capture the correlation among the alternatives by adding a correction term to the utility function of the MNL model to correct the calculated choice probabilities. Although these models are heuristic approximations in handling correlations, they have the advantage of adopting the well-known and simple logit structure and outperforming the MNL model.

Generally, C-logit and PSL present the problem that the adopted correction is a heuristic concept missing a well-founded derivation. Assumptions implicitly used are not known and no satisfactory derivation based on theoretical arguments is presented in the literature, thus raising doubts about the correct specification of the correction terms.

In the general framework of the urban networks, routes between an origin and a given destination show often a high degree of overlap with other routes in the choice set.

Due to this fact, correlations between alternative routes significantly influence the choice probabilities. Thus, different route choice model structures with different degrees of complexity were proposed to account for the correlation (Bovy et al., 2008).

In the before mentioned work, correlation is addressed by adding a specific term in the utility function. This term rates the alternatives’ utility depending on the overlapping levels in the choice set. Which is the expression that addresses the best with correlation due to spatial overlap is a key-point in the model specification. Even not being a general agreement in the literature using a unique way to measure correlation, somehow there is a general agreement that random error variance is akin to the route’s length or size measure. Again, the magnitude or the way on how is measured is not straightforward derived. Number of links, length of the route measured in perhaps, time, distance or any other featuring route characteristics.
In random utility modeling, alternative routes are assumed to have random utility distributions with variances and covariances. It is generally agreed that the random error variance in the case of routes somehow is related to some length or size measure of the route; see Daganzo and Sheffi (4) for a sound motivation. Alternatively, the length or size measure may be the number of constituting links, as the route length in distance or time.

Taking a subset of a network involved a single origin-destination pair, some of the links within routes from origin to destination, might partly overlap. i.e. For all routes, exists some link belonging to more than a single route.

$C_n$ as the set of paths between origin and destination taken by an individual, $n$. Consider a net structure based on common links, which nest corresponds to aggregate alternative where routes are grouped. An aggregation subset is formed by all routes using a certain link $a$. Thus, each route is related to its links. Links used only by a unique route; the aggregation subgroup only contains one route. Conversely, if a link is used by more than one route, the aggregation subgroup consists of the routes attached to that link.

$C_{an}$ is the subset of paths using link $a$, $C_{an} \subseteq C_n$.

Path size is determined by considering the choice of the elemental routes made by the individuals and regarding the size of the aggregation of routes by their common link $a$ that equalize the number $M_{an}$ paths using link $a$.

**Size Factor for a Single Common Link in the Choice set**

Given a single link $a$ and its respectively subset $C_{an}$ of routes with size $M_{an}$ using common link $a$.

It is possible to define size factor of $C_{an}$ conveying an additional attractiveness of each alternative in this subgroup explained because of its size.

The larger a subgroup $C_{an}$ is, the higher the probability that this subgroup and one of its members will be chosen by an individual.
The subset utility $U_{an}$ (sometimes mentioned wrongly link utility in the literature) for individual $n$ is described by the maximum utility of routes $j$ belonging to the subset \((7, 8)\):

$$U_{an} = \max_{j \in C_{an}} (V_{jn} + \varepsilon_{jn}) + \varepsilon_{an} \quad a = 1, \ldots, n$$ \hspace{1cm} (29)

Utility $U_{an}$ alternatively might be formulated as the sum of expectations $V_{an}$ and its random term $\varepsilon_{an}$.

Therefore, average deterministic utility for path using link $a$ might be defined as:

$$M_{an} = \sum_{j \in C_{n}} \delta_{aj}$$ \hspace{1cm} (30)

Being $\delta_{aj}$ the link path incidence matrix, with entries 1 in case link $a$ is in route $j$, and 0 any other case.

As shown in (Bovy et al., 2008) under the following random utility choice assumptions:

- If alternatives do not follow a Gumbel random distribution. The size of subset $C_{an}$ is large
- Utility of the subset of paths using link $a$, have equal $V_{jn}$
- For all $j \in C_{n}$, random terms $\varepsilon_{jn}$ are independent and i.d.d, i.e. All routes $j$, no other overlaps between some other routes.

And, taking into consideration random utility choice theory explained in (4.7.3), utility of an individual with respect to all the routes formed by link $a$ equals to:

$$U_{an} = V_{an} + \frac{1}{\mu_a} \ln M_{an} + \varepsilon_{an}$$ \hspace{1cm} (31)

where $\mu_a$ is the scale parameter of link $a$ equal to variance parameter of the error term distribution $\varepsilon_{an}$. $\frac{1}{\mu_a} \ln M_{an}$ represents the additional utility due to the size factor of the subgroup. Notice that whenever $M_{an} = 1$, i.e. link $a$ is only trespassed by a single route, the whole term vanishes.
Size Factor of an Elemental Alternative

While size factor is a positive correction for the size of an aggregate alternative to each route of the subgroup, to find the correct utility value of the elemental alternatives a corresponding negative correction has to be made. In this case, the subgroup $C_{an}$ if the choice model assumes mutual independence among the routes (such as MNL).

The following expression for an alternative $j$ might be defined as:

$$SC_j = -\frac{1}{\mu_a} \ln M_{an} = -\frac{1}{\mu_a} \ln \sum_{l \in C_n} \delta_{aj} \tag{32}$$

Which express a utility reduction for the route. However, in order to make (26) insensitive to the network’s link definition and ensure links’ additive in route, overlap between routes is preferred to be expressed in terms of length $l_a$ of the common links instead of its cardinality. Additionally, common length is related to the overall route length $L_j$ to standardize all routes in the choice set. As defined by (Cascetta).

$$SC_j = -\frac{1}{\mu_a} \frac{l_a}{L_j} \sum_{l \in C_n} \delta_{aj} \tag{33}$$

Where $\frac{l_a}{L_j}$ measures the level of correlation amongst overlapping routes.

Utility of each elemental route in those cases of only a single common link in the choice set is given by:

$$U_{jn} = V_{jn} - \frac{1}{\mu_a} \frac{l_a}{L_j} \sum_{l \in C_n} \delta_{aj} + \varepsilon_{jn} \tag{34}$$

Size Factor for Multiple Common Links in the Choice Set

Whenever more than a single common link, besides the assumptions exposed previously, two additional assumptions must hold;
- All elemental routes of $C_n$ that share links have the same systematic utility $V_{jn}$.
- All $\varepsilon_{jn}$ terms are identically independent distributed. Hence, $\forall a, \mu_a = \mu$.

For all routes with more than one link in common with other routes, the negative size correction $\frac{1}{\mu_a} \ln M_{an}$ are weighted by the length of the commonality. (Bovy et al., 2008) defined the path size corrected written as:

$$U_{in} = V_{in} - \frac{1}{\mu_a} \sum_{i \in \Gamma_i} \frac{l_a}{L_i} \ln \sum_{i \in C_n} \delta_{ai} + \varepsilon_{in} \quad (35)$$

Because correlations are proportional to the common length, the sum over links is natural. The path size correction factor $PSC_i$ of route $i$, which is different from the traditional path size factor because of its derivation, is established as:

$$PSC_i = \frac{1}{\mu_a} \sum_{i \in \Gamma_i} \frac{l_a}{L_i} \ln \sum_{i \in C_n} \delta_{ai} \quad (36)$$

From equation (*), common links’ length is weighted by the expression $\sum_{i \in C_n} \delta_{ai}$, which the following properties are derived:

- The PSC estimates depends on the number of common links in the route, common links’ length $l_a$, and the numbers $M_a$ of distinct routes using each common link.
- Autonomous routes with any single common links have a maximum PSC value of 0.
- There is no upper bound on $M_a$ therefore there is no lower bound on the PSC value.
- The PSC specification is independent to the network specification.

Utility reduction increases as:

- The number of common links in a route is increased.
- Lengths of the common links are larger.
- The number of overlapping routes with links of the route increases.

Keeping all those assumptions and derivations, the heuristic layout must ensure that are all fulfill, and that after its performance, the possible solution contributes to insight those properties. The heuristic evaluation relies on assessing the commonality by the factor size without any kind of aggregation, and being compared with the resulting from the heuristic.
4 Methodology

In this section, methods and arguments applied during the course of this master thesis are explained. First of all, pre-processing techniques tools applied to the data set sample are defined. Secondly, map-matching techniques are introduced. As it is mentioned below, this procedure is essential in order to commit the route choice since in this step points in the space are assigned to their corresponding link of the network.

Then, trip distribution modeling theory is expounded, since is the basis on what the model arises. Definitions and notations in terms of transport modeling are introduced. Following the argumentation, the general framework for discrete choice models are explained in the overall transport application context as well as inconsistencies, or likely arising problems and limitations of the model.

Afterwards, particularities of the problem tackled are bring to the light and addressed in its particularities.

4.1 Pre-processing data and outlier detection

Collected data is susceptible to present imperfections such as errors, outliers, and missing data. In order to correct these inaccuracies, it is necessary to use different tools, just as specific analytics techniques might need particular conditions for the data set (e.g., only binary variables, centred data, normality, only qualitative variables, etc.). In this situation there are tools to verify those specific requisites or instruments to transform data to achieve those conditions. Processes were data is prepared are complex and time consuming and there is a lack of information in articles
dealing with this topic, basically because the methodology should be created specifically to each application and human interaction is required.

A description of the methodology of a pre-processing method is described in Gibert et al., 2016. This method is developed with basic descriptive statistics and univariate and bivariate visual representations of data. Other multivariate techniques, such as Principal Component Analysis, are effective in the detection of outliers and nonlinearities.

The most complex pre-processing techniques are the ones used for identifying imperfections or confirming if an established assumption is applied in a specific analysis. Those techniques are related with diagnoses on data and therefore to tasks associated with pre-processing, for instance outlier, missing data and error detection, identifying relevance or redundancy, feature weighting, independence assessment, detection of influential observations, and normality/linearity assessment.

During the recodification new variables might be involved, and that step is accomplished through discretization, centring, standardization, normalization and, finally, the creation of new variables through aggregation, feature extraction, building indicators and dealing with multivalued variables.

Meta-data supplied by INRIX was used to establish variables and INRIX data undertook different pre-processing steps. Trip attributes can be seen in Table 2. Regarding internal trips to Barcelona’s Metropolitan Area, 53.9% of total trips corresponded to passengers’ cars, and almost 70% of the trips pertain to Endpoint Type 0, showing that the trip does not start or end at a stop, and therefore meaning that the trip trajectory is not an origin-destination trio. Moreover, this reflects that the sample of OD trips included in INRIX does not correspond to the OD trip pattern for internal trips in the study area.

<table>
<thead>
<tr>
<th>Trip attributes</th>
<th>Definitions</th>
</tr>
</thead>
</table>

Table 2. Trip variable attributes
<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TripID</td>
<td>A trip's unique identifier (string)</td>
</tr>
<tr>
<td>DeviceID</td>
<td>A device's unique identifier (string)</td>
</tr>
<tr>
<td>ProviderID</td>
<td>A provider's unique identifier (string)</td>
</tr>
<tr>
<td>Mode</td>
<td>Not available</td>
</tr>
<tr>
<td>StartDate</td>
<td>The trip's start date and time in UTC, ISO-8601 format, example: &quot;2019-09-01T08:33:35.000Z&quot; (string)</td>
</tr>
<tr>
<td>StartWDay</td>
<td>Values from 1 to 5 (Monday to Friday) (integer)</td>
</tr>
<tr>
<td>EndDate</td>
<td>The trip’s end date and time in UTC, ISO-8601 format, example: &quot;2019-09-29T08:33:35.000Z&quot; (string)</td>
</tr>
<tr>
<td>EndWDay</td>
<td>Values from 1 to 6 (Monday to Saturday) (integer)</td>
</tr>
<tr>
<td>StartLocLat</td>
<td>The latitude coordinates of the trip’s start point in decimal degrees (floating point f6.3)</td>
</tr>
<tr>
<td>StartLocLon</td>
<td>The longitude coordinates of the trip’s start point in decimal degrees (floating point f6.3)</td>
</tr>
<tr>
<td>EndLocLat</td>
<td>The latitude coordinates of the trip’s end point in decimal degrees (floating point f6.3)</td>
</tr>
<tr>
<td>EndLocLon</td>
<td>The decimal degree longitude coordinates of the trip’s end point in decimal degree (floating point f6.3)</td>
</tr>
<tr>
<td>Geospatial Type</td>
<td>Describes the trip’s geospatial intersection with the requested zones. Polythomic factor with 4 levels (EE, EI, IE, II)</td>
</tr>
<tr>
<td>Provider Type</td>
<td>Numerical representing the provider type (Consumer, Fleet, Mobile) – Polythomic factor</td>
</tr>
<tr>
<td>Provider Driving Profile</td>
<td>Numerical representing the provider driving profile. Polythomic factor (Consumer, Taxi, LocalDeliver and Trucks)</td>
</tr>
<tr>
<td>Vehicle Weight Class</td>
<td>Numerical representing the vehicle weight class. Polythomic factor (Cars, Vans and Heavy Trucks)</td>
</tr>
<tr>
<td>Origin Zone Name</td>
<td>The origin zone of the trip, if the trip started in a zone – this must be filled in (linked to VISUM-VML TAZ – Zones) (integer)</td>
</tr>
<tr>
<td>Destination Zone Name</td>
<td>The destination zone of the trip, if the trip ends in a zone – this must be filled in (linked to VISUM-VML TAZ – Zones) (integer)</td>
</tr>
<tr>
<td>Endpoint Type</td>
<td>Indicates if the trip starts and ends at a detected stop</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Tri Mean Speed Kph</td>
<td>Average speed (km/h) – Floating point data (5 decimals)</td>
</tr>
<tr>
<td>Trip Max Speed Kph</td>
<td>Max speed (km/h) – Floating point data (5 decimals)</td>
</tr>
<tr>
<td>Trip Distance Meters</td>
<td>Trip distance (m) – Floating point data (1 decimal)</td>
</tr>
<tr>
<td>Movement Type</td>
<td>1 = Moving Trip, 0 = Non-moving Trip – Binary factor</td>
</tr>
<tr>
<td>tt.min</td>
<td>Travel time (min) – Field added by UPC – Floating point</td>
</tr>
</tbody>
</table>

Trips are composed of GPS coordinates (longitude and latitude and additional fields). According to INRIX, latency in Barcelona’s Metropolitan data is usually lower than 10 sec. Each observation registered includes the next fields: trip identifier, waypoint sequence, date, time, longitude, latitude, segment identifier, ZoneName, DeviceId, RawSpeed, RawSpeed Metric and LinkID.

In order to build the data matrix, target population has been determined for the analysis. The process of filtering aims to select subsamples from the main data matrix in order to eliminate observations from other fields unrelated with the study and also to limit the range of the analysis. The selection of the target population has been done in compliance with the following filters:

- Trip registers pertaining to working days (728,060 out of 846,295 trips).
- Internal trips to the first crown of Barcelona’s metropolitan area (456,751 out of 728,060 trips).
- Trips pertaining to private INRIX data consumers (245,728
- Waypoint data for working trips (6,758,275)
- Removed attributes of trip data: Mode, IsStartHome, IsEndHome, ProbeSourceType, MultipleCorridors, MultipleZones, MovementType,
OriginCbg, DestCbg, GeospatialType, ProviderType, ProviderDrivingProfile and VehicleWeightClass.

OriginZoneName and DestinationZoneName were assigned to the trip working data matrix and ZoneName into the waypoint working data matrix. This was based on projecting onto the Transportation Analysis Zone (TAZ) a shapefile loaded in RStudio with sf package and using coordinate points and waypoint registers. 27,821 trips were eliminated due to unsuccessful matching of origin or destination in the TAZ. Intrazonal trips were also rejected, leading to 185,432 trips whose trajectories account for 13,005,532 waypoints. Imputation of SegmentID and LinkID in waypoint registers was initially supported by the maptools RStudio package and snapPointsToLines() function. XD-Segment shapefile by INRIX and VI-SUM-VML link shapefile were uploaded. After cleaning the data, LinkID imputation has been used because the information still showed a lot of contradictions. PostGIS (Anon 2018), was used to define map-matching of the nearest XD-Segment to a GPS track register (longitude, latitude, time).

Occasionally, a trip may have indistinctively the start-off or the end outside the study area. In consequence, some variables as travel time or distance, which ultimately are present in the model might distort the values, yielding errors and mislead the interpretation of the model.

In case of having trips with origin or destination in external zones, or even trips that in some midpoint might exit the study zone, they are entirely eliminated from the sample.

Some commercial software allows the user to add/subtract terminal costs to facilitate better modeling of these trips; the idea is that by manipulating these intra-zonal costs one would make the gravity model fit better. However, this is not a very good option; it is actually preferable to remove intra-zonal trips from the synthetic modeling process and to forecast those using even simpler approaches. This typically assumes that intra-zonal trips are a fixed proportion of the trip ends calculated by the trip generation models.
Moreover, intra-zonal trips are not normally loaded onto the network as they move from a centroid to itself. This makes it less essential to model them in detail. However, in reality, some of these trips use the modeled network. Even though, this problem is probably of significance only for rather coarse zoning systems.

4.2 Map point-segment matching

Global Position Systems is a satellite-based radio-navigation which along several other purposes, it provides geo-location and time information in a defined space (Mintsis et al., 2004). The signal processed by the receivers suffers from signal processing, propagation and other circumstances yielding an error that makes the point drawn to suffer a dispersion causing difficulties when it comes to match those points in the segments that define the urban structure of roads and streets. However, this step is essential in order to define the choices that determine the utility function for the subjects in the model, as long as these choices represent the set of links in the traffic network that corresponds to a chosen path for any individual.

In consequence, any waypoint corresponding to a trip must be matched with the most likely link in which corresponds to the network. In the dataset available instead of links, which are mostly well connected, the spatial definition of the network is made through XDSegments. These are characterized for being points in the two dimensional space, with latitude and longitude as the waypoints, but defined in a line i.e. with a starting coordinate and ending coordinate. Those are referenced by a unique identification, which is the key to be matched with waypoints in order to allocate them in the network. Furthermore, XDSegments, contain forwarding adjacent segments.

Since these are defined according the Primary Crown urban facilities and corresponds to accessible roads, there are lots of them that intersect. Because of this fact, and due to the proximity between each segment, the map-matching procedure might become cumbersome and mistaken.
Figure 12. Map showing the network segments

In order to do so, a criterion to measure the proximity or the closeness from the waypoint to a certain segment and the corresponding matching accuracy must be defined. There are several algorithms in the literature to tackle this problem although there is not “Swiss-knife” that works better than the other, but rather the adequacy of the performance depends on the problem idiosyncrasy itself. In any case, the algorithm chosen must satisfy both, the accuracy criteria required and previously defined for the concerning problem (define or mentioned this accuracy) and to be computationally feasible.

In general, is not possible to solve the problem by brute force comparing each waypoint with the whole set of links in the network. For instance, in the sample there are more than 6.75 million of waypoints to be compared with 7964 segments which would lead to approximately $5.37 \cdot 10^{10}$ comparisons. Hence, in order to reduce the amount of operations several heuristic algorithms such the heading of the observations can be compared with the direction of the links, which implies that links located in the opposite directions, can be discarded (Tang et al., 2015). Other methods focus on observations close to challenging areas, like intersections and parallel links that can prevent the map matching process to estimate the position correct if the heading and position of the observations are deceptive (Hashe-
mi and Karimi, 2016). All of these heuristics together with the addition of constrained local search based on distances might reduce substantially the number of link candidates to be matched and make the problem affordable from the computation time view.

Although some predefined functions are implemented in R to snap points into lines, those do not show a good performance in its commitment. Hence, a hand-craft function must be defined. This work has not been carried out in this context since it has been proven to be tricky and out of the scope of the main objectives of this master thesis. Alternatively, and since is strictly necessary in order to define path-choices, it has been carried out with the help of QGIS.

Map-matching among XD Segments and waypoints posed great difficulties. There are different paths to prepare a map-matching algorithm with advantages and disadvantages. The situation where the only usable information is the network structure in the form of XD segments and the spots of observations conforms the most complicated version of the map-matching dilemma (Greenfeld, 2002). The use of PostGIS provides admissible results, but nowadays there is no implementation of map-matching algorithms based on this methodology (Jagadeesh et al., 2004; Jagadeesh and Srikanthan, 2016; Montero and Ros-Roca, 2019).

Despite that GPS tracks during one moth accumulate large quantities of waypoints and trips registers, the methodology proposed decreases trip samples and OD choice group of paths are usually very small, for this reason a subgroup of the most observed OD pairs has to be pick out for path size and choice set classification. Fewer than five to ten different paths are estimated in all selected OD pairs.

In the next figure, two random trips are plotted with all waypoints contained in the trip put on top of the segments matched. The segments matched were the larger in distance.
As it may be noticed, for the vast majority of waypoints, they fit quite properly in the underlying network of segments. Nevertheless, trips with low latency in the signal might suffer from higher disturbances and worst fit in the map-matching.

The result of the matching procedure for 6,758,275 waypoints into its corresponding segments turns out in a resulting 2,216,148 waypoints that could not be matched and are left as missing data.

Notice that, within the set of waypoints contained in a given trip, it might be a subset of them matched to the same segment. Initially, this should not be an inconvenience, as in relation with GPS latency, and accordance with the length of the segment and the speed to pass through the given segment, a succession of some points might truly be in the same segment.

Cutting off straight away those succession of segments that are equal, it might have its limitation since certain route information might be lost.
For instance, some information might be obtained through analyzing the number of waypoints assigned to each segment, since in accordance the latency, an approach of the time spent in a section might be approximated. Moreover, some segments can be categorized through the number of waypoints assigned to segments could be an insight about the congestion of the road.

Unfortunately, as shown in Table 5, latency changes in most of the devices set up in the vehicles; hence it is difficult to estimate any measure. Thus, in this work not repeated segments are considered to form the set of or $C_n$.

### 4.3 Sparse Matrices

Observed trip matrices are usually sparse. For example, even being aggregated into OD District Municipality the 612 different TAZ Zones defined for the sample into larger 216 zones as in the sample, turns into over 46656 cells. Considering that in the sample there are about 68246 trips. This yields an average of approximately 1.5 trips per cell. This amount, however, is reduced since some OD pairs are more likely to contain trips than others, for instance, due to peak hours i.e. the peak hour of the sample 3 p.m. with more than 4800 trip in comparison with off-peak 1 a.m. with scarcely 200 trips. Adding spatial distribution, as for instance, residential or household quarters where density in the cities may differ from zone to zone, yields with high probability to find no observations on a particular OD pair.

### 4.4 Treatment of External Zones

Occasionally, a trip may have indistinctively the start-off or the end outside the study area. In consequence, some variables as travel time or distance, which ultimately are present in the model, might distort the values yielding errors and misleading the interpretation of the model.
Trips with origin or destination in external zones, or even with trips that in some midpoint might exit the study zone, are entirely eliminated from the sample.

4.5 Intra-zonal Trips

A similar problem occurs with intra-zonal trips. Given the limitations of any zoning system, the cost values given to centroid connectors are a very crude but necessary approximation to those experienced in reality. The idea of an intra-zonal trip cost is then poorly represented by these centroid connector costs.

Some commercial software allows the user to add/subtract terminal costs to facilitate better modeling of these trips; the idea is that by manipulating these intra-zonal costs one would make the gravity model fit better. However, this is not very good; it is actually preferable to remove intra-zonal trips from the synthetic modeling process and to forecast those using even simpler approaches. This typically assumes that intra-zonal trips are a fixed proportion of the trip ends calculated by the trip generation models.

Moreover, intra-zonal trips are not normally loaded onto the network as they move from a centroid to itself. This makes it less essential to model them in detail. However, in reality, some of these trips use the modeled network. Even though, this problem is probably of significance only for rather coarse zoning systems.

4.6 The Stability of Trip Matrices

Trip matrices’ stability is essential for the goodness of model response. It is well-known by the literature that traffic flow along the network varies day-to-day. Variations flow occurs on similar days, or even same weekday and on the same day of the week over similar weeks. Furthermore, even in samples more extended over time, as those contained in a whole year or any longer period, seasonal variations may take place. The question comes out, therefore, about the possibility to spread of day-to-day
variations at the level of trip matrix cell values so estimations over time
might hold for a period longer than the one data sample was collected.

Aggregation process tends to compensate some of the random variations
at the trip matrix level since it is possible to fill the sparse of the matrix.
On the other hand, with the aggregation it is possible to lose representa-
tiveness by treating the averages of the resulting individuals aggregated.

The results also suggest that one should be cautious when testing how
the value of a scheme or plan changes with variations in the estimated
trip matrix used during assessment. Sensitivity analysis seems a particu-
larly appropriate way to investigate the effects of varying the trip matrix.

4.7 Temporal Stability

Transport modelling purpose is to assist in the formulation and evalua-
tion of transport plans. Model parameters must be ensured that will re-
main stable over a given period between base and the transport model
implementation.

This assumption has been analyzed in different studies, concluding that
it can’t be refused when trips of all forms are grouped together (Kannel
and Heathington 1973; Smith and Cleveland 1976) or in crude-zonal based
models.

In order to assess temporal stability, trips are aggregated by different set
of temporal aggregation. First, counting for type of days, such as, Mon-
day, Tuesday,..., Friday and comparing between them whether the same
distribution for the variables holds. The same process is applied for
different aggregations as, per rank, hours. Since most of them are two-by-
two comparison. After ranks hours, the hours have been compared isolate
e.g. hour 1 a.m with 1 a.m and between consecutive hours.

4.8 Geographic Stability
The examination of temporal stability presents some difficulties because data with similar quality is required at two different points in time for the same area. On the contrary, analyzing geographic stability might be less difficult due to the possibility to require data on different locations (e.g., conducting a shared research project by two different organizations based in different regions).

In any transport model, geographic transferability should play a key role because its presence advocates that repeatable regularities in travel behavior exist and therefore can be reflected by the model. It also will show a major probability that temporal stability exists, and this is essential for any prediction model. Finally, the demand for large-scale transportation surveys between different urban areas might also be reduced.

Travel attributes are not always exchangeable between different locations. For instance, it is clear that the duration of the work trip depends on the context, and therefore it should be a function of the profile and distributions of workplaces, area size and residential areas over space.

Despite this, the transitivity of trip prices should be seen as feasible. If the trips express the needs of individuals in different activities away from home and trip costs are related to groups of people with the same characteristics, it can be predicted that those trip costs will remain steady and geographically interchangeable inside an equal cultural context.

The analysis of the transferability of trip generation models is scarce and its results often inadequate (Caldwell and Demetski, 1980; Daor, 1981). Successful tests had only examined sections of the trips, for instance trips made by car (see Ashley 1978). Transferability of the personal-category trip generation model was successfully reported by Supernak (1979, 1981).

The transferability of route choice model was studied by Rose and Koppelman (1984) and they used local data to adjust modal constants. They concluded that similarities in the context were determinant in model transferability. However, due to the large variability obtained in their re-
sults, they recommend to be cautious as a way to guarantee that the transferred model is appropriate in a new context.

4.9 Routes choice definition

Modeling route choice is used in several areas such as Stochastic User Equilibrium, traffic assignment, and plenty of other transport research fields. Theoretical research often assumes all individuals act homogeneously and they will choose the shortest travel cost, when actually does not necessarily occur in real applications.

Regarding to route choice, Cascetta proposed a modified specification of C-logit model, which overcomes the overlap issue, by introducing the commonality factor in the utility function. However, as shown in Cascetta et al., 1996, the C-logit model requires the explicit enumeration of alternative paths, and the model fits better with a limited of reasonable paths while keeping comparable cost.

There are multiple ways of finding these paths, one is to compute a sufficient large amount of overall shortest path, i.e. path enumeration method, selecting those path that satisfy certain set of constraints and getting rid of the rest.

Path enumeration algorithms are, in general, time consuming and CPU expensive. Authors as Hu, X., and Chiu, Y.C (2015 suggest K-Shortest path algorithm to provide a multiple feasible path options, while additional k-paths in order of increasing travel costs balancing overlap and travel time deviation.

However, in this master thesis, instead of performing this search, a different heuristic has been approached since some assumptions do not hold as the authors propose for link cost computation. For instance, there is
not available data about time in the segments in order to build up alternative routes, nor any measure of capacity or congestion, and for that is difficult to set the network’s cost links.

Under some primitive way, some values could be approximated by averaging the values according the data available; for instance, time spent in a given segment and for a concrete time could be approximated through trip variables however those are dismissed because of lack of data. Moreover, authors make use of a link network i.e. all pair are well connected which is not the case.

Instead, a heuristic based on path size defined in (Bovy et al., 2008) has been implemented:

Being $C_{an}$ set of segments in trip $n$ $\delta_{aj} \in \{0,1\}$ whether there is a coinciding segment from segment $a$ to trip $j$

<table>
<thead>
<tr>
<th>Step 1) Initialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_n := \bigcup_{a \in C_{an}} s_{an}, \forall (n \neq j) \in C_n, S := s_n \bigcup s_j$</td>
</tr>
<tr>
<td>$M_a = \sum_{j \in C_n} \delta_{aj}$</td>
</tr>
<tr>
<td>$PS_o := a \in S \left{ \frac{M_a}{\max} \right}$</td>
</tr>
<tr>
<td>$R_0 := PS_o$</td>
</tr>
<tr>
<td>$RM_o := j \in C_n, \forall n \in t: \bigcup_{i \in P_{S_i}}(s_{in}) \bigcap_{j \in C_n} s_j := \frac{l}{\sum_{j \in P_{S_o}} l_j} \ln \sum_{j \in C_n} \delta_{aj}$</td>
</tr>
<tr>
<td>$PSR_{ij} := \frac{l}{\sum_{j \in P_{S_o}} l_j} \ln \sum_{j \in C_n} \delta_{aj}$ $i \in R_0, j \in C_n$</td>
</tr>
<tr>
<td>$p: \forall n \neq j \in C_n, S_i := s_j \bigcap s_n, a \in S_i$</td>
</tr>
</tbody>
</table>

| Step 2) Stopping Criteria: $\inf\{x \in RM\}: p \leq F x$ |
| $S_j := S \setminus (S \bigcap PS_o)$ |
| $s_a := \sum_{j \in C_n} \sum_{a \in P_j} \delta_{aj} > |PS_o| + 1$ |

\[ \text{if } \exists a \in S: S_a \neq \{\emptyset\} \cup p \leq \frac{l}{\sum_{j \in P_j} l_j} \ln \sum_{j \in C_n} \delta_{aj} \text{ then} \]
\[ PS_i = PS_i \cup s_a; \arg\max \frac{l_a}{\sum_{j \in PS_i} L_j} \ln \sum_{j \in C_n} \delta_{aj} \]

\[ R_i = P \cup PS_i \]

\[ RM_{i+1} := j \in C_n, \forall n \in t: \bigcup_i (s_{in}) \bigcap_{j \in PS_i} s_j := \frac{l_a}{\sum_{j \in PS_i} L_j} \ln \sum_{j \in C_n} \delta_{ej} \]

\[ S_i := S \setminus (S \bigcap PS_i), \forall a \in S_i \sum_{j \in C_n} \delta_{aj} \geq |PS_i| + 1 \]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

Step 3) Matching trips to route set

\[ RM_{ij} \forall j \in C_n \]

In the first step, sets and variables are initialized. Set \( C_n \), contains for each trip in the pair OD all segments, \( s_{in} \), that constitutes the trip \( n \). In relation with this set, each segment distance is stored in \( l_n \), for each segment in \( s_{in} \). The algorithm starts by finding the most common segment in trips, i.e. for all trips, count the time for each segment that appears for different trips and set \( P \) with the given segment. Once initialization is done, a search to find which segments can be included in \( P \) is started. The set \( R_i \) constitutes the routes created and on which trips must be agreed according their commonality measured by route matching threshold \( RM_i \). The search is made in such a way that commonality is preserved.
in a threshold; while at the same time the route is the largest as possible. The way on which commonality is computed, does not consider the order of the segments i.e. any permutation of the same path is considered to share the same commonality.

Segments are added to $PS_{i+1}$, until there is no coinciding segment, with any of $PS_i$ segments such that commonality between all trips with any segments a $PS_i$ has a minimum threshold. The heuristic, especially for the first stages of the algorithm might create permutation of some routes, i.e. different routes contain the same set of segments but with other order. Conceptually, does not contribute, since those permutations could be eliminated before assigning trips to routes.

However, time computation, especially whenever the commonality between segments is high, might be deprecated. Constraints of connectivity might solve this problem ensuring that segment $a$ is added to the route only if, for any trip with a segment coinciding with $a$, denoted as $C_{an}$, the resulting intersection, of $C_{an}$, with a given route, resulting segments are consecutive in both routes and in the trip. This process however, has been supress since, computationally is expensive, and remotely happens that a route created influences in the way on which is created for the assignment of the trip on it.

This logic under which the heuristic operates is to guarantee that, the larger numbers of trips are assigned to each route, i.e. trip reduction into routes is as larger as possible while, at the same time, those routes are the most common as possible to the trips keeping a certain threshold of commonality between trips assigned to the route.

A key point for the matching route to trips procedure is to decide on what to base the segment’s election. On one hand, choosing a candidate’s segment to enter the route that best contributes to the commonality between route and the trip that containing the segment, i.e. the segment with maximum commonality with the route, and thus, the best representing a given trip. Or, on the other hand, to choose the one that appears more times for different trips above the threshold.
Every time none of the segments are finding out to be in commonality with the whole path, within the threshold defined, a new path is created such that incorporates the segments in the set of routes. Henceforth, a new route is initialized if not all of trips are covered by a route. In this case, trips are initialized in pairs of segments, by assessing commonality with all trips none assigned, and choosing those that has more commonality, unless there is not such possibility because any other alternative has been tested and only trips of 1 segment may be add as starting point for a route. Likewise, the next segments’ candidates are constrained to be present in any path formed with the segments added. And the search for new segments starts again as initially.

However, a problem arises in the definition of commonality. Figure out a set of segments defining a route, for which a given trip contains all segments in the route. If a new segment is added to the route, it turns out in a loss of similarity. That’s the reason why a route commonality measure needs to be computed between trips and the routes. Even though, for highly different trips, for instance a long one, with few or slightly none existence of commonality with other trips, the algorithm creates a lot of routes, that barely do not contribute to aggregate to routes the trips left to be matched with a route.
1. Initialization
Find segment with max commonality

Add path to set routes
For any two segments not in $PS_i$, go to step 2

2. Calculate PSC
Compare $PS_i$ with all the segments in $S$ not in $PS_i$

Add path to set routes
For any two segments not in $PS_i$, go to step 2

If any trip does not hold commonality threshold

3. Choose the segment with highest PSC,
Check if above threshold for commonality with trips

If the threshold is hold for all trips

4. Match trips routes
Check if above threshold for commonality with trips

Commonality threshold definition

The threshold setting is likely the most sensitive part of the algorithm since the solution might vary highly according this criterion.

In general, for a tight threshold, high commonality is expected to be found in commonality between trips and their respective routes. The algo-
Algorithm tends to create routes with less variance in number of segments’ composing routes.

However, as the commonality threshold increases, the resulting routes are much more similar to the original ones, and thus fewer trips are grouped to routes. On the contrary, the less restrictive with respect to commonality threshold, more difference is allowed between trips and routes, and thus, more trips are allocated to each of the routes.

In cases where trips are highly different between them, for a given threshold, the number of trips to assign reaches the threshold and thus the number of routes increases. As a matter of fact, the algorithm shows a bad performance in those cases where trips are very different, since adds many routes that do not add contribution to satisfy commonality.

The threshold must be defined according to the trade-off paid by aggregating more trips to a route, and thus, less similar the route produced with the original trip, and to reduce the amount of trips enough in order to be capable of managing for the modelling tool decided to be used in the route choice model.

Although, not a closed-form has been found out to properly set a threshold in function of the desired number of route choice set cardinality, the strategy followed in order to decide the threshold is the following:

For any trip, compute the path size corrected with itself in the following manner: the first segment is subtracted from $C_n$, afterwards the last one and ultimately, if the trip allows it, both. This strategy is followed recursively for all the segments forming the trip. Hence, each trip has its own threshold. The number of times to perform this procedure is highly conditioned by the size of segments for the trip and what it seems reasonable with the characteristics of the trips in relation to segments.

5 Results
In this section all the concluding results are exposed. First of all, results regarding the foundation on which the model is based on are shown. Basically, all the results to ensure that random utility model’s assumptions are hold; pre-processing of data, stability, clustering.

Afterwards, as long as multinominal logit models make use of discrete alternatives and those might be representative, and finite, trips must be gathered into routes. The approach in order to obtain a choice set such that is suitable from the point of view of commonality, as developed by (Bovy et al., 2008), is performed.

5.1 Pre-processing

Outlier detection for trip travel time, trip mean speed and trip distance has been developed to obtained two data matrices for GPS tracking data, and also an analysis of the number of waypoints per trip, the number of trips per device and waypoint latency. This work is focused on OD spatial distribution stability over time ad OD route selection. Short trips are no adequate for the objective of the study, consequently they have been discarded from the data set, as recommended d by other authors (Montini et al. 2017).

Some statistics about trip variables are shown in Table 3 and 4:

- Trip distance: up to 25% of the trips have a total distance of less than 1km and 5% are greater than 18km. Trip distance median is 2,490m.

- Trip Mean Speed (km/h): Median is 16.85 km/h; the 2.5% and 97.5% percentiles are 3.75 and 74.51 km/h, respectively.

- Trip Travel time (min): Median is 9.61min, which is a really short trip duration; 2.5% and 97.5% percentiles are 3.75 and 74.51 min, respectively.
Table 3. Variable descriptors based on moments

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Trip Mean Speed</th>
<th>Trip Max Speed</th>
<th>Distance</th>
<th>Transit Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>3,121</td>
<td>8,043</td>
<td>2,500,319</td>
<td>0,950</td>
</tr>
<tr>
<td>Max.</td>
<td>197,393</td>
<td>199,614</td>
<td>23,995,400</td>
<td>133,533</td>
</tr>
<tr>
<td>Median</td>
<td>19,298</td>
<td>64,700</td>
<td>6,283,388</td>
<td>19,600</td>
</tr>
<tr>
<td>Mean</td>
<td>23,432</td>
<td>70,384</td>
<td>8,108,127</td>
<td>23,729</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>14,280</td>
<td>30,572</td>
<td>5,253,125</td>
<td>15,414</td>
</tr>
<tr>
<td>Skewness</td>
<td>3,159</td>
<td>1,267</td>
<td>1,147</td>
<td>1,850</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>27,830</td>
<td>5,960</td>
<td>3,478</td>
<td>7,932</td>
</tr>
</tbody>
</table>

Table 4. Quantiles of the trip variables

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Trip Mean Speed</th>
<th>Trip Max Speed</th>
<th>Distance</th>
<th>Transit time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>3,121</td>
<td>8,042</td>
<td>2500,319</td>
<td>0,95</td>
</tr>
<tr>
<td>10%</td>
<td>10,868</td>
<td>39,205</td>
<td>3015,317</td>
<td>9,40</td>
</tr>
<tr>
<td>20%</td>
<td>13,475</td>
<td>48,363</td>
<td>3631,356</td>
<td>1,181</td>
</tr>
<tr>
<td>30%</td>
<td>15,342</td>
<td>53,992</td>
<td>4340,978</td>
<td>1,435</td>
</tr>
<tr>
<td>40%</td>
<td>17,297</td>
<td>58,891</td>
<td>5235,669</td>
<td>1,687</td>
</tr>
<tr>
<td>50%</td>
<td>19,298</td>
<td>64,699</td>
<td>6283,388</td>
<td>1,960</td>
</tr>
<tr>
<td>60%</td>
<td>22,309</td>
<td>72,136</td>
<td>7702,088</td>
<td>2,290</td>
</tr>
<tr>
<td>Quantiles</td>
<td>0%</td>
<td>15%</td>
<td>30%</td>
<td>45%</td>
</tr>
<tr>
<td>-----------</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>N° Waypoints / Trip</td>
<td>2</td>
<td>6</td>
<td>11</td>
<td>24</td>
</tr>
<tr>
<td>Latency</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

The ultimate working test to get OD matrices at distinct spatial aggregations has 68,246 trips and 6,758,275 waypoints. Just 10,505 trips correspond to OD pairs that have more than 1 trip. Scarcity is normal in OD matrices, but if the TAZ spatial resolution is discounted, and it is taken into account a district in municipality resolution, then 73% of OD pairs have more than 1 trip while about 9% have more than 50 trips. A speed profile analysis is applied to classify day-type and hour in proportion to similar congestion, given that congestion level has to be administrated. The subsequent action after that has been the aggregation of trips in OD pairs among districts with the same congestion pattern and choosing OD pairs with more trips to tackle OD path tendency. During 8 to 10 am
stage, only 22 OD pairs accomplish the condition of having more than 50 OD trips and they account for 1,086 total trips and 108,498 waypoints.

Regarding the aggregations, there are two main dimensions of it, spatial and temporal. Trip Distribution matrix are set up, according initial and ending GPS position trip. Each initial and ending point is mapped into 612 different TAZ regions. Below are displayed in generation, regions where trips start, and attractions i.e. trips whenever trips end.
Unfortunately, TAZ matrix defined according this zoning set up leads to a matrix with most of its entries empty as shown in the following frequency plot:

Figure 3. Frequency plot for the trip distribution matrix aggregated by TAZ zones and hours.

In order to properly address this inconvenience issue in the spatial dimension, trips might be aggregated by Municipality and District.
Trips consist of 18 municipalities, with up to 11 districts. In the case of, \( M_n = \{1,\ldots,18\} \), and District, \( D_{an} = \{1,\ldots,11\} \) \( D_{an} \subseteq M_n \). Another subset of spatial data is defined as quarters. However, is only defined for the municipality of Barcelona. Hence, quarters cannot be used as a distinctive spatial sense.

Aggregating trips by its origin and destination:

![Figure 4. Frequency plot for the trip distribution matrix aggregated by OD Municipality (time dimension not considered)](image)

![Figure 5. Frequency plot for the trip distribution matrix aggregated by OD Municipality and District (time dimension not considered)](image)
The spatial aggregation of observations into more extensive regions might be interpreted as centroids, since weighted and grouped by the averaging the variables that define them. Nevertheless, once temporality is added the seizure of the matrix boosted-up. Grouping trips by Hours, i.e. trips aggregated by the hour when they are generated. The number of cells reaches the amount of 216,600. Unhopefully, the number of empty cells reaches close to 90% of the total.

Due to this fact, hours are grouped in the following sets: [0,5], [5,6], [6,9], [9,10], [10,13], [13,15], [15,17], [15,17], [17,18], [18,20], [20,24]

The way how these elements are grouped is asymmetric. i.e. In the first element [0,5] are considered 5 hours, while in the following [5,6] only 1 hour. In the hopeless task of finding the best arrangement, since there are too many ways of sorting the elements of the set. This specific way on what the hours are arranged has been made considering peak and valley hours considered by the literature, and by observing by eye, when most of the trips are generated.

The total number of cells after the arrangement in the rank of hours specified reduces to 90,250, with approximately 17% of filled cells for the same spatial grouping.

![Figure 6. Frequency plot for the trip distribution matrix aggregated by OD Municipality and District (time dimension grouped in ranged sets)](image)

The aggregation process turns a set of individual observations into a single represented as the average. In order to get more information about the group of observations, also variance of the individuals aggregated has
been taken as aggregation group information.

Table 6. Waypoints defining trips

<table>
<thead>
<tr>
<th>Trip ID</th>
<th>Unique trip identifier in which the waypoint belongs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waypoint Sequence</td>
<td>Waypoints order in the trip, starting from 1 and incrementing by one till the last of trip sequence.</td>
</tr>
<tr>
<td>Capture Data</td>
<td>The capture date and time of the waypoint in UTC, ISO-8601 format i.e. seconds precision</td>
</tr>
<tr>
<td>Latitude</td>
<td>Waypoint latitude coordinates (floating point f6.3)</td>
</tr>
<tr>
<td>Longitude</td>
<td>Waypoint longitude coordinates (floating point f6.3)</td>
</tr>
<tr>
<td>Segment ID</td>
<td>Identifier extracted from the snapping points defined in Latitude, Longitude space into the segments defining the Primary Crown Metropolitan Area of Barcelona.</td>
</tr>
<tr>
<td>Latency</td>
<td>Update time</td>
</tr>
</tbody>
</table>

Outliers’ detection and data preparation is deeply important to not get biased and mislead by values out of the data distribution and yield to untrustworthy conclusions.

Those outliers might be caused for several reasons, for instance, loose information due to the GPS devise gets out of range for a while, out of ordinary congestion due to a one-off problem in the network because of a traffic accident or any other reason. Furthermore, in the process a first approach of the characterization of the data available is done. First of all, univariate descriptive analysis is done for all of the variables separately.

Since there are such a large amount of OD pairs, is not easy to discern whether a certain consistency in the spatial data holds. One way to approach the difficulty is through the assessment of trips for the same OD. However, there is not an automatic way to see whether it might be outli-
ers, since there are too much to supervise all the pairs. Nor is appropriate to fix a threshold for pairs with similar distance between centroids, since network topology might differ significantly between them and be the reason for the dissimilarity. Although there are unsupervised methods based on choice of featured selection, similarity measures or clustering methods, approaches such as in the article of Cai et al., 2013. Mahalanobis distance might be used to discover similar properties from an unknown sample dataset. It measures a distinction between two random vectors from the identical distribution by covariance matrix, and might help to outlier detection in univariate tools.

For example, this metric would identify a fraction of 1-quantile of data declared as potential multivariate outliers based on classical mean and covariance in a normal distribution. Divergences from multivariate normality centre and covariance have to be estimated in a robust way, such as the MCD estimator. The final Mahalanobis distance is suitable for outlier detection.

![Figure 7. Plot of the classical and the robust (based on the MCD) Mahalanobis distance with outliers represented](image-url)
Table 7. Main Statistics of trip variables with raw data

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Trip Mean Speed Kph</th>
<th>Trip Max Speed Kph</th>
<th>Trip Distance (m)</th>
<th>ttmin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>2.316</td>
<td>6.478</td>
<td>1500</td>
<td>0.500</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>34.873</td>
<td>79.853</td>
<td>2945</td>
<td>3.112</td>
</tr>
<tr>
<td>Median</td>
<td>50.169</td>
<td>93.826</td>
<td>8232</td>
<td>12.500</td>
</tr>
<tr>
<td>Mean</td>
<td>50.866</td>
<td>96.682</td>
<td>9950</td>
<td>17.911</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>66.850</td>
<td>110.223</td>
<td>16245</td>
<td>27.492</td>
</tr>
<tr>
<td>Max.</td>
<td>198.196</td>
<td>199.993</td>
<td>24000</td>
<td>183.917</td>
</tr>
</tbody>
</table>

Table 8. Main Statistics of trip variables after outlier cut-off (0.999) detected by robust Mahalanobis distance

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Trip Mean Speed Kph</th>
<th>Trip Max Speed Kph</th>
<th>Trip Distance (m)</th>
<th>ttmin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>3.586</td>
<td>6.44</td>
<td>1500</td>
<td>2.600</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>14.043</td>
<td>36.19</td>
<td>2508</td>
<td>9.155</td>
</tr>
<tr>
<td>Median</td>
<td>18.804</td>
<td>54.43</td>
<td>4121</td>
<td>13.667</td>
</tr>
<tr>
<td>Mean</td>
<td>21.304</td>
<td>55.99</td>
<td>5142</td>
<td>15.274</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>26.572</td>
<td>73.34</td>
<td>7068</td>
<td>19.562</td>
</tr>
<tr>
<td>Max.</td>
<td>61.198</td>
<td>182.79</td>
<td>16379</td>
<td>52.000</td>
</tr>
</tbody>
</table>

The first table shows statistics of trip variables with raw data while in the second table the same statistics are shown but after eliminating the
further observations by the robust Mahalanobis distance measure.

In general terms, it can be observed that in the second table results are much lower than in the first table (such as the difference between the Trip Mean Speed with a 50.169 (raw data) towards the Trip Mean Speed of table 7 with 21,304 (outlier cut-off). A similar pattern can be observed for the rest of the values. After eliminating outliers’ candidates, variables seem to be more reliable, since those are ranged in a narrower interval, and closer to their mean.

For instance, in the first table the rate of values of the travel time are within a range from 0 to 184, while the travel time on the second table shows a rate between 2 to 52, which is a highly difference, since most of the shortest trips that are highly inconsistent due to the brevity of the trip, and those longest trips. Variables regarding the speed, and knowing beforehand, the speed limits in the area of influence, after outliers detection, they look more according what is expected in reality, means of speed values distributed around the mean, and with a standard deviation apparently trustfully. In the case of trip max speed, still there are values for which does not seem to be reliable, values of 6.44 km/h are not expected for vehicles, but since they might be trips with high congestion and short distance, with no other reason of evidence to eliminate them, they are proven to be in the sample.
Figure 8. Boxplots for each variable per each day

As it can be seen, there are no major differences between the variables per each day. For trip mean speed values, it is appreciated that very high values might be outliers, because the average speed in the crown cannot be very high and therefore extreme values can be eliminated with facility. For trip distance, very high values might correspond to very long trips, but it has been found that some of them can be eliminated since they come from another location outside the study area. And for travel time, there is a difference between Wednesday and the rest of the days, possibly due to different levels of congestion according to the day. Trips showing high travel times (more than one hour) are probably outliers since they are atypical. This is related to high values in the trip distance variable, where high values also correspond to long-distance trips. This correspondence among both variables is not surprising because of the degree of correlation that share.
Figure 9. Boxplots for each variable per each hour

The box plot shows that when the working sample is segregated in several hours, differences between rank hours can be appreciated. These differences are more obvious in the values that overcome the 3r quartile, such as the pick hour (6-9, with the maximum number of observations) of trip mean and travel time variables. Furthermore, peak hours, between (6,9] and (10,13] according the sample are those ranks that accumulate more outliers. Although, it might be normal, since are the rank with more values by far, the proportion of outliers is even higher in comparison with the rest. Some insight about the congestion might be the reason for those extreme values, as the travel time is the variable more affected. Even though, mean speed, which could be used as an indicator of congestion, or at least is correlated i.e. the best is the flow in a network, the higher can be the mean speed, does not behave much more different than other related hours (those not out of normal traffic behavior, e.g. (0,5]. Although translated to a subset of the network this could not be the case.
5.2 Origin-Destination pattern stability

INRIX data is passively collected from the corresponding GPS devises of the available private car vehicles sample. This fact implies necessarily that with high probability, each trip is generated in different moments, since each vehicle is independent of each other. Even though, there might be trips that coincide simultaneously in the same instant, or are close enough between the end of the last moment with the start of the next one.

Furthermore, as long as the GPS receiver devices belong to one and only one vehicle, this might be compressed in certain spatial location. This fact happens, since private vehicles are straightforward linked to owners’ mobility patterns.

In order to make easier to understand the characteristic that features the corresponding trips, these are grouped in order to gather trips that could be similar in characteristics and hence, they share features that are inherent in them to facilitate studying the immanence that defines them.

Firstly, as known by the literature and thoroughly been studied, working days as Monday and Friday differ from the rest of working days (Ortúzar and Willumsen, 2011). Furthermore, along the hours of the whole day, trips are not distributed uniformly but on rashes. This phenomenon is the cause for which congestion is produced since the main generation trips are produced in what is known as peak hours and, due to the massive and sudden demand, road capacity might exceed its capacity turning out in congestions and traffic jams.

The stability pattern assessment has been developed in two different approaches, the first one consisting in perform an explanatory analysis in order to permit to see a first insight on how that trip might differ and could be characterized. The explanatory analysis consists in computing some descriptive statistics to give an overall idea of the data, and supplementary plots to see their respective distribution.
Table 9. Statistics showing the moments (1 – 4) of each variable

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Trip Mean Speed Kph</th>
<th>Trip Max Speed Kph</th>
<th>Trip Distance Meters</th>
<th>tt.min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>29.45</td>
<td>67.96</td>
<td>6.723447e+07</td>
<td>16.00042</td>
</tr>
<tr>
<td>Variance</td>
<td>1282.8</td>
<td>5701.01</td>
<td>6.755347e+07</td>
<td>388.72942</td>
</tr>
<tr>
<td>Skewness</td>
<td>75896.44</td>
<td>5.753484</td>
<td>9.454087e+11</td>
<td>13513.85277</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>55849896</td>
<td>672344700</td>
<td>1.547989e+16</td>
<td>673192.34903</td>
</tr>
</tbody>
</table>

Table 10. Correlation between trip variables

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Trip Mean Speed Kph</th>
<th>Trip Max Speed Kph</th>
<th>Trip Distance Meters</th>
<th>tt.min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trip Mean Speed Kph</td>
<td>1</td>
<td>0.63601</td>
<td>0.2685</td>
<td>-0.3540</td>
</tr>
<tr>
<td>Trip Max Speed Kph</td>
<td>0.6360</td>
<td>1</td>
<td>0.4814</td>
<td>-0.0492</td>
</tr>
<tr>
<td>Trip Distance Meters</td>
<td>0.2685</td>
<td>0.4814</td>
<td>1</td>
<td>0.5992</td>
</tr>
<tr>
<td>tt.min</td>
<td>-0.3540</td>
<td>-0.0492</td>
<td>0.599</td>
<td>1</td>
</tr>
</tbody>
</table>
Afterwards, under the guideline that the explanatory analysis provides, Kolmogorov tests are performed with different grouping levels by the categorical variables. A Kolmogorov-Smirnov test is performed in all cases in order to check the similarities amongst the temporality trip characteristics. The Kolmogorov-Smirnov (KS) two-sided test is commonly used to calculate the goodness-of-fit between the empirical distribution of a subgroup of observations and a specific continuous probability distribution (Simard and Ecuyer, 2011) and it develops a comparison of cumulative distribution functions.

Initially, trips are split by each day typology, i.e Monday to Friday, and afterwards within this separation, the test divides them and groups them by each starting trip hour. Afterwards, trips are grouped by ranked hours and finally compared.
Table 11. Significative rang hours (green for non-significant differences, blue for significant differences)

<table>
<thead>
<tr>
<th>Ranked hours</th>
<th>Mean Speed</th>
<th>Maximum Speed</th>
<th>Distance</th>
<th>Tt</th>
</tr>
</thead>
<tbody>
<tr>
<td>'9,10,13'</td>
<td>0.156</td>
<td>0.265</td>
<td>0.333</td>
<td>0.092</td>
</tr>
<tr>
<td>'9,10,13'</td>
<td>5.273e-03</td>
<td>7.273e-02</td>
<td>8.521e-02</td>
<td>1.033e-05</td>
</tr>
<tr>
<td>'10,13,15'</td>
<td>5.569e-12</td>
<td>4.594e-05</td>
<td>0.129</td>
<td>5.329e-15</td>
</tr>
<tr>
<td>'13,15,18'</td>
<td>0</td>
<td>0</td>
<td>0.157</td>
<td>1.886e-09</td>
</tr>
<tr>
<td>'15,17,18'</td>
<td>7.864e-08</td>
<td>0</td>
<td>3.057e-01</td>
<td>1.536e-15</td>
</tr>
<tr>
<td>'17,18,20'</td>
<td>9.320e-09</td>
<td>0.364</td>
<td>0.303</td>
<td>1.069e-10</td>
</tr>
</tbody>
</table>

Fortunately, only two-by-two comparisons are assessed in working days, turning into \( \frac{5 \times 4}{2} = 10 \) comparisons.

Table 12. Results obtained amongst working days (green for non-significant differences, black for significant differences)

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Mean Speed</th>
<th>Maximum Speed</th>
<th>Distance</th>
<th>Travel time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday - Tuesday</td>
<td>0.167</td>
<td>0.838</td>
<td>0.598</td>
<td>0.169</td>
</tr>
<tr>
<td>Monday - Wednesday</td>
<td>2.814e-10</td>
<td>0.051</td>
<td>0.146</td>
<td>3.52e-4</td>
</tr>
<tr>
<td>Monday - Thursday</td>
<td>3.1 e-4</td>
<td>0.077</td>
<td>0.092</td>
<td>0.186</td>
</tr>
<tr>
<td>Monday - Friday</td>
<td>3.848e-15</td>
<td>4.04e-4</td>
<td>6.98e-3</td>
<td>1.05e-4</td>
</tr>
<tr>
<td>Tuesday - Wednesday</td>
<td>4.199-08</td>
<td>0.020</td>
<td>0.234</td>
<td>0.018</td>
</tr>
<tr>
<td>Tuesday - Thursday</td>
<td>0.015</td>
<td>0.085</td>
<td>0.190</td>
<td>0.312</td>
</tr>
<tr>
<td></td>
<td>1.693e-13</td>
<td>1.58e-4</td>
<td>0.025</td>
<td>1.875e-3</td>
</tr>
<tr>
<td>----------------------</td>
<td>----------------</td>
<td>-------------</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td><strong>Tuesday - Friday</strong></td>
<td>0.083</td>
<td>0.646</td>
<td>0.280</td>
<td>0.096</td>
</tr>
<tr>
<td><strong>Wednesday - Thursday</strong></td>
<td>0.010</td>
<td>0.087672</td>
<td>9.629e-4</td>
<td>0.634</td>
</tr>
<tr>
<td><strong>Wednesday - Friday</strong></td>
<td>1.55e-4</td>
<td>0.038</td>
<td>0.515</td>
<td>0.024</td>
</tr>
<tr>
<td><strong>Thursday - Friday</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Computing some basics statistics is not enough to assess whether they are different, since there is not a clear measure of dissimilarity between the groups selected. Hence, multivariate tools as PCA with the use of clustering methods are used in order to ensure stability and homogeneity.

### 5.3 Principal Component Analysis

Once OD Municipality District TAZ aggregation is defined, a principal component analysis is done on the matrix previously set up. The aggregation of data by the pair of origin and destination, as well the temporal dimension in rank hours makes possible to group data and calculate variables as mean, and variance, which characterizes the whole aggregation.

PCA is a technique that allows explaining variation in the variables and bringing to the light patterns in a dataset. Furthermore, dimensionality reduction helps to understand the relationship between those with respect the axis on what the variables are projected based on orthogonal transformation.

In order to decide how many components should be chosen in order to adequately explain the data set, two general criteria are implemented. On one hand, a minimum threshold for total variance explained is defined for instance, 70% of the total explained variance (although it may vary depending on the accuracy needed and the interpretability of the dimensions), the criteria might be focused on the average criteria cutoff of the eigenvalues, retaining up to the last component above the average (see Table 13). On the other hand, based on scree plot, which are the eigenvalues sorted and its evaluation for the largest difference between them
once the threshold of 70% of variance captured is overpassed, dimensions of the projection are chosen.

According these two criteria, the numbers of principal components to be kept are 3, getting a total of 70.79% variance retained.

Table 13. Variance explained by the four first components of the PCA

<table>
<thead>
<tr>
<th>Eigenvalues</th>
<th>Dim. 1</th>
<th>Dim. 2</th>
<th>Dim. 3</th>
<th>Dim. 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance</td>
<td>2.185</td>
<td>1.931</td>
<td>1.548</td>
<td>0.848</td>
</tr>
<tr>
<td>% of Var.</td>
<td>27.31</td>
<td>24.139</td>
<td>19.345</td>
<td>10.602</td>
</tr>
<tr>
<td>Cumulative % of Var</td>
<td>27.31</td>
<td>51.449</td>
<td>70.794</td>
<td>81.396</td>
</tr>
</tbody>
</table>

Figure 16. Percentage of explained variances per each one of the components of the PCA

Figure 16 shows a significant difference among the variance of the third and fourth components.
Figure 17. Variables projected on the two first components of the PCA

To interpret each PCA components, the magnitude and direction of variables’ coefficients are examined between each other and w.r.t PCA components. The larger the absolute value of the coefficient, the larger contribution of the variable in calculating the component.

Aggregated variables as mean trip speed and, maximum speeds which are highly correlated with the second component and variance is well captured by the first component. Mean trip travel time as Mean trip distance are well correlated between each other and w.r.t second principal component, which is apparently reasonable since, one might expect for the same conditions i.e. same trip’s mean speed, the longer trip’s route is, the larger is the time to take the same route.

Regarding the third component, this one is captured through the variance of the before mentioned variables, above all, mean trip maximum speed and trip distance; the first one with a slightly positive correlation with the second component, and the last one with negative.
Figure 18. Variables projected on the second and third components of the PCA

Figure 19. Biplot of projected variables in the two first principal components. Trips characterized by ranked hours are plotted upon the space.

Principal Component Analysis is effective in the identification of outliers since they can analyze the individuals that are isolated from the components by analyzing the contribution of the variables to each of the components. Afterwards, it can be seen what characterizes those individuals and if they belong to a specific group.
5.4 Clustering

In section (1.2), RUM assumptions were exposed. Recalling the first one, all individuals belong to the same homogeneous population.

In order to guarantee this assumption, a clustering procedure is performed. Cluster analysis is useful to group individuals based only on information in data that describes them and their relationship. The objective is to form groups of individuals such that within-group individuals’ similarity is as higher as possible, while difference between groups is as large as possible.

In the concerning case homogeneity is ensured, first by spatial sense, i.e. trips, or individuals, that to be modeled have a common initial and ending region. Intuitively, considering two trips, the first one with the origin and destination the most separated as possible, and the second one with colliding frontier between origin-destination, variables as distance or travel time, which are shown to be correlated, will be affected by the spatial dimension.

Likewise, two trips from different OD, but with similar distance between the respective OD centroids, might differ significantly in the before mentioned variables. This might be caused by congestion or any other network topology, which is expressed in good measure due to temporal dimension.

This imposition takes its sense as long as the number of possible routes, or alternatives for a given OD pair are, in general, lower for a given since is related to a certain TAZ zone. Conversely, in general, the wider the region extension is, the larger the set of segments that shape the network and higher trips’ dissimilarity. Intuitively, for a given pair of origin-destination, the numbers of possible routes taken by any trip compressed in the OD, in general, are lower than if considered any possible route in the network would be considered. This is explained because congestion over the network seldom is distributed uniformly, neither the road topology, e.g. number of lanes for different links may differ significantly, speed limits and some other influent factors having an impact on density and traffic flow.
In general, as more constraints added to the case, as for instance, segments, more similar those groups would be.

Despite the effectiveness of the hierarchical clustering, this method cannot be applied in our study due to lack of memory (CPU).

![Individuals factor map - PCA](image)

Figure 20. Projected clusters on each component of the PCA

PCA Component Analysis has been applied for a better interpretation of the outliers, because we can notice the relation the contribution of the variables with each one of the components of the PCA and therefore identify cluster’s characteristics. Characterization of the clusters according to original spatial-temporal aggregated variables shows that Cluster 1 and 2 are related to V TT Mean and V trip distance (meters) variables, meaning that the values of those variables are higher in the individuals from the sample. The cluster number 3 is well differentiated from the second component, meaning that the distances among the values of the V trip mean variables are quite homogeneous. Regarding the fourth cluster, its longitude is the highest and is not well defined by the components due to the high variability in the mean speed. Observing in detail the scores trips, individuals’ projection in the first two component some outliers are detected:
Table 14. Statistics for outliers found in thought projection of individuals (N=51)

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Trip Mean Speed Kph</th>
<th>Trip Max Speed Kph</th>
<th>Trip Distance Meters</th>
<th>tt.min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>7.68</td>
<td>23.85</td>
<td>1622</td>
<td>1.66</td>
</tr>
<tr>
<td>1st. Qu.</td>
<td>16.24</td>
<td>39.82</td>
<td>2636</td>
<td>8.87</td>
</tr>
<tr>
<td>Median</td>
<td>21.64</td>
<td>55.07</td>
<td>4321</td>
<td>14.83</td>
</tr>
<tr>
<td>Mean</td>
<td>26.95</td>
<td>61.85</td>
<td>6690</td>
<td>16.55</td>
</tr>
<tr>
<td>3rd. Qu</td>
<td>39.7</td>
<td>80</td>
<td>10110</td>
<td>21.27</td>
</tr>
<tr>
<td>Max</td>
<td>59.35</td>
<td>118.2</td>
<td>17680</td>
<td>44.9</td>
</tr>
</tbody>
</table>

![TAZ Generation Cluster: 1](image1.png)

![TAZ Generation Cluster: 2](image2.png)
Figure 21. Primary Crown trips generated for each TAZ Zone in relation with each cluster.
Plots above show graphically how clusters might change in the spatial dimension. Each cluster accounts for a number different of observation, there are zones as for instance, TAZ zones, 391 that contains 1594 trips generation, meanwhile others as 139 contains only 2. This difference, makes undisguised the congestion in the spatial sense.
Furthermore, crossing tables with Generation against Attraction frequencies, show that although they are not symmetric i.e. does not generate in the same proportion as attracted, in most of them it holds e.g. the for 3 TAZ zones that generates a greater number of trips, also attracts in the same order the number of trips. That is likely explained since those zones are the most congested in the network.

Table 15. Variables variance mean for each cluster and according population

<table>
<thead>
<tr>
<th>Cluster</th>
<th>VTMS</th>
<th>VTMx</th>
<th>VTD</th>
<th>VTT</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19,382</td>
<td>4300.27</td>
<td>4.96E+21</td>
<td>1011,795</td>
<td>101</td>
</tr>
<tr>
<td>2</td>
<td>2,875</td>
<td>80.503</td>
<td>1.24E+18</td>
<td>139</td>
<td>12768</td>
</tr>
<tr>
<td>3</td>
<td>2,159</td>
<td>37.380</td>
<td>2.41E+18</td>
<td>205</td>
<td>5303</td>
</tr>
<tr>
<td>4</td>
<td>129.130</td>
<td>508.258</td>
<td>3.86E+18</td>
<td>224</td>
<td>8572</td>
</tr>
<tr>
<td>5</td>
<td>756</td>
<td>14.840</td>
<td>1.79E+18</td>
<td>101</td>
<td>9059</td>
</tr>
<tr>
<td>6</td>
<td>104</td>
<td>8.886</td>
<td>3.58E+18</td>
<td>1.583</td>
<td>8972</td>
</tr>
<tr>
<td>7</td>
<td>203</td>
<td>15.096</td>
<td>1.60E+17</td>
<td>216</td>
<td>18757</td>
</tr>
<tr>
<td>8</td>
<td>350</td>
<td>14.224</td>
<td>3.97E+18</td>
<td>1.336</td>
<td>3362</td>
</tr>
</tbody>
</table>

Legend Table 15

VTMS: Variance trip mean speed
VTMx = Variance trip maximum speed
VTD = Variance trip distance
VTT = Variance travel time
N = number of observations
Table 16. Within sum of square for each cluster and the mean for each one of the cluster
OD Municipality- District

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Within-SS</th>
<th>MTS</th>
<th>MaxTS</th>
<th>MTD</th>
<th>MTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2685</td>
<td>32</td>
<td>249</td>
<td>4389</td>
<td>51</td>
</tr>
<tr>
<td>2</td>
<td>4709</td>
<td>117</td>
<td>699</td>
<td>15815</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>4942</td>
<td>85</td>
<td>215</td>
<td>6465</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>7966</td>
<td>316</td>
<td>367</td>
<td>6889</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>4623</td>
<td>169</td>
<td>899</td>
<td>15995</td>
<td>230</td>
</tr>
<tr>
<td>6</td>
<td>5603</td>
<td>71</td>
<td>412</td>
<td>13193</td>
<td>125</td>
</tr>
<tr>
<td>7</td>
<td>4634</td>
<td>168</td>
<td>673</td>
<td>12871</td>
<td>31</td>
</tr>
<tr>
<td>8</td>
<td>5655</td>
<td>295</td>
<td>545</td>
<td>9936</td>
<td>20</td>
</tr>
</tbody>
</table>

Legend

Within-SS: Sum of Squares
MTS: Mean Trip Speed
MaxTS: Maximum Trip Speed
MTD: Mean Trip Distance
MTT: Mean Trip Time

Table 17. Variance means for each variable and the corresponding number of totals individuals classified in the cluster.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>VTMS</th>
<th>VTMaxS</th>
<th>VTD</th>
<th>VTT</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.925</td>
<td>370710.28</td>
<td>3.59E+1 9</td>
<td>7.150</td>
<td>20305</td>
</tr>
<tr>
<td>2</td>
<td>65.441</td>
<td>3274027.59</td>
<td>3.87E+2 0</td>
<td>65.265</td>
<td>2948</td>
</tr>
<tr>
<td>3</td>
<td>7.527</td>
<td>71380.01</td>
<td>7.22E+1 8</td>
<td>581</td>
<td>6162</td>
</tr>
<tr>
<td>4</td>
<td>87.847</td>
<td>192747.84</td>
<td>8.74E+1 8</td>
<td>483</td>
<td>7623</td>
</tr>
<tr>
<td>5</td>
<td>282.988</td>
<td>7128061.22</td>
<td>3.74E+2 1</td>
<td>1.266.871</td>
<td>165</td>
</tr>
<tr>
<td>Cluster</td>
<td>Within-SS</td>
<td>MTMS</td>
<td>MTMS</td>
<td>MTD</td>
<td>MTT</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>2448</td>
<td>78</td>
<td>568</td>
<td>1024,4806</td>
<td>136</td>
</tr>
<tr>
<td>2</td>
<td>2979</td>
<td>74</td>
<td>220</td>
<td>2400,530,7</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>2702</td>
<td>205</td>
<td>954</td>
<td>1353,3646</td>
<td>43</td>
</tr>
<tr>
<td>4</td>
<td>2641</td>
<td>408</td>
<td>718</td>
<td>7416076,6</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>1522</td>
<td>108</td>
<td>446</td>
<td>6710,880,7</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>2517</td>
<td>23</td>
<td>324</td>
<td>3601715,8</td>
<td>57</td>
</tr>
<tr>
<td>7</td>
<td>2427</td>
<td>18</td>
<td>205</td>
<td>898028,2</td>
<td>23</td>
</tr>
<tr>
<td>8</td>
<td>4039</td>
<td>62</td>
<td>644</td>
<td>14803249,7</td>
<td>250</td>
</tr>
</tbody>
</table>

Table 18. Within sum of square for each cluster and the mean for each one of the cluster TAZ Zone

Stability has been calculated again after clustering the data for the five selected OD.

Table 19. KS test for the first five OD to assess stability. Results only for OD comparisons with non-significant differences, the rest has been discarded for a better understanding (green for non-significant differences, black for significant differences). Only significant combinations from all combinations tested are shown.
5.5 Routes choice set algorithm

The algorithm has been tested in a selected OD, which is the one with largest number of trips aggregated by TAZ Zone. The following properties and some statistics describing the OD trips’ segments are summarized below:

Table 20. Descriptive statistic for all trip variables concerning to the OD TAZ selected. (Speeds measured in km/h, trip distance)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Trip Mean Speed</th>
<th>Trip Max Speed</th>
<th>Trip Distance Meter</th>
<th>Travel time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>22,816</td>
<td>39,758</td>
<td>1581,371</td>
<td>1,166</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>51,602</td>
<td>69</td>
<td>2254,156</td>
<td>1,7</td>
</tr>
<tr>
<td>Median</td>
<td>75,041</td>
<td>90,476</td>
<td>2569,775</td>
<td>1,8</td>
</tr>
<tr>
<td>Mean</td>
<td>69,061</td>
<td>86,327</td>
<td>2858,548</td>
<td>3,380</td>
</tr>
<tr>
<td>3rd Qu</td>
<td>86,340</td>
<td>103,107</td>
<td>2941,500</td>
<td>3,127</td>
</tr>
<tr>
<td>Max.</td>
<td>112,288</td>
<td>129,534</td>
<td>5781,702</td>
<td>15,204</td>
</tr>
</tbody>
</table>

- Number of OD trips: 51
- Number of segments in Cn non repeated: S: 55
Figure 23. Segment plotted for the first OD in spatial-temporal aggregation corresponding to the major number of trips

Table 21. Number of times that a segment overlaps with other segments

<table>
<thead>
<tr>
<th>Segment</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>18</th>
<th>19</th>
<th>22</th>
<th>29</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nº overlapping times</td>
<td>25</td>
<td>11</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Furthermore, there is a subset of segments without defined distance in INRIX: 9 segments, of which 8 of them are 1 segment trip. Thus, do not disturb excessively the performance since those are eliminated with the resulting trips to be 42.

After the comparison two by two between OD trips, the subsequent results are obtained:

- Number of times of coincidence: 320
- Mean of overlapping segments on those trips that coincide: 2.25
- Number of times of no coincidence between segments: 540 (out of 861 comparisons)
The heuristic developed and applied to this data, before starting, searches those trips that does not have any commonality between other rest of trips, i.e. none of the segments in the trip are found out in any other trip.

For this particular OD TAZ Zone, 1 trip has been discarded. This trip, however, must be considered in the resulting route set definition.

As explained before, not a proper way to find a commonality threshold has been found out. Hence,

The resulting routes are plotted in relation with the succession of segments that form the route.
Figure 24. Routes created from trips. Most of them are overlapped, hence cannot be plotted.

In this plot it might be notice how different some routes are between each other, and by assessing in particular some of them, the overlap is highly noticed since,

Figure 25. Route 2: With the most trips aggregated to it plotted for the 4 segments that compounds it.

Figure 26. Two trips belonging to Route 2 (Pictured in Figure 25). Red trip is contained in blue. Hence, they share commonality because of the overlap.
Table 23. Quantiles for the OD TAZ before applying algorithm (Trips) and after routes in terms of segments commonality.

<table>
<thead>
<tr>
<th>Quantiles</th>
<th>0%</th>
<th>20%</th>
<th>40%</th>
<th>60%</th>
<th>80%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trips PSC&gt;0</td>
<td>0.08</td>
<td>1.31</td>
<td>3.01</td>
<td>5.43</td>
<td>8.2</td>
<td>16.33</td>
</tr>
<tr>
<td>Routes PSC&gt;0</td>
<td>0.22</td>
<td>0.22</td>
<td>0.64</td>
<td>0.69</td>
<td>0.69</td>
<td>1.25</td>
</tr>
<tr>
<td>Trips</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.66</td>
<td>4.28</td>
<td>16</td>
</tr>
<tr>
<td>Routes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.25</td>
</tr>
</tbody>
</table>

The results suggest that, while trips have a far variance in terms of segments, routes are more homogeneous. That’s reasonable since, routes are trips aggregated into a common form, this gain in variance does not come at any price, since those routes might override some important properties from the point of view the trips characteristics.

In relation with routes set’s characteristics; this is formed by 15 different routes, plus the one not explored in the heuristic discarded because of lack of commonality with the rest of trips. Within the 15 routes created as output by the algorithm, 5 are actually aggregated trips, the rest 10 are trips isolated which the heuristic did not find out a possible combination for those to be aggregated with. The reason that likely best explains this phenomenon is due to the dissimilarity between most of them. Those trips are mostly scarcely overlapped with other trips, or trips far away from the mean from the point of view of variables and number of segments. (i.e small trips formed by two 1 or 2 segments with those segments only common for few trips, and in the opposite side, long trips far away from the mean that, although might contain few segments in common with other trips, in proportion does not suffice to add it to some route without losing its shape according the threshold taken.

The results in terms of segments are summarized in the following table:
Table 24. Description of routes grouping the trips

<table>
<thead>
<tr>
<th>Route ID</th>
<th>Number of segments</th>
<th>Commonality average</th>
<th>Number of trips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route 1</td>
<td>2</td>
<td>2.16</td>
<td>5</td>
</tr>
<tr>
<td>Route 2</td>
<td>4</td>
<td>9.73</td>
<td>11</td>
</tr>
<tr>
<td>Route 3</td>
<td>3</td>
<td>3.11</td>
<td>4</td>
</tr>
<tr>
<td>Trip 1</td>
<td>2</td>
<td>1.36</td>
<td>1</td>
</tr>
<tr>
<td>Route 4</td>
<td>3</td>
<td>7.51</td>
<td>4</td>
</tr>
<tr>
<td>Route 5</td>
<td>3</td>
<td>7.21</td>
<td>3</td>
</tr>
<tr>
<td>Trip 2</td>
<td>3</td>
<td>4.55</td>
<td>1</td>
</tr>
<tr>
<td>Trip 3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Trip 4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Trip 5</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Trip 6</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Trip 7</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Trip 8</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Trip 9</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Trip 10</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Regarding the number of segments these are summaries into:

Table 25. Number of segments

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd. Qu</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>OD trips</td>
<td>3</td>
<td>4</td>
<td>3,941</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>OD routes</td>
<td>2</td>
<td>2</td>
<td>2,467</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
As it can be observed, in comparison with the original sample, OD routes have much less variance in the number of segments compared to OD trips.

As shown in section 3.5.5 Multinomial model usually needs numerical variables as an input to characterizes each of the choices, in relation with the alternatives. As the route choice set, as already seen, is formed by routes, to characterize the route is imperative. In order to do so, trips considered as substituted by route, are taken as input to compute means and variances respect to each of trips assigned to a given route. Means are computed since there is no clear way on the usage of segments for lack of kind of information, otherwise, some weighting according the segments utilization could be used instead. Even though, and taking into the sample size, for the same route, trips assigned could be weighted according the commonality between each trip with the final route.

On this way, it might be seen as clustering data, by the route morphology so all trips assigned in a road should be not highly different.

Table 26. Calculation of the mean of trips associated with each of the routes, grouped by the mean of each of the trips associated to each one of the routes.

<table>
<thead>
<tr>
<th>Route Variables</th>
<th>Trip Mean Speed</th>
<th>Trip Max Speed</th>
<th>Trip Distance</th>
<th>Travel time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route 1</td>
<td>90,01</td>
<td>104,45</td>
<td>2356,19</td>
<td>1,57</td>
</tr>
<tr>
<td>Route 2</td>
<td>76,55</td>
<td>107,95</td>
<td>2516,14</td>
<td>2,07</td>
</tr>
<tr>
<td>Route 3</td>
<td>70,36</td>
<td>86,41</td>
<td>2236,73</td>
<td>1,94</td>
</tr>
<tr>
<td>Route 4</td>
<td>76,05</td>
<td>100,74</td>
<td>2366,65</td>
<td>1,93</td>
</tr>
<tr>
<td>Route 5</td>
<td>93,72</td>
<td>110,85</td>
<td>2679,9</td>
<td>1,72</td>
</tr>
<tr>
<td>Route 6</td>
<td>60,72</td>
<td>77,67</td>
<td>1889,2</td>
<td>1,87</td>
</tr>
<tr>
<td>Route Variables</td>
<td>Var Mean Speed</td>
<td>Var Max Speed</td>
<td>Var Distance</td>
<td>Var Travel time</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------</td>
<td>---------------</td>
<td>--------------</td>
<td>----------------</td>
</tr>
<tr>
<td>Route 1</td>
<td>185</td>
<td>66,84</td>
<td>21844</td>
<td>0,05</td>
</tr>
<tr>
<td>Route 2</td>
<td>465</td>
<td>475</td>
<td>304</td>
<td>0,38</td>
</tr>
<tr>
<td>Route 3</td>
<td>256</td>
<td>263</td>
<td>40897</td>
<td>0,37</td>
</tr>
<tr>
<td>Route 4</td>
<td>192</td>
<td>319</td>
<td>61114</td>
<td>0.26</td>
</tr>
<tr>
<td>--------</td>
<td>-----</td>
<td>-----</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Route 5</td>
<td>188</td>
<td>87.4</td>
<td>142000</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Table 27. Calculation of the variance of trips associated with each of the routes, grouped by the variance of each of the trips associated to each one of the routes.

For a different OD TAZ, concretely, the third in the amount of trips, with 35 trips after eliminating non-commonality trips, and as it shows in terms of number of segment, trips have less variance.

<table>
<thead>
<tr>
<th>0%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.5</td>
<td>4</td>
<td>5</td>
<td>9</td>
</tr>
</tbody>
</table>

After applying the algorithm the results are:

<table>
<thead>
<tr>
<th>Route ID</th>
<th>Number of segments</th>
<th>Number of trips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route 1</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Route 2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Route 3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Route 4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Route 5</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Route 6</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Route 7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Route 8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Route</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>---------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Route 9</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Route 6</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>
6 Conclusions

Discrete route choice models are a cornerstone in the scope of transport modeling for those assignment or demand equilibrium problems. These are based on the general theory of random utility models, and thus some theoretical hypothesis must hold in the dataset to be applied in order to get a righteous and truthful result. Pre-processing techniques, such as outliers’ detection, spatial and temporal aggregation, stability assessment, and clustering has been carried out in order prepare the base and the conditions on which the route choice set generation must be defined. The expected results for this phase had to let to a scenario on which trips are similar in their variable description, and thus route choice models assumptions and properties such as mutual independence of the routes must hold.

Trips’ variables have been selected in order to draw as much as possible properties of the trips that could extract the best route’s utility concept. These are related to route’s definition through the segments that constitute the path that each trip follows and that eventually defines the route choice. Despite of this, segments characteristic do not necessarily are consistent with trip attributes, e.g. trips large in segments might contained segments without distance definition, or, not having matched the segment due to a problem in the map-matching procedure.

Since Multinomial Logit models are constrained to be limited in the number of choices to model, and these are usually fewer than the population necessary to train models, trips need to be joined in a fancy manner that ensures independence of routes while nor distortion about the original data trips is significantly produced. The heuristic proposed aims to gather different trips into more general identities throughout commonality definition.

It seems a cumbersome task to find a holistic heuristic which might be suitable for all sorts of dataset, in this case, for several ODTAZ. Thus, the full process is difficult to be automatized systematically. Between the most relevant limitations is the imposition of defining a threshold that best suits all OD TAZ. Hence, for each route setting, a threshold has to
be accurately estimated according the inspection of segments and its dis-
tance.

Possible enhancement of the algorithm considers the trip matching pro-
cedure with routes. Route independence, might be improved throughout
choosing different criteria according the routes obtained i.e. matching in-
stead of the maximum commonality, adding some other criteria for those
cases where the maximum is shared between different routes as for in-
stance, to choose a route still not matched. Unfortunately, the number of
routes might increase not controlled and a successfully way to approach
this inconvenience has not been found out.

Considering continuity between segments in the path size computation
(only considering those relatively successions) has been solved, although
not applied in the results for the time spent and check that results do not
vary widely.. Clustering in the OD in relation with the segments charac-
terization might be an option to tackle those samples with different types
of commonality between trips although not successfully accomplished.
7 BIBLIOGRAPHY


9. Chodrow, Philip S., Zeyad al-Awwad, Shan Jiang, and Marta C. González. (2016) Demand and Congestion in Multiplex Transportation Networks edit-
ed by Y. Moreno. PLOS ONE 11(9): e0161738.


Annex

The heuristic code is presented below;
routes <- list(list())
i = 1
for (i in 1:which.min(abs(diff(TAZC$N[1:10])))) {
  require("dplyr")
  seg < trip %>% filter(pca.TAZC$OriginZoneName[i] == trip$OriginZoneName
  & pca.TAZC$DestinationZoneName[i] == trip$DestinationZoneName &
  pca.TAZC$rankHours[i] == trip$rankHours) %>% se-
l ect(TripId, TripMeanSpeedKph, TripMaxSpeedKph, TripDistanceMeters, tt.min
  )

  Cn < sapply(as.character(seg$TripId), function(x) waypoints[which(x == waypoints$ TripId), 'SegmentId'])

  Cn <- sapply(1:length(Cn), function(x) Cn[[x]] < Cn[[x]][lis.na(Cn[[x]])])

  Cn <- sapply(1:length(Cn), function(x) Cn[[x]] <- unique(Cn[[x]])

  names(Cn) <- as.character(seg$TripId)

  Cn[which(names(find.segment) %in% names(sapply(find.segment[which(sappl y(la, function(x) length(x)) == 0]), function(x) length(x))))]

  find.segment <- sapply(unique(unlist(Cn)), function(y) unlist(sapply(1:length(Cn), function(y) names(Cn[y])[any(Cn[[y]] == x)] )))

  names(find.segment) <- as.character(unique(unlist(Cn)))

  PSC.pair <- sapply(1:dim(combn(Cn,2))[2], function(x)
  psc.route(combn(Cn,2)[,x][1], combn(Cn,2)[,x][2]) )

  ## Filtering trips

  sapply(1:length(Cn), function(x)
  length(find.segment[which(names(find.segment) %in% unlist(Cn[[x]])])))

  sapply(1:length(Cn), function(x) sapply(1:length(x), function(y) ))

  ## Deleting those trips that do not have any segment in common

  SID <- sort(table(unname(unlist(Cn)))), decreasing = T)
  SID <- SID[which(SID > 1)]
  SID <- names(SID)
## Initialization

```r
routes<-list(list())
subpath<-SID[1]
  aux<- sapply(SID[which((SID%in%as.character(subpath))==F)],function(x) sapply(1:length(Cn),function(y) sum(Cn[[y]]%in%c(subpath,x)) ))
W<-sapply(1:ncol(aux),function(x) which(aux[,x]==(length(subpath)+1)) )
which(sapply(1:length(W),function(x) length(W[[x]]))/length(SID) > quantile(sapply(1:length(W),function(x)length(W[[x]]))/which(sapply(1:length(W),function(x)length(W[[x]]))>1))/length(SID),prob=0.5)[1] ]
ID<-SID[which((SID%in%as.character(subpath))==F)] candidate<- table(which(sapply(ID,function(x) sapply(Cn,function(y) sum(y%in%c(x,subpath)))) )===(length(subpath)+1),arr.ind = T)[,2]
ID<-ID[as.numeric(names(candidate))] subpath <- c(subpath,ID[which.max(psc(ID,subpath))])
routes[[i]][1]<-subpath
route.match <- sapply(Cn,function(x) sapply(1:2,function(y) psc.route(x/routes[[i]][1] ) ))
ID<-SID[which((SID%in%as.character(subpath))==F)] candidate<- table(which(sapply(ID,function(x) sapply(Cn,function(y) sum(y%in%c(x,subpath)))) )===(length(subpath)+1),arr.ind = T)[,2]
ID<-ID[as.numeric(names(candidate))] p=0.1
while(any(sapply(1:length(Cn),function(x) route.match[|x|]<quantile(PSC.pair[which(PSC.pair!=0)], probs = p)[1] ) )
if(length(ID)>1){
  if(any(unname(table(which(sapply(ID,function(x) sapply(Cn,function(y) sum(y%in%c(x,subpath)))) )===(length(subpath)+1),arr.ind = T)[,2])) & any(psc(ID,subpath)>quantile(PSC.pair[which(PSC.pair!=0)], probs = p)[1]) )
    # # Comparison whether the route with x elements coincides y times with other trips
    if(any(unname(table(which(sapply(ID,function(x) sapply(Cn,function(y) sum(y%in%c(x,subpath)))) )===(length(subpath)+1),arr.ind = T)[,2]))>1 )
```

103
any(psc(ID,subpath) > quantile(PSC.pair[which(PSC.pair != 0)], probs = p)[1]))

if(length(routes[[i]]) == 1) {routes[[i]][[length(routes[[i]])]] <- subpath <-
  c(subpath, ID[which.max(psc(ID, subpath))])
}

route.match <- sapply(Cn, function(x) sapply(1:2, function(y)
  psc.route(x, routes[[i]][[1]])) )

ID <- SID[which((SID%in%as.character(subpath)) == F)]
candidate <- table(which(sapply(ID, function(x) sapply(Cn, function(y)
  sum(y%in%c(x, subpath)))) ) == (length(subpath) + 1), arr.ind = T)[,2])

#ID <-
ID[as.numeric(names(candidate)[unnname(which(table(which(sapply(ID, function(x) sapply(Cn, function(y)
  sum(y%in%c(x, subpath)))) ) == (length(subpath) + 1), arr.ind = T)[,2]) > 1))])]

ID <- ID[as.numeric(names(candidate))]

} else {
  routes[[i]][[length(routes[[i]])]] <- subpath <-
  c(subpath, ID[which.max(psc(ID, subpath))])
  route.match <- sapply(Cn, function(x) sapply(1:length(routes[[i]]), function(y)
    psc.route(x, routes[[i]][[y]])) )
  ID <- SID[which((SID%in%as.character(subpath)) == F)]
  candidate <- table(which(sapply(ID, function(x) sapply(Cn, function(y)
    sum(y%in%c(x, subpath)))) ) == (length(subpath) + 1), arr.ind = T)[,2])

  #ID <-
  ID[as.numeric(names(candidate)[unnname(which(table(which(sapply(ID, function(x) sapply(Cn, function(y)
    sum(y%in%c(x, subpath)))) ) == (length(subpath) + 1), arr.ind = T)[,2]) > 1))])]

  ID <- ID[as.numeric(names(candidate))]

  print(ID)

  }
}
else if(length(routes[i]) == 1) {

    if(sum(route.match > quantile(PSC.pair[which(PSC.pair != 0)], probs = p))[1]) > 2) {
        routes[i][length(routes[i]) + 1] <- subpath <-
        SID[!SID%in%sapply(routes[[i]], function(x) x[[1]])[1]
        route.match <- sapply(Cn, function(x) sapply(1:length(routes[[i]]), function(y) psc.route(x, routes[[i]][[y]])) )
        ID <- SID[which((SID%in%as.character(subpath)) == F)]
        candidate <- table(which(sapply(ID, function(x) sapply(Cn, function(y)
            sum(y%in%c(x, subpath))) ) == (length(subpath) + 1), arr.ind = T)[,2])
        #ID <-
        ID[as.numeric(names(candidate)[unname(which(table(which(sapply(ID, function(x) sapply(Cn, function(y)
            sum(y%in%c(x, subpath))) ) == (length(subpath) + 1), arr.ind = T)[,2]) > 1)])]
        ID <- ID[as.numeric(names(candidate))]
    }
    else{
        print("Corregir quan no arriba el threshold")
    }
}

else

    if(sum(route.match[dim(route.match)[1],] > quantile(PSC.pair[which(PSC.pair != 0)], probs = p)[1]) > 1) {
        # comprobar que el commonality respecte la ruta a almenys més gran del llindar per a dos
        routes[i][length(routes[i]) + 1] <- subpath <-
        SID[!SID%in%sapply(routes[[i]], function(x) x[[1]])[1]
        route.match <- sapply(Cn, function(x) sapply(1:length(routes[[i]]), function(y) psc.route(x, routes[[i]][[y]])) )
        ID <- SID[which((SID%in%as.character(subpath)) == F)]
        candidate <- table(which(sapply(ID, function(x) sapply(Cn, function(y)
            sum(y%in%c(x, subpath))) ) == (length(subpath) + 1), arr.ind = T)[,2])
        #ID <-
        ID[as.numeric(names(candidate)[unname(which(table(which(sapply(ID, function(x) sapply(Cn, function(y)
            sum(y%in%c(x, subpath))) ) == (length(subpath) + 1), arr.ind = T)[,2]) > 1)])]
        #ID <-
        ID[as.numeric(names(candidate)[unname(which(table(which(sapply(ID, function(x) sapply(Cn, function(y)
            sum(y%in%c(x, subpath))) ) == (length(subpath) + 1), arr.ind = T)[,2]) > 1)])]
    }
}
tion(x) sapply(Cn[,function(y)
sum(y%in%c(x,subpath)) ) == (length(subpath)+1),arr.ind =
T)[,2]>1)]) ]
    ID<as.numeric(names(candidate))
} else {

    com<-combn(SID!SID%in%sapply(routes[[i]],function(x
x[[1]]),2)
    # connectivity<- sapply(1:ncol(com), function(w)
sum(sapply(Cn[,function(x) sapply(1:ncol(com), function(z
diff(which(x%in%com[,z] ) )==1)[w,], na.rm=T) )
    #
    # if(length(connectivity)!=0){
    #    com<-com[,which(connectivity>1)]
    # }
    if(ncol(com)==1){
        break;
    }
    else{
        aux<-sapply(Cn[,function(y) sapply(1:ncol(com),function(x
psc.route(com[,x],unlist(y)))
        non.assigned<-sapply(1:length(Cn), function(x
sum(route.match[,x])<=p)
        if(ncol(com)==1 & length(which(aux!=0 &
non.assigned==T))>0){
            print("linea 820")
        }
    }
    #else{
    #com[,which.max(sapply(1:ncol(com), function(x
sum(aux[x,which(sapply(1:length(Cn), function(x sum(aux[x]))!=0 &
non.assigned==T))])
    #com[,which.max(sapply(1:ncol(com), function(x
sum(aux[x,which.max(sapply(1:length(Cn), function(x
sum(aux[x]))!=0 ) )])))
Selecionar algún segment màxim de commonality preferable si no está triat en el sapply(1:ncol(com), function(x)
sum(unlist(aux[x,])))

#routes[[i]][[length(routes[[i]])+1]] <- subpath <-
com[,which.max(sapply(1:ncol(com), function(x)
sum(aux[x,which.max(sapply(1:length(Cn), function(x)
sum(aux[.,x]))!=0 ) ) ) )]

#route.match <- sapply(Cn,function(x) sapply(1:length(routes[[i]]),function(y psc.route(x,routes[[i]][[y]]) ))
#ID<-SID[which((SID%in%as.character(subpath))==F)]
#candidate<-table(which(sapply(ID,function(x) sapply(Cn,function(y)
sum(y%in%e(x,subpath)))) )==length(subpath)+1,arr.ind = T)[,2])
#ID<-ID[as.numeric(names(candidate))]

#

#else if( length(ID)>1){
# else if( length(ID)==0 ){

#Take those that are has never start a route
com<-combn(unlist(SID[!SID%in%sapply(routes[[i]],function(x
x[[1]]),2),2)

aux<-sapply(Cn,function(y) sapply(1:ncol(com),function(x
psc.route(com[.,x],unlist(y) ) )

non.assigned<-sapply(1:length(Cn), function(x
sum(route.match[,x])>p)

if(ncol(com)==1)

add<-
sapply(find.segment[SID[!SID%in%sapply(routes[[i]],function(x
x[[1]]),length(x)]
if(length(which(add==max(add)))>1){

add[which(add==max(add))] &
unique(unlist(Cn[non.assigned]))

if(unique(unlist(Cn[non.assigned]))%in%names(add[which(add==max(add
))])}{

107
SID[unique(unlist(Cn[non.assigned])) %in% names(add[which(add==max(add))])]

} {
} {
}

else {

routes[[i]][[length(routes[[i]]]+1]] <- subpath <-
com[,which.max(sapply(1:ncol(com), function(x)
sum(aux[x,which(sapply(1:length(Cn), function(x) sum(aux[x]))! = 0 &
non.assigned==T]])))]

route.match <- sapply(Cn,function(x) sap-
ply(1:length(routes[[i]]),function(y) psc.route(x,routes[[i]][[y]]))

ID<-SID[which((SID%in%as.character(subpath))==F)]
candidate<-table(which(sapply(ID,function(x) sap-
ply(Cn,function(y)
sum(y%in%c(x,subpath)))) )== (length(subpath)+1),arr.ind = T)[,2]

#ID<--
ID[as.numeric(names(candidate[unname(which(table(which(sapply(ID,function(x) sapply(Cn,function(y)
sum(y%in%c(x,subpath)))) )== (length(subpath)+1),arr.ind =
T)[,2])>1)]))] ]

ID<-ID[as.numeric(names(candidate))]

}

# connectivity<- sapply(1:ncol(com), function(w)
sum(sapply(Cn,function(x) sapply(1:ncol(com), function(z)
diff(which(x%in%com[z] ) )==1)[w], na.rm=T) )

# if(length(connectivity)! =0){
# com<-com[which(connectivity>1)]
# }

#If possible assign someone not assigned
# aux<-sapply(Cn,function(y) sapply(1:ncol(com),function(x)
psc.route(com[,x],unlist(y))
    # nonassigned<-sapply(1:length(Cn), function(x)
    sum(route.match[x]) == 0)
    # com[,which.max(sapply(1:nrow(com), function(x)
    sum(aux[,which(sapply(1:length(Cn), function(x) sum(aux[,x]))] == 0 &
        nonassigned == T)))]
    #
    # # Seleccionar algun segment màxim de commonality preferible si no está triat en el sapply(1:nrow(com), function(x)
    sum(unlist(aux[,]))
    # routes[[i]][[length(routes[[i]]+1]] <- subpath <-
    com[,which.max(sapply(1:nrow(com), function(x)
    sum(aux[,which(sapply(1:length(Cn), function(x) sum(aux[,x]))] == 0 &
        nonassigned == T)))]
    # route.match <- sapply(Cn,function(x) sapply(1:length(routes[[i]]),function(y) psc.route(x,routes[[i]][[y]]))
    # ID<-SID[which((SID%in%as.character(subpath))==F)]
    # candidate<-table(which(sapply(ID,function(x) sapply(Cn,function(y)
        sum(y%in%c(x,subpath))) == (length(subpath)+1),arr.ind = T)[,2])
    # ID<-as.numeric(names(candidate[['unname(which(table(which(sapply(ID,function(x) sapply(Cn,function(y
        sum(y%in%c(x,subpath))) == (length(subpath)+1),arr.ind =
            T)[,2])>1)))']]
    # ID<-ID[as.numeric(names(candidate))]
    } else if(any(unname(table(which(sapply(ID,function(x) sapply(Cn,function(y
        sum(y%in%c(x,subpath))) == (length(subpath)+1),arr.ind =
            T)[,2])!=1) &
        any(psc(ID,subpath) > quantile(PSC.pair[which(PSC.pair!=='0)],[probs =
            p])[1]) ){
    # The last element is assigned
    routes[[i]][[length(routes[[i]]))] <- subpath <-
    c(subpath,ID[which.max(psc(ID,subpath)])]
A new element is assigned
com<-combn(SID||SID%in%apply(routes[[i]],function(x)
x[[1]]),2)
  # connectivity<- sapply(1:ncol(com), function(w)
  sum(sapply(Cn,function(x) sapply(1:ncol(com), function(z)
  diff(which(x%in%com[,z]) ==1)[w,], na.rm=T))
  #
  # if(length(connectivity)!==0){
  #   com<-com[which(connectivity>1)]
  # }
  aux<-sapply(Cn,function(y) sapply(1:ncol(com),function(x)
  psc.route(com[1,unlist(y)]))
  non.assigned<-sapply(1:length(Cn), function(x)
  sum(route.match[,x])==0)
  com[1,which.max(sapply(1:ncol(com), function(x)
  sum(aux[x,which(sapply(1:length(Cn), function(x) sum(aux[,x]))==0 &
  non.assigned==T)])]}
  
  # Selecció amb segment màxim de commonality preferible si
  # no està triat en el sapply(1:ncol(com), function(x) sum(unlist(aux[x,.])))
  routes[[i]][[length(routes[[i]])+1]] <- subpath <-
  com[1,which.max(sapply(1:ncol(com), function(x)
  sum(aux[x,which(sapply(1:length(Cn), function(x) sum(aux[,x]))==0 &
  non.assigned==T)])]
  route.match <- sapply(Cn,function(x) sapply(1:length(routes[[i]]),function(y) psc.route(x, routes[[i]][[y]]))
  ID<-SID[which(ISID%in%as.character(subpath))==F]
  candidate<-table(which(sapply(ID,function(x) sapply(Cn,function(y)
  sum(y%in%c(x,subpath))) ==(length(subpath)+1),arr.ind = T)[,2])
  #ID<-
  ID[as.numeric(names(candidate)[unname(which(table(which(sapply(ID,function(x) sapply(Cn,function(y)
  sum(y%in%c(x,subpath))) ==(length(subpath)+1),arr.ind =
  T)[,2]>(1)]))])
  ID<-ID[as.numeric(names(candidate))]
else{
    com<-combn(SID$SID%in%apply(routes[[i]], function(x)
    x[[1]]),2)
    # connectivity<- sapply(1:ncol(com), function(w)
    sum(sapply(Cn, function(x) sapply(1:ncol(com), function(z)
    diff(which(x%in%com[,z] ) ==1)[w,], na.rm=T) )
    #
    # if(length(connectivity)!=0){
    # com<-com[,which(connectivity>1)]
    # }}
    if(ncol(com)==0){
        break;
    } else{
        aux<-sapply(Cn, function(y) sapply(1:ncol(com), function(x)
        psc.route(com[,unlist(y)]))
    non.assigned<-sapply(1:length(Cn), function(x)
    sum(route.match[x]==0)
    com[,which.max(sapply(1:ncol(com), function(x)
    sum(aux[x,which(sapply(1:length(Cn), function(x) sum(aux[x,]))!=0 &
    non.assigned==T)))))]
    }
    }
    ##Seleccionar algun segment màxim de commonality preferible si
    no està triat en el sapply(1:ncol(com), function(x) sum(unlist(aux[x,])))
    routes[[i]][[length(routes[[i]])+1]] <- subpath <-
    com[,which.max(sapply(1:ncol(com), function(x)
    sum(aux[x,which(sapply(1:length(Cn), function(x) sum(aux[x,]))!=0 &
    non.assigned==T)]))]
    route.match <- sapply(Cn, function(x) sapply(1:length(routes[[i]]), function(y) psc.route(x, routes[[i]][[y]]))
    ID<-SID[which((SID%in%as.character(subpath))==1)]
    candidate<-table(which(sapply(ID, function(x) sapply(Cn, function(y)
    sum(y%in%e(x,subpath))) )==(length(subpath)+1), arr.ind = T)[,2])}
### ID

ID[as.numeric(names(candidate[unname(which(table(which(sapply(ID, function(x) sapply(Cn, function(y)
sum(y%in%c(x, subpath)))) ) == (length(subpath) + 1), arr.ind =
T)[,2]) > 1))])]

ID <- ID[as.numeric(names(candidate))]

}