

Joni Sylvin

# HOW DOES ACCESSIBILITY PREDICT TRIP PRODUCTION?

# Abstract

Joni Sylvin: How does accessibility predict trip production?

Master of Science Thesis

Tampere University

Master's Programme in Computing Sciences

November 2024

---

The aim of this work is to study the effect of accessibility as well as other explanatory factors on trip production with statistical modelling. Previous zonal and national traffic models used in Finland have been found to be lacking for which reason the Finnish Transport and Communications agency Traficom has been assigned to build a national traffic modelling system which this work is part of. Transport models are typically divided to four levels including trip generation, distribution, modal split and assignment with distribution and modal split being often together nowadays. Trip production refers to number of trips made by person during a day or a year.

The data for this study are a survey of a three week period of long distance leisure trips of distance over 100 kilometres with the explanatory variables in data including e.g age, income level, car ownership and employment status. Location information is also included in form of zone id and areal classification variable.

Accessibility measures how people value different opportunities available to them in their action zone. The accessibility indicator for this work is derived from random utility theory and multinomial logit model. Concerning accessibility, the area of leisure apartments, number of jobs in hotel and restaurant industry and population are taken into account in the measure as well as price and time of trip.

The methods used compare different statistical models of which zero-inflated and hurdle model variations proved to be the most fitting ones for this study. The results indicated that the used accessibility indicator does not have a statistically significant effect on trip production in any of the tested statistical models. The limit of 100 km for trips already in trip production stage is limiting the direction of trips and might therefore result in distortions in models. It was also found that income level, car ownership and employment status have a significant role in trip production levels. These findings would indicate that in addition to increased economical resources, available time resources also contribute increasingly to the number of long distance leisure trips made by an individual. For future studies, it could be considered to change the limit of trip length or for instance study a different trip category.

Keywords: accessibility, hurdle models, negative binomial regression, traffic modelling, trip production

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

# USE OF AI IN THESIS

I have utilised AI tools in my thesis:

- No
- Yes

The AI tools utilised in my thesis and their purposes are described below:

Names and versions of AI tools:

Purpose of using AI tools:

Sections where AI tools were used:

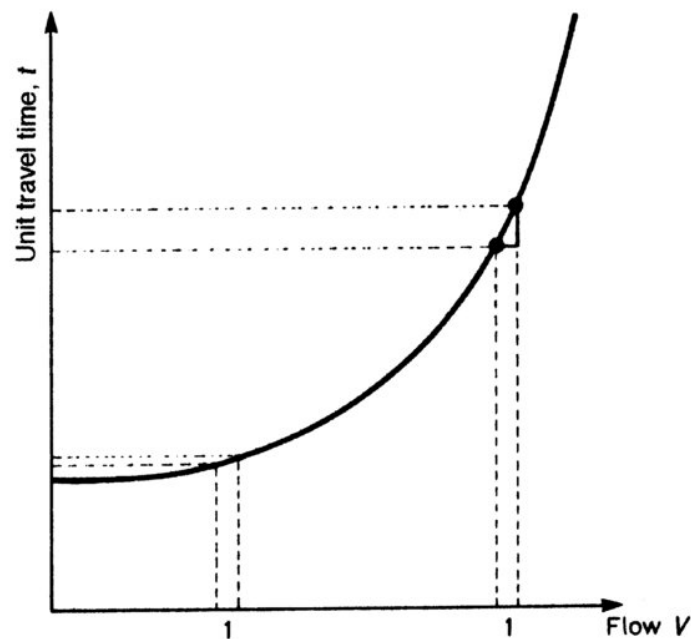
I acknowledge that I am fully responsible for the entire content of my thesis, including the parts generated by AI, and accept accountability for any violations of ethical standards in publications.

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Classical modelling approach . . . . .	6
1.2	Traffic modelling . . . . .	7
1.3	Structure of this thesis . . . . .	8
<b>2</b>	<b>Previous studies and background</b>	<b>9</b>
2.1	Multinomial logit models . . . . .	10
2.2	Random utility theory . . . . .	11
2.3	Accessibility . . . . .	11
<b>3</b>	<b>Methods</b>	<b>15</b>
3.1	Accessibility measures in this study . . . . .	15
3.2	Poisson regression . . . . .	17
3.3	Quasi-Poisson models . . . . .	17
3.4	Negative binomial models . . . . .	18
3.5	Zero-inflated models . . . . .	18
3.6	Stop-go hurdle model . . . . .	19
3.7	Resampling data: cross validation . . . . .	20
<b>4</b>	<b>Analysis</b>	<b>21</b>
4.1	Data description . . . . .	21
4.2	Explanatory variables and attribute selection . . . . .	22
4.3	Goodness of fit measures . . . . .	25
4.4	Results of different models and model selection . . . . .	26
4.5	Testing of the final model candidates . . . . .	28
<b>5</b>	<b>Results and conclusions</b>	<b>31</b>
5.1	Implications and conclusions . . . . .	35
5.2	Final words . . . . .	35
	<b>References</b>	<b>37</b>
	<b>APPENDIX A: Zero inflated negative binomial model</b>	<b>38</b>
	<b>APPENDIX B: Stop-go hurdle negative binomial model</b>	<b>40</b>

# 1 Introduction

In the evolving and ever-changing world and environment, traffic modelling and planning remain as some of the key issues to pay attention to, in order to avoid subsequent problems of lack of quality in them. Congestion in areas with high amount of traffic (illustrated in figure 1.1), delays, poor access and connections, accidents, financial deficits, pollution and lack of infrastructure and travel options are just some of these potential resulting problems. There have been research and different methods on these topics for decades and varying preferred ways and models have been used in different times.



**Figure 1.1.** Effect of addition of one car to congestion in different flow levels. (Ortuzar 2011)

Instances of bad modelling leading to previously mentioned problems include failure to properly include dynamic elements like time in the modelling process. As the division of traffic volume is not even during a day but is instead concentrated on a few hours of a day, failure of taking it into account in traffic modelling and planning could easily lead into a situation where a transport system would not be able to cope with peak periods of the day despite being able to work well with average demand for travel in a zone. Poor modelling may also lead to poor planning and as traffic related investments like building new transport systems, for example a new metro line, do affect the area for decades and require substantial economical investments, that way decisions made with incorrect or insufficient information produced by modelling could result in significant economical losses. It should also be taken into account that the building period of new infrastructure itself often causes traffic related problems like added congestion when certain traffic routes are out of use during construction.

When it comes to this work and its focus area, Finland, all the way from 1990s there have been several zonal and national traffic models which have been produced with varying sets of methods. However earlier models have been found to be lacking for which reason the Finnish Ministry of Transport and Communications has assigned Finnish Transport and Communications agency Traficom to build a national traffic modelling system which this work is part of. With the new traffic modelling system the trips by Finnish people are described, including the number of trips conducted, the direction of them as well as mode and route choices.

The national traffic system consists of both zonal as well as long distance traffic. Here long distance trips are defined to be trips over 100 kilometres long that tend not to be daily trips. Even while these long distance trips make only 3% of all domestic trips, they still make 40% of all travelled kilometres by Finnish people. This results in long distance trips contributing largely in the context of climate policy. (Pastinen 2020)

The aim of this work is to study what kind of a model and explanatory factors explain trip production best as well as to find what connection does accessibility have with trip production if any. This kind of modelling and understanding could prove to be useful in traffic planning issues. Different potential statistical models and comparing their fit to the data are used to find the best selection of a model and the explanatory variables.

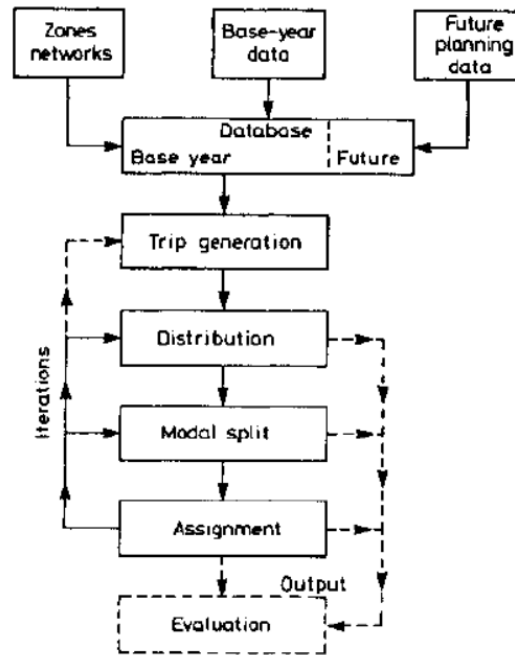
## **1.1 Classical modelling approach**

Despite various developments in transport modelling methods in years and decades, the classical modelling structure can be distinguished with clear different parts included in the process, which is described in figure 1.2.

In the beginning a system is needed for both networks and zones. When it comes to that data, different demographical information as well as information of facilities are needed, for instance shopping and employment related data. In order to be able to conduct the first level modelling which is modelling the trip generation, that data is used. Trip generation part is also the focus of this work when calculating trip production and trip rates and the explanatory variables as factors affecting it. Trip rate is usually measured by number of trips by one individual in one day or one year.

The second level picks up from the first stage of trip generation and divides the trips among destinations, so that the modeller ends up with a trip matrix with trip origins as rows and destinations as columns, where each element tells the volume of trips from one origin to one destination. The third stage is modal split where according to the previous levels a transportation mode is then chosen for each trip, with all transport methods available (such as car, bus, train, or airplane) including public and private ones. The classic model then finishes with an assignment stage, where the volume of travel is computed for a defined network of possible travel connections and for each mode of the trip as well.

It is important to realize that this model itself does not correspond to all possible trip generation changes flexibly but for instance congestion could affect all of route, mode, time or also destination or frequency. For instance one could pick a different



**Figure 1.2.** Transport model classically consists of four main stages including trip generation, distribution, modal split and assignment(route). Trip production is typically defined as number of trips made by person during a day or a year. (Ortuzar 2011)

gym or a restaurant according to the actual situation even if the alternative options would have not been first hand choices. A notable aspect of the model is also that it includes an iterative process between the stages of trip distribution and assignment while trip generation stage remains without changes. That way it can be seen as a shortcoming of the approach that changes of the network are not resulting in effects on trip productions and attractions. This assumption would for instance lead to an outcome that an extension of an underground line being continued to a new location where service was not available previously would not have effect on trips between that zone and the rest of the zones. This is also illustrated in figure 1.2 where the feedback arrow to trip generation stage is marked with a dashed line.

The way such a model works in practice is that it needs to be frequently updated with new data, including for example new flow values (traffic volume values, for instance number of vehicles passing a point in an hour) which in turn then affect travel times and these following sub models of the model will be needed to be re-run again. What is also notable is that the position of modal split has typically changed in different studies where it has sometimes been located before trip distribution as well as next stage after or together with trip generation. (Ortuzar 2011)

## 1.2 Traffic modelling

In traffic modelling there are several characteristics of the domain that usually have to be accounted for. Typically the space dimension derives the need of having to

divide areas into zones, coding them together with the relevant transport networks. Trip length is usually measured by travel time or number of kilometres. The explanatory variables in the related studies may usually include variables like age, gender, household type and size, income level, living area and so on. The modelling is often started with the full model and then eliminating irrelevant variables one at the time.

The level of aggregation is also something to be taken into account affecting the modelling. Aggregated data was in the earlier years used more often until the disaggregated models started to appear more in the 1980s. The disaggregated models tend to be more demanding for the analyst and also demand more data in some cases. The advantages of disaggregated person or household level data come into play often when the model needs to be used and maintained over after for instance demographic areal changes. As it includes more detailed and personalized data attached to more explanatory variables, the predicting power could be more accurate also in the ever-living demographical population type and number.

In traffic modelling there are also several domain specific characteristics to take into account when it comes to planning the study and traffic in general. Transport supply for example is not a product but a service which cannot be stored for later use but has to be consumed in the production location and time instead of being able to store it for later use. Regarding infrastructure it is not possible to build for instance half a bus station or one third of a traffic center so certain lumpiness is also attached in the domain. One has to also bear in mind that investments concerning infrastructure require typically several years of time, a substantial amount of money as well as disruptions to the current traffic scheme in the area of building e.g a new metro line. (Ortuzar 2011)

### **1.3 Structure of this thesis**

The goal of this work is to study the effect of accessibility as well as other explanatory variables on trip production for which we have data with trip counts for persons who participated in a three-week survey questionnaire. The first chapter introduces the topic and its main points and issues. The second chapter in this work consists of going through some of the previous related work, random utility theory and different known and used definitions for accessibility. The relevant statistical methods used and their application in traffic context are defined and explained in the third chapter whereas in the fourth chapter the analysis part follows and fitting the data to different models in practice as well as characteristics of data which is also discussed there. The fifth chapter is a discussion about the obtained results as well as implications of them and their usability in transport planning and the chapter and work are concluded by suggesting ways to possibly enlarge this work further.

## 2 Previous studies and background

Within relevant fields, previous studies have been conducted, one of them being an article from Daly (2006), where advances in modelling traffic generation are discussed. The work in an interesting background for this work for various reasons. In the paper the linkages of different sub-models (including modal split, route choice etc.) of the classic traffic demand model by logsum variables are explained, with those resulting in as a composite measure of travel utility according to the members in choice set selections which form the accessibility measure. In addition to that, definition is given for multinomial logit model of which the logsum is derived. Discussion is included about suitability of different model types regarding if there is a need to differentiate modelling the probability of making one or more trips from making multiple trips or if a single model is sufficient with conclusion being that the former would be usually needed for urban and regional models and the latter would be sufficient when modelling long-distance trips.

One UK-related research interesting from the point of view of this study is the work by Jahanshahi, Williams and Hao (2009) in which the effects of different socio-economic and accessibility factors on trip rates were inspected as well as interactions of chosen variables. The methods included negative binomial regression as alternative for the Poisson regression which according to the study has been found to overestimate the significance of certain explanatory variables in the models. Findings pointed out that relative importance of variables is tied to the trip purposes as well and inclusion of socioeconomic class and individual income as explanatory variables could improve the Transport's National Trip End Model (NTEM). Regarding the accessibility definition of that study, area type and population density were found to be strongly correlated and the area type had a more significant effect when it comes to trip rates.

Krasic and Novacko (2015) studied the effect of transport network accessibility on the number of generated trips using the gravity or entropy definition or the so called "Hansen's measure". The tools used in the work were correlation and regression analysis and connection was found between the accessibility and trip amounts. The study includes an interesting summary of other previous articles which studied the effects of accessibility showing how there are different definitions for it available depending on the situation as well and how the significance of accessibility has varied a lot between studies, models and definitions.

Particularly interesting study is the one made by Jahanshahi, Daly et al. (2017) about travel frequency with UK National Travel survey (NTS) data. Commuting and business trips were modelled with different models suitable for count data type of which many are also relevant for the study of this paper. In addition, the set of explanatory variables in that study include several explanators that are common or at least very similar with this study. Stop-go hurdle model, zero-inflated Poisson model and zero-inflated negative binomial model were used in the paper and are typical choices when modelling such traffic data with a lot of potential zero values in trip counts. The goodness of fit and information criteria used including AIC and

log-likelihood values share common ground between the two studies as well. There are differences in the setting, as in this paper the trips are narrowed down to long distance leisure trips with distance over 100 km instead of business and commuting trips and areal differences are also present as the data in the travel frequency data is from the United Kingdom instead of Finland which is the location of the data in this study. For instance, increase in income seems to be associated with making more trips in the travel frequency study findings.

In a work by Kristoffersson, Berglund and Algiers (2020) where tour generation study was done. In the study daily tour pattern of travellers are included instead of trying to model tours with varying purposes completely independently of each other. Daily limitations are taken into account and findings are made about car access and driving license affecting positively to the probability of multiple trips per day. Interestingly a connection is found between accessibility and non-mandatory trips correlating positively as well as effect of weekdays and holiday seasons on trips. High income was found to decrease the probability for tours of inexpensive activities. Many of similar explanatory variables are available in the context of this work as well.

## 2.1 Multinomial logit models

The accessibility measure used in this study is based on multinomial logit models, which are defined in this section. In these models, the response variable  $Y_i$  is categorical with binary or multiple possible outcomes. That being said,  $Y_i$  is following categorical distribution

$$(2.1) \quad \text{Cat}(\theta_{i1}, \theta_{i2}, \dots, \theta_{im}),$$

where

$$(2.2) \quad P(Y_i = "1") = \theta_{i1}, P(Y_i = "2") = \theta_{i2}, \dots, P(Y_i = "m") = \theta_{im}$$

The form of the multinomial logit model is:

$$(2.3) \quad \log\left(\frac{\theta_{ik}}{\theta_{i1}}\right) = \beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \dots + \beta_{pk}x_{ip}, \quad k = 2, 3, \dots, m.$$

Then the probabilities  $P(Y_i = "k") = \theta_{ik}$  have the following forms under the multinomial logit model:

$$(2.4) \quad \theta_{i1} = \frac{1}{1 + \sum_{j=2}^m e^{\beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2} + \dots + \beta_{pj}x_{ip}}}, \quad j = 2, 3, \dots, m.$$

$$(2.5) \quad \theta_{ik} = \frac{e^{\beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \dots + \beta_{pk}x_{ip}}}{1 + \sum_{k=2}^m e^{\beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \dots + \beta_{pk}x_{ip}}}, \quad k = 2, 3, \dots, m.$$

(Isotalo 2023)

## 2.2 Random utility theory

The random utility theory is an important framework for discrete choice models. There are some general assumptions it includes. It is stated that an individual in population  $Q$  has all the information in order to maximise the net personal utility for himself considering physical, legal, social and budget type of restrictions. There is also a set of alternatives  $\mathbf{A} = \{A_1, \dots, A_j, \dots, A_N\}$  and set  $\mathbf{X}$  of vectors which include values regarding the individual and the alternatives. In addition the choice made by the individual is assumed taken beforehand with the restrictions already taken into account.

Every option  $A_j \in \mathbf{A}$  also includes a net utility  $U_{jq}$  for an individual  $q$  and all information affecting the decision taken by the individual cannot possibly be known by the analyst so one has to divide  $U_{jq}$  into two different parts. One of the parts is the component which represents a function of the attributes that can be measured,  $V_{jq}$ , a function which consists of the measured attributes  $\mathbf{x}$  (for an individual there is a set of attributes  $\mathbf{x} \in \mathbf{X}$  and a choice set  $A(q) \in \mathbf{A}$ ). The other component  $\varepsilon_{jq}$  is the random part consisting of individual tasters and idiosyncrasies as well as observational and measurement error from the analyst.

Compiled, the net utility  $U_{jq}$  can be formed by equation  $U_{jq} = V_{jq} + \varepsilon_{jq}$ , to include the 'irrationalities' occurring in the process of picking a choice. That allows for instance two individuals to have same values of attributes yet remaining with the possibility of them ending up making different choices. (Ortuzar 2011)

## 2.3 Accessibility

Accessibility is one of the key parts of this work and it is important to understand it in the traffic context in order to implement its effect on traffic production. In transport, accessibility is defined as potential for traveling to an activity or to interact with a person. Higher level of accessibility for instance restaurant wise means a larger number of restaurants of different type in the action zone of the individual. Multiple measures exist for it of which four are presented in this work. The usefulness of each of them varies by the situational context.

### Distance to nearest location

One of the simplest accessibility measures is distance to nearest location which measures the distance to the closest location of interest group, for instance medical services, shopping center or subway station. This approach does not take into account the size of the attractions leading to all locations being treated as equally valuable. Unlike in many other measures, distance to nearest location does not account for cumulative effect of multiple locations regarding accessibility. If there are multiple similar accessible locations available within a reasonable distance, only the closest one will be accounted for. Accessibility in this definition is measured by:

$$(2.6) \quad A^{ip} = \min_{j \in L^p} (d_{ij})$$

in which  $A^{ip}$  is the measure of accessibility of zone  $i$  with location type  $p$ .  $L^p$  on the other hand is a set which includes locations which are of a type  $p$ . Finally  $d_{ij}$  describes the distance between locations  $i$  and  $j$  in set  $L^p$ .

As can be seen this restricted location model results in closest location being always chosen with probability 1, regarding location  $j$  for purpose  $p$  with the condition of the person being located in zone  $i$ .

$$(2.7) \quad P_j^{ip} = \begin{cases} 1 & \text{if } d_{ij} = \min_{j' \in L^p} (d_{ij'}) \\ 0 & \text{otherwise} \end{cases}$$

in which  $P_j^{ip}$  describes the probability in a situation where one is located in zone  $i$  and about to choose location  $j$  for travel purpose  $p$ .

### **Isochrone measure/cumulative count measure**

Another measure is the isochrone measure which is also known as the cumulative count measure. It differs in multiple ways from the distance to nearest location measure. Firstly, it sets the size of the accessibility zone by defining the maximum distance for locations that can be considered to be contributed to the accessibility of a zone  $i$  for an activity purpose  $p$ . The size of the activity type  $p$  is measured in the definition as well. This could for instance be a size of an amusement park which could be measured by number of employees or other relevant value. However, in this approach there is no differentiation between distances inside the action zone inside the defined distance limit. For example, two destinations within the reach of five and thirty minutes will be valued the same distance wise as long as the limit is set to at least to thirty minutes. On the other hand, if the limit is e.g thirty minutes, a purpose destination within a 35 minute distance and a 28 minute distance would be handled in a totally different way, with the former not being accounted at all in the accessibility measure whereas the latter would be taken fully into account which may not make sense intuitively. There is also neither theoretical nor empirical background or standard for setting the maximum limit but typically 30, 40 and 45 minutes have often been used.

Below are the form of the accessibility and the probability form of making a choice in this definition of accessibility, where  $X_j^p$  is the size of activity type  $p$  (eg. number of stores or restaurants) at location  $j$  and  $L_{D|i}^p$  is a set that includes the locations of activity type  $p$  that are located within either a maximum travel time or distance  $D$  of zone  $i$ :

$$(2.8) \quad A^{ip} = \sum_{j \in L_{D|i}^p} X_j^p$$

$$(2.9) \quad P_j^{ip} = \begin{cases} \frac{X_j^p}{\sum_{j' \in L_{D|i}^p} X_{j'}^p} & \text{if } j \in L_{D|i}^p \\ 0 & \text{otherwise} \end{cases}$$

### The gravity measures/entropy measure

The gravity measures which are also known as entropy measures extend the previous model by adding an impedance function  $f(d_{ij})$  which has the role of weighting the value of a destination inside the action zone by weighting closer distance more and distances located further away but yet in the action zone less. Another aspect to notice in this measure is that the set of locations are not formed by a threshold that would be cut-off arbitrarily but instead a choice set,  $L^{ip}$  is formed. Choosing the choice set depends on situational context including practical considerations and goals of the analysis and therefore varies case-by-case. In this approach, the choice probability of a single option is continuously decreasing when the time or distance to the location is increasing. The total accessibility can then be measured as a sum of the choice probabilities.

$$(2.10) \quad A^{ip} = \sum_{j \in L^{ip}} X_j^p f(d_{ij})$$

The location choice model of this measure is of the following form:

$$(2.11) \quad P_j^{ip} = \frac{X_j^p f(d_{ij})}{\sum_{j' \in L^{ip}} X_{j'}^p f(d_{ij'})}$$

Previously gravity models have been in usage for instance in geography and travel demand modelling cases. The reason why the models are called gravity models is because they have been derived by analogy to Newton's Law of Gravity which states that there is a spatial interaction between two points in space which is proportional to the size of the location of attraction  $X_j^p$  and the relation turns inverse when it comes to time or distance between the two points ( $f(d_{ij})$ ).

### Random utility based measures

In random utility based measures the destination choice is most commonly defined by the general form of the multinomial logit model (MNL). In the model, there are explanatory variables included and the corresponding parameters explaining the local choice but in the total utility an error term is added to  $U_{jq}$  to measure the assumption that the most rational choice will not always be taken by the individual. Individual differences of perception of utility in choice making may occur in addition to the population level explanatory variables. In the case of multinomial logit model as the form of random utility model, the general form of the destination choice model is:

$$(2.12) \quad P_j^{ip} = \frac{e^{V_j}}{\sum_{j' \in L} e^{V_{j'}}} = \frac{e^{\beta Z_j}}{\sum_{j' \in L} e^{\beta Z_{j'}}}$$

where  $V_j = \beta Z_j$  is the systematic utility of alternative  $j$ ,  $Z_j$  is the vector of explanatory variables, and  $\beta$  is the (row) vector of parameters.

Perceived utility by a decision maker is:

$$(2.13) \quad U_{jq} = V_{jq} + \varepsilon_{jq}$$

In these measures it is assumed that the person picks a choice that results in producing the maximum perceived utility which is represented by  $U_j$ . Expected maximum utility regarding the Multinomial Logit model connected to the choice is given by:

$$(2.14) \quad I^{ip} = E[\max_j(U_j)] = \ln \left( \sum_{j \in L^{ip}} e^{\beta Z_j} \right)$$

In formula 2.12 where the numerator of it is the logarithmic choice option and the denominator measures the logarithmic sum of all the options, the denominator which measures the total accessibility of travel purpose  $p$ , is called logsum. This expected maximum utility can be shown to be the consumer's surplus for this choice which makes it a standard measure of economic benefit. That way, accessibility can be defined with this same measure as can be seen by comparing formula 2.14 and formula 2.15.

$$(2.15) \quad A^{ip} = \ln \left( \sum_{j \in L^{ip}} e^{\beta Z_j} \right)$$

(Miller 2020)

## 3 Methods

### 3.1 Accessibility measures in this study

As the data in this work are long distance leisure trips that are over 100 kilometres long, the accessibility measure has been formed accordingly. Of the accessibility definitions the accessibility indicator in this work is based on the random utility based one which is based on multinomial logit model and the logsum(see eq. 3.2). Accessibility measures are available in data for aerial travelling, long distance buses, for train and car trips and there is a merged accessibility available for all of them combined as well as so called sustainable accessibility measure with car accessibility excluded.

#### Calculation of the accessibility parameters

The calculation of the parameters of accessibility used in this study is based on the mode destination choice model, following the random utility theory approach referenced from Miller in the previous chapter. Tree structure logit model which is also known as nested logit model and is in connection to the multinomial logit model, is used:

$$(3.1) \quad p_{md} = p_m p_{d|m} = \frac{\exp V_m}{\sum_{m'} \exp V_{m'}} * \frac{\exp V_{md}}{\sum_{d'} \exp V_{md'}},$$

with logsum

$$(3.2) \quad V_m = \theta \log \sum_{d'} \exp V_{md'},$$

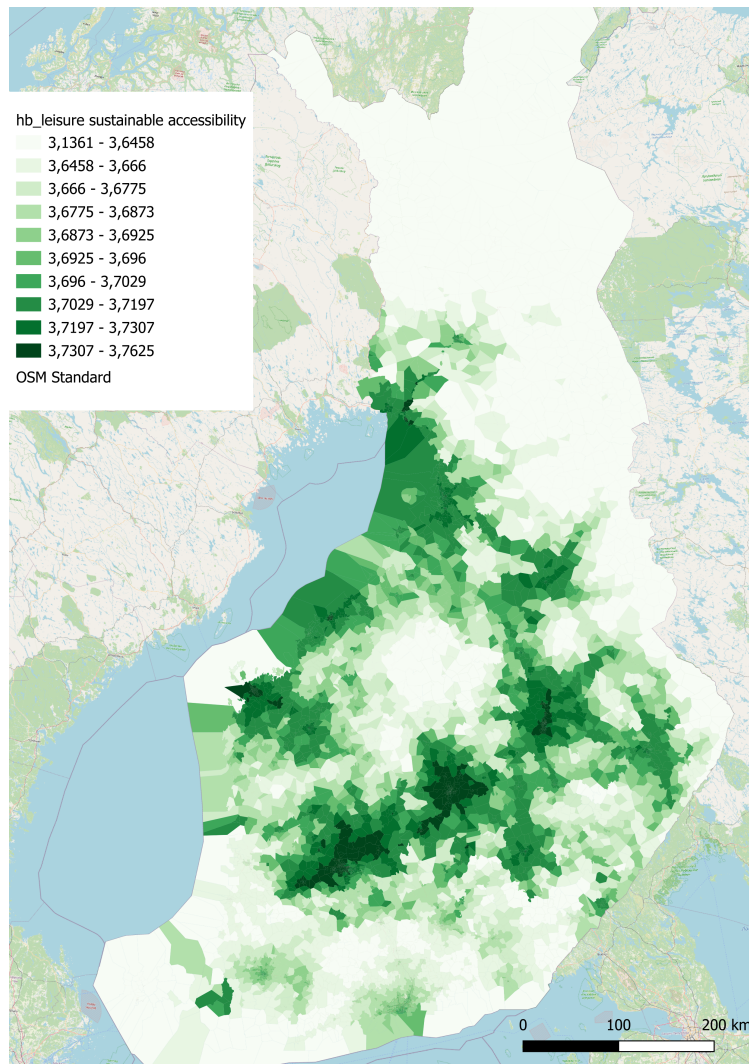
being the merged utility for selecting mode  $m$  and  $V_{md}$  being the utility for  $m$  and destination  $d$  while  $\theta$  is a weight coefficient. The attraction variables describing the size or quantity of model zones are included in the utility functions and referred as size variables.

The form of the utility function with size variables is defined as:

$$(3.3) \quad V_{md} = \sum_r \beta_r x_{rmd} + \phi \log S_d$$

where  $x_{rmd}$  gives component  $r$  of the level of service offered by mode  $m$  to destination  $d$ .  $S_d$  is a measure of the quantity or size of the attraction of zone  $d$  and  $\phi$  is a weight coefficient for  $S_d$ .

Regarding  $S_d$ , the area of leisure apartments, number of jobs in hotel and restaurant industry and population(visiting trips, summer house trips) have been included as its parameters. The impedance function  $\beta_r x_{rmd}$  includes price of trip and time of trip as impedance parameters. As mentioned, the utility function includes the size



**Figure 3.1.** Accessibility of long distance leisure trips in Finland. Accessibility in this work is a continuous explanatory variable.

of the target as well (in  $S_d$ ). For instance, if assuming two targets within the same distance from home, given that the amount of employees in two target companies (e.g amusement parks) would be the same, the utility would be double sized for the target with double number of employees.

As can be seen from the accessibility measure in figure 3.1 the highest accessibility values are in locations and zones which are far enough from the highest population and number of jobs values (in other words far enough from eg. Helsinki) which are in the accessibility measure as parameters. This naturally results from using the distance of 100 kilometres as a limit of a long distance leisure trips. For instance for residents in Helsinki the long distance travel connections are good but with this accessibility indicator the centralized population and job number areas are too close to contribute to the accessibility. As another example, zones that are located over 100 kilometres from Helsinki would benefit of the utility of attractions in Helsinki in this definition of accessibility of long distance trips as only attractions with minimum distance of 100 kilometres are contributing to the value of accessibility according

to the distance definition. This can also clearly be concluded from the accessibility map in figure 3.1. That should be noted as it could be a limitation of functionality of the accessibility measure in this study.

### 3.2 Poisson regression

In a situation where count data is modelled statistically, Poisson models are the starting ground. In such situations the response variable  $Y_i$  gets values consisting of integers that are nonnegative of a form  $y_i = \{0, 1, 2, 3, \dots\}$  Poisson distribution and negative binomial distribution are the most usual options for modelling count data. The probability mass function (PMF) of the Poisson distribution is given by:

$$(3.4) \quad P(Y_i = y_i) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

where  $y_i$  is the number of events ( $y_i = 0, 1, 2, \dots$ ) and  $\lambda$  is the average number of events in a given interval. In Poisson models, there is also a constraint where the expected value and variance are equal as can be seen from their definitions.

$$(3.5) \quad E(Y_i) = \mu_i$$

and

$$(3.6) \quad Var(Y_i) = \mu_i$$

Examples of link functions for the response variable include identity link, log link and square root link:

$$(3.7) \quad \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip},$$

$$(3.8) \quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip},$$

$$(3.9) \quad \sqrt{\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip},$$

(Isotalo 2023)

### 3.3 Quasi-Poisson models

Quasi-Poisson is a differentiation of Poisson models and accounts for overdispersion in count data. The difference between the two is the assumption of variance. In Poisson models, the amount of variance is set and assumed to be equal to the expected value but in practice this might not always be a valid assumption. So the

quasi-Poisson models allow overdispersion and underdispersion which mean that the variance might be growing at a different rate than the expected value and be of a different value. If  $\phi$  is larger than one, overdispersion is found from the model whereas values below one indicate underdispersion.

$$(3.10) \quad E(Y_i) = \mu_i$$

and

$$(3.11) \quad Var(Y_i) = \phi \mu_i$$

The form of unbiased estimate for  $\phi$ :

$$(3.12) \quad \tilde{\phi} = \frac{\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}}{n - (p + 1)} = \frac{X^2}{n - (p + 1)}.$$

(Isotalo 2023)

### 3.4 Negative binomial models

In the case of clear overdispersion found, it is also a possibility to use negative binomial model. When the form of the probability mass function is

$$(3.13) \quad f(y_i | \mu_i, \theta) = \frac{\Gamma(y_i + \theta)}{\Gamma(\theta) y_i!} \cdot \frac{\mu_i^{y_i} \theta^\theta}{(\mu_i + \theta)^{(y_i + \theta)}, \quad y_i = 0, 1, 2, 3, \dots$$

the random variable  $Y_i$  follows negative binomial distribution  $Y_i \sim \text{NegBin}(\mu_i, \theta)$ , and then

$$(3.14) \quad E(Y_i) = \mu_i$$

with variance structure:

$$(3.15) \quad Var(Y_i) = \mu_i + \frac{\mu_i^2}{\theta}$$

(Isotalo 2023)

### 3.5 Zero-inflated models

Sometimes the case is that zero value occurs more frequently than it tends to under Poisson or negative binomial distribution. In this case zero-inflated models are a modelling option in the event of modelling non negative count data response variable

$Y_i$  with realisations  $y_1, y_2, \dots$  having the value zero too frequently. The probability structure of the zero-inflated Poisson model is:

$$(3.16) \quad P(Y_i = 0) = \theta_i + (1 - \theta_i)e^{-\mu_i}$$

$$(3.17) \quad P(Y_i = y_i) = (1 - \theta_i) \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, \quad y_i = 1, 2, 3, \dots$$

$$(3.18) \quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

The form of the expected value  $\mu_i$  with the commonly used log link is:

$$(3.19) \quad \mu_i = \theta_i \cdot 0 + (1 - \theta_i) \cdot e^{(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})}$$

The logit link structure in the case of parameter  $\theta_i$  depending on the explanatory variables:

$$(3.20) \quad \text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1 - \theta_i}\right) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \dots + \alpha_p x_{ip}$$

where  $\alpha_0, \alpha_1, \dots, \alpha_p$  are unknown parameters.

(Isotalo 2023)

So in zero-inflated models there are two processes of which the first one produces zeros only and the second one follows Poisson distribution and non negative integer values occur starting from zero. This is different from stop-go hurdle models explained in the next section, where the second process is truncated at zero making the first process of the mixture the only one producing zeros.

### 3.6 Stop-go hurdle model

Just like in zero-inflated models, in stop-go hurdle models there are two processes where the first one is producing zero values and the second one produces positive count values. The difference to zero-inflated models is that the model is truncated at zero so only the first process can produce zero values.

In the first stage there are two ways to formulate the probability of making a trip at all, which is based on the logit model with following specifications:

$$P(y_i > 0) = \frac{1}{1 + \exp(-v_i)} \quad \text{or} \quad P(y_i = 0) = \frac{1}{1 + \exp(v_i)}$$

$v_i$  being the utility function of making at least one trip:

$$(3.21) \quad v_i = \gamma_0 + \gamma_1 z_{1i} + \dots + \gamma_m z_{mi} = Z_i' G$$

with  $Z_i$  being the vector that includes the explanatory variables (for instance household socioeconomic characteristics and similar factors) while  $G$  is a vector of a form:

$$(3.22) \quad G = (\gamma_0, \gamma_1, \dots, \gamma_m)'$$

consisting of an array of unknown coefficients to be estimated.

In the next phase the number of trips are estimated with condition  $\Pr(y_i > 0)$ ; recursive method is used in the default stop-go for the estimation of trips: first choice decides whether exactly one trip or two or more trips will be made; then, given that at least two trips are to be made, the next level is then a choice between exactly two trips or at least three trips to be made and so on.

When  $\Pr(\text{stop})$  is assumed to be the probability of stopping at each level (meaning stopping at one relative to doing more than one trip, stopping at two relative to doing more than two trips and so on) and being the same for all levels, the stop-go process ends up being a geometric count model. The probability of making  $y_i$  trips is given by:

$$(3.23) \quad \Pr(y_i | y_i > 0) = \Pr(y_i > 0) \Pr(\text{stop})(1 - \Pr(\text{stop}))^{y_i - 1}$$

where

$$(3.24) \quad \Pr(\text{stop}) = \frac{1}{1 + \exp(v_{gi})}$$

and  $v_{gi}$  is the utility of 'go' (meaning making more trips, relative to 'stop' at certain level) given by:

$$v_{gi} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ki} = X_i' \beta$$

Regarding equations containing  $z_{i1}$  to  $z_{im}$ , those consist of  $m$  regressor variables (vector of  $Z_i$ ) and explain the probability of making at least one trip and  $x_{i1}$  to  $x_{ik}$  form a set of  $k$  explanatory variables (vector of  $X_i$ ) resulting as the utility of making more trips (go) when an individual makes at least one trip. It is also notable that the  $Z$ s and the  $X$ s may or may not have common terms.

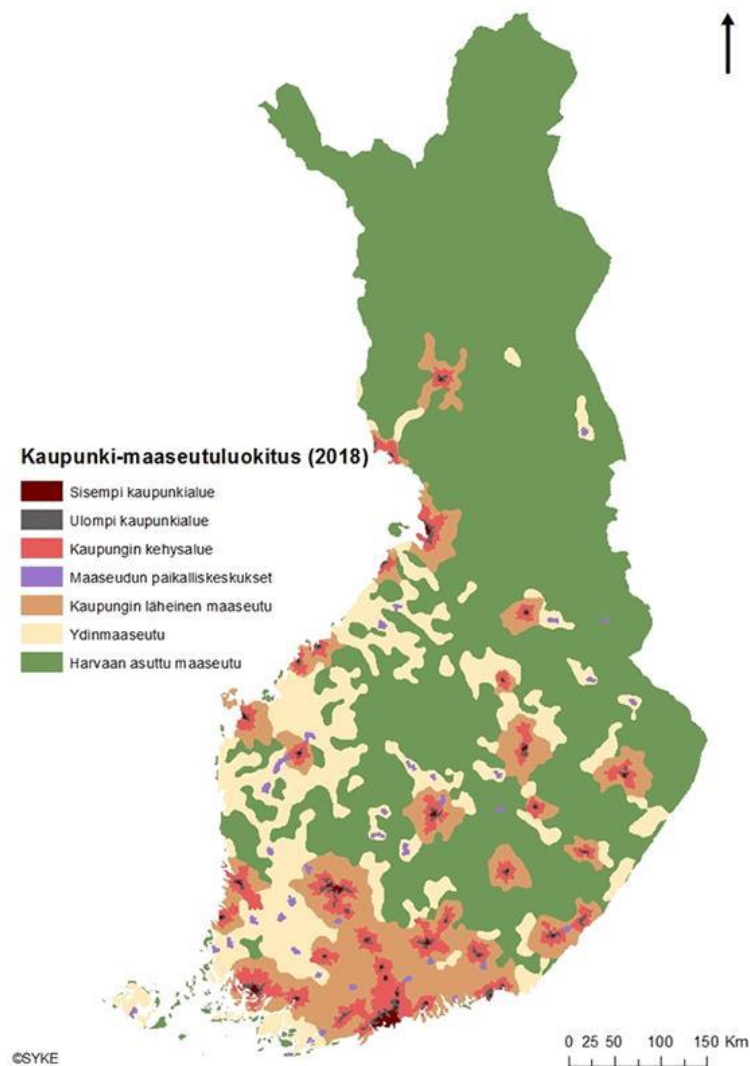
(Jahanshahi 2017)

### 3.7 Resampling data: cross validation

As a method of assessing the fit of a statistical model more accurately a different set of tools are available. In this work, cross validation is used. In order to achieve this, the data is split into training data and test data. The split data segments are kept separate and the size of the training data should typically to be as large as possible as it is intended to be used for estimating the parameters of the model.

Test data, on the other hand, is used for testing the fit of the model. Naturally it is also supposed to be as large as possible so compromise is needed when splitting the data. In  $K$ -fold crossvalidation the method is to split data in a  $K$ -fold partition where for a single  $k$  experiment the training part consists of  $k - 1$  folds and the fold not included is kept for testing. That leads to  $k - 1$  folds resulting in  $k - 1/k * 100\%$  of the data and in this work 10-fold crossvalidation is used resulting in 90% of the data being used for training the model per experiment while the remaining 10% is used for testing. (Emmert-Streib 2023)

## 4 Analysis



**Figure 4.1.** YKR classification consists of seven different area classes, including in the order listed next to the map: inner urban areas, outer urban areas, peri-urban areas, local centres in rural areas, rural areas close to urban areas, rural heartland areas and sparsely populated rural areas. (SYKE 2018)

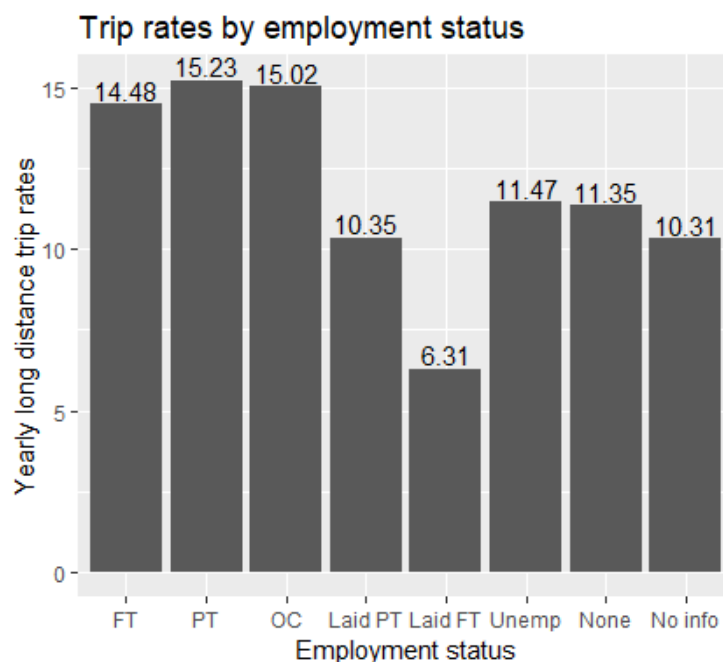
### 4.1 Data description

The data used in this work consists of HLT 2016 personal traffic research data. (Pastinen et al. 2016) The respondents were asked questions about their trips and their background information during the tracking period of three weeks. The focus in this work are the long distance leisure trips (distance of trip being over 100

kilometres). The data was initially in several different sets including for instance trips, persons, YKR areal classification (figure 4.1), zones and accessibility tables. By combining them, the total amount of trips by one person during the tracking period were calculated and these were transferred into one day and one year trip rates by dividing and multiplying them with appropriate calculation operations for data rows and columns and multiplying them with the weight parameters of the explanatory groups.

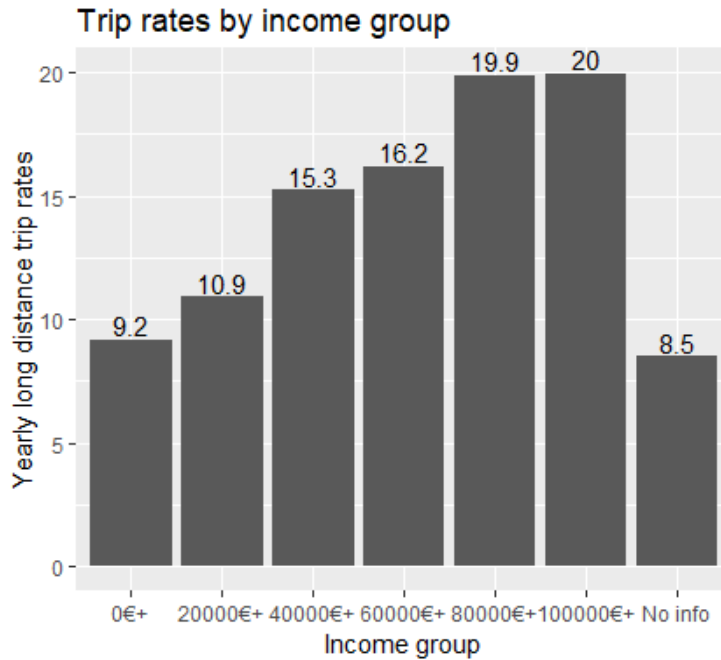
Due to low amount of observations in some explanatory variable groups some merging was conducted for the explanatory variables in order to get more coherent and reliable results statistically. Income groups were compiled into smaller number of groups as in the highest income groups there was shortage of observations. The same was applied to the number of owned cars variable transforming it into a binary car ownership variable by merging all the positive integers (owned cars) into one value of owning a car and zero values formed the group with persons without a car. There was also other merging done by forming three groups of the 7 YKR classification groups due to the similar results in between the subgroups in urban, rural and sparse rural areas. In the statistical modelling stage, also the non-working groups were merged by adding laid off part time, laid off full time, unemployed and outside workforce groups into one group of non-workers. Further details of this merging process will be given in section 5.1.

## 4.2 Explanatory variables and attribute selection



**Figure 4.2.** Employment status groups include full time, part time, occasional, laid off part time, laid off full time, unemployed, outside of work force and group who decided not to disclose this information.

In the first stage of analysis, comparisons of trip rates were conducted based on the available explanatory variables in the data. These include car ownership, income level, employment status and YKR areal classification. The trips data, persons data and zonal classification data were compiled together in order to obtain the total number of trips and the resulting trip rates for long distance leisure trips. Most of the time the basis for development of traffic models is based on compiling such a cross table which is to be used as a production model for trip rates.

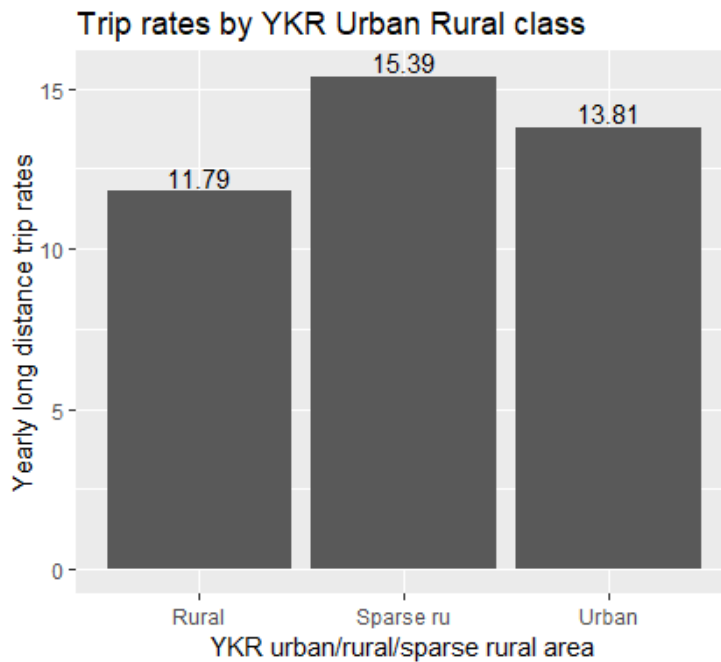


**Figure 4.3.** Trip rates by income group.

In the figures the chosen plots of the trip rates using different explanatory variables are included. The first one shown in figure 4.2 is the effect of employment status to the trip rates for the chosen trip group which are the long distance (over 100 km distance) leisure trips. Higher income tends to lead to a higher trip rate as well. On the other hand those working part time had higher trip rate (15.23) compared to those laid off from work (10.35 and 6.31 respectively for laid off part time and full time workers).

An increasing effect on long distance leisure trips can also be seen when comparing the trip rates for different income groups in figure 4.3. Those with the income of less than 20 000 euros in a year made on average 9.2 trips in a year compared to 20 trips per year for those with earning over 100 000 euros a year. It is worth noting that income and employment are likely to be somewhat correlated in this case as most of the time working tends to lead to a higher income as well.

One of the background explanatory variables of this work is the YKR classification which divides different zones into 7 different categories starting from inner urban city areas and leading all the way to sparsely populated rural areas. Differences were found from these trip rates as well. The 7 classes were merged into 3 new categories consisting of rural, sparse rural and urban areas. Those living in the

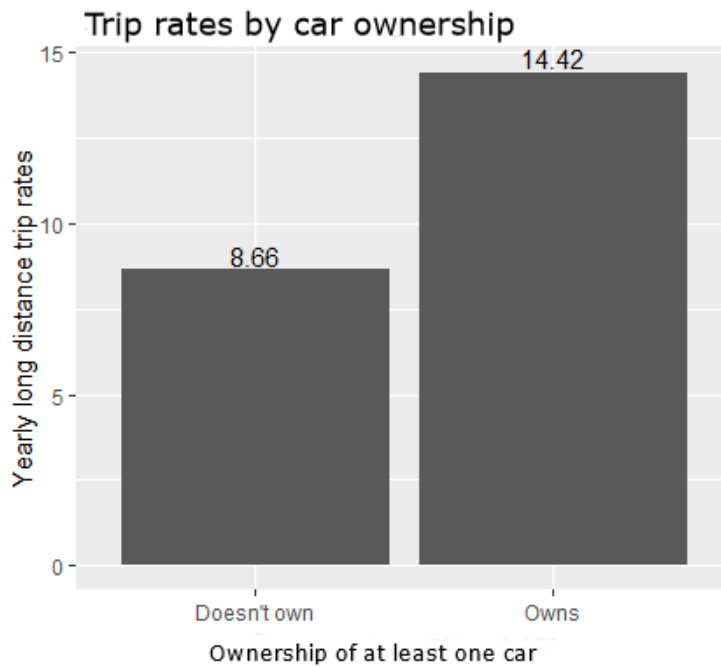


**Figure 4.4.** Trip rates by YKR areal classification, with the 7 groups having been merged into 3 groups of rural, sparse rural and urban areas.

sparsely populated rural areas made the most trips on average with a trip rate of 15.39 trips per year whereas those living in the most populated three city categories made the second most travelling group with trip rates around 13.81 approximately what is illustrated in figure 4.4. The middle population rural area categories made the least trips. This could be possibly explained by those living in the most sparse areas living furthest away from leisure locations and having an immediate need to travel to reach such destinations when making leisure trips. On the other hand those living in the most populated areas could be thought to have more economic capabilities to make longer trips compared to those living further away from city centres but yet having a fair amount of leisure trip destinations in the immediate closure to their living area. Also something to take into consideration is that YKR areal classification is likely to correlate with different measures of accessibility even if it is based mostly on population density. City areas with the highest density tend to have trip targets closer by them. Those owning at least one car made 14.42 trips on average compared to 8.66 trips for those who do not own a car as can be seen in figure 4.5.

When studying the effect of having a driving license to long distance leisure trips, it is perhaps unsurprising that possessing one tends to lead to a way higher number of trips compared to not having one. Those with the license make an average of 14.7 trips in a year compared to those not having one who resulted with 6.23 trips. Those who used to have a license make only 4.29 trips per year which could likely be explained by them tending to be older people on average. It was found on other comparison and plotting that older people in general make less such trips. The driving license explanatory variable was however left out from later statistical modelling as it is believed to correlate fairly with car ownership and same goes for the age variable as significant difference was only seen comparing the oldest age

group with others which ended up other factors to be prioritized in the modelling stage. Tested mean squared error(MSE) resulted in not being improved either when testing models with those variables together with the prioritized variables.



**Figure 4.5.** Trip rates by car ownership

### 4.3 Goodness of fit measures

In order to measure the fit of the tested statistical models, different distributions and link functions and several measures of goodness of fit were used. Regarding the link function of the response variable there were log, identity, square root and inverse links available and the relevant measure is Akaike information criterion (AIC). Smaller AIC values indicate better compatibility of the link function compared to another link function with a greater AIC value. AIC value can however only be compared between different link functions, distributions and models with the variable selection remaining the same between the compared link functions and distributions in the models.

Another goodness of fit measure between models and distributions is the mean squared error measure. It calculates the mean of the squared summed residuals between the values in the data and estimated fitted values of the model fitted to the data.

Regarding testing if the distribution of the model is the correct one of the data, one can conduct a statistical test with the so called Pearson residuals. They are calculated by counting the residuals and dividing them by the standard deviation. That way when testing the Pearson residuals values with the estimated values of the model as an explanatory variable, the estimated values should not be a significant explainer as the values of Pearson residuals should not vary according to the fitted

values. If the hypothesis ends up being rejected, then it most likely indicates that the distribution itself would not be suitable for the data. By inspecting these residuals it is possible to test if the model is producing a systematic error.

#### 4.4 Results of different models and model selection

**Table 4.1.** AIC values in order to select the link function

Link function	AIC value	Model
Log	27200.7	Poisson
Identity	27201.03	Poisson
Square root	27200.88	Poisson
Inverse	27200.25	Poisson

\*Poisson regression model was used for choosing the link function and only sustainable accessibility (with car accessibility excluded) was in the model as an explanatory variable as it is the only one that has a continuous value. For categorical explanatory variables the link function does not play as important role.

**Table 4.2.** MSE values for link function comparison

Model	MSE	Distribution
Log	2.134477	Poisson
Identity	2.134514	Poisson
Square root	2.134497	Poisson
Inverse	2.134425	Poisson

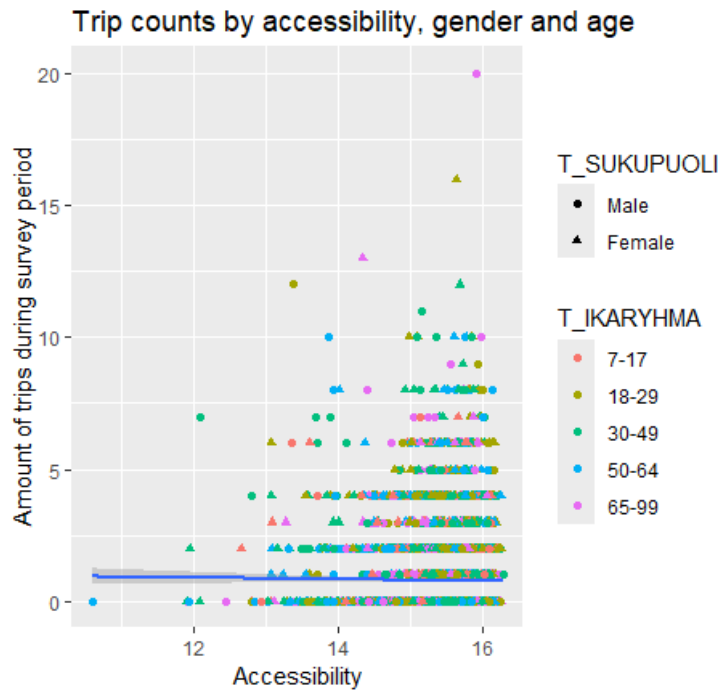
\*Poisson regression model was used for choosing the link function and only accessibility was in the model as an explanatory variable as it is the only one that has a continuous value. For categorical explanatory variables the link function does not play as important role.

For data modelling, R was used as the statistical software because of its features in modelling when it comes to libraries for statistical models and also because of its extensive plotting capabilities. As the data is in count data form regarding the number of trips as a response variable, Poisson regression models were the primary option for this case. Quasi-Poisson and Negative binomial regression models differ from Poisson models in terms of the variance structure as Poisson models assume equal mean and variance which is often not the case in practical situations.

Because of the nature of the data being count type data, it was possible to test different link functions with log, identity, square root and inverse link being the alternatives. For categorical explanatory variables the link function is not as important so the choice was made while having the sustainable accessibility(excluding car accessibility) continuous variable as the only explanatory variable in the models. AIC values found in table 4.1 resulted in inverse link having the smallest value so it was chosen to be the link function as long as Poisson models were concerned.

Different Poisson models were tested with the sets of variables and because of the seemingly large amount of zero values in the data, also zero-inflated and stop-go hurdle model were tested with the full models with other explanatory variables included as well.

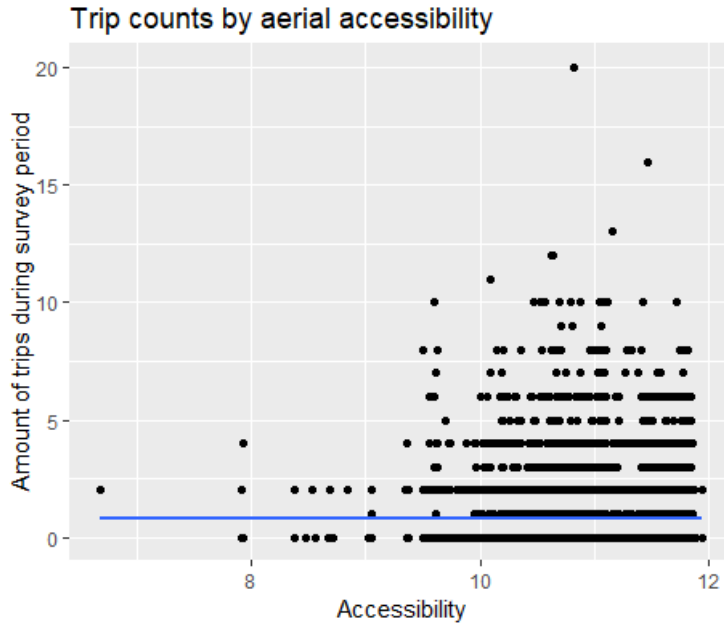
As can be seen from table 4.2, the tested inverse link also gives the best fit when comparing the mean squared errors of the link function options.



**Figure 4.6.** Trips by sustainable accessibility (which does not include car accessibility), gender and age. The blue plot line corresponds to a fitted linear regression line of trips by sustainable accessibility with all groups and genders included.

The accessibility variables were included in the tested models as explanators, with all the combined accessibility value as well as aerial, long distance bus and train accessibility values separately and as a final choice, the sustainable accessibility (plotted in figure 4.6). The result was however that any of the variables were not significant in the tested models either alone or together with other variables. The aerial accessibility (plot in figure 4.7) produced relatively closely significant p-value (0.057) for increased effect on trips that value being slightly over 0.05 in the hurdle model zero producing process. Of the tested models, that was the most significant value found for aerial accessibility.

The next step was to compare more models by MSE, which was calculated for Poisson, Quasi-Poisson and negative binomial models with the chosen inverse link function but also add zero-inflated and hurdle models which were tested with both Poisson distribution and negative binomial distribution choices, the link functions being logit link for the zero generating process and log link for the count process of the respective models. The results indicated the zero-inflated model with negative binomial distribution to produce lowest MSE as can be seen from table 4.3.



**Figure 4.7.** Trips by airplane accessibility. The blue plot line corresponds to a fitted linear regression line of trips by airplane accessibility.

**Table 4.3.** MSE values for model and distribution comparison

Model and distribution	MSE	Link
Poisson	2.09786	Inverse
Quasi-Poisson	2.09786	Inverse
Negative binomial	2.099665	Inverse
Zero-Inflated (Poisson distribution)	2.096284	Logit/Log
Zero-Inflated (Negative binomial distribution)	2.096235	Logit/Log
Hurdle (Poisson distribution)	2.096284	Logit/Log
Hurdle (Negative binomial distribution)	2.096304	Logit/Log

\*Zero-inflated models resulted in lowest MSE-values in comparison of the full models. Link function was decided according to previous AIC value comparisons.

## 4.5 Testing of the final model candidates

In the next phase, it was tested with Pearson residuals if the prediction coefficient between the predicted value and the residual deviates from zero in order to find out how well the models and distributions fit the data and if there is systematic error found in them as explained in chapter 4.3. Results indicate the models to be quite well fit to the data generally with the exception of using the negative binomial model with the chosen inverse link where the p-value is as low as 0.000147 indicating that the prediction coefficient between the predicted value and the residual deviates from zero implying wrong choice of model and possible systematic error with it in regards to the data. Poisson and Quasi-Poisson models result in highest p-values of 0.537 for both, as shown in table 4.4.

**Table 4.4.** Test of Pearson’s residuals for fitness/systematic error of distribution

Model	P-value	Link
Poisson	0.537	Inverse
Quasi-Poisson	0.537	Inverse
Negative binomial	0.000147	Inverse
Zero-Inflated (Poisson distribution)	0.208	Logit/Log
Zero-Inflated (Negative binomial distribution)	0.25	Logit/Log
Hurdle (Poisson distribution)	0.219	Logit/Log
Hurdle (Negative binomial distribution)	0.27	Logit/Log

\*Only Negative Binomial model resulted in a low p-value indicating bad fit of distribution and possible systematic error in the model.

**Table 4.5.** AIC comparison of the final model candidates

Model	AIC	Link
Poisson	24147.69	Inverse
Quasi-Poisson	NA	Inverse
Negative binomial	19352.18	Inverse
Zero-Inflated (Poisson distribution)	18439.85	Logit/Log
Zero-Inflated (Negative binomial distribution)	18405.27	Logit/Log
Hurdle (Poisson distribution)	18440.09	Logit/Log
Hurdle (Negative binomial distribution)	18405.76	Logit/Log

\*Zero-Inflated and Hurdle model had lowest AIC-values in comparison of the full models.

Also AIC-values were calculated for the same model candidates with results in table 4.5 showing zero-inflated model and hurdle model both with negative binomial distribution resulting in lowest AIC-values of 18405.27 and 18405.76 correspondingly. AIC is not defined in the same way for Quasi-Poisson models which is the reason why its value was neglected in this case.

After the previous tests and goodness of fit measures, the final model candidates were chosen to be measured more precisely by calculating MSE-values again but this time with using 10-fold crossvalidation and splitting the data into training data consisting 90% of the data and leaving the remaining 10% to be used as test data. With the created R crossvalidation function, MSE was calculated for ten different splits for each model and then the average MSE of those splits for each final model candidate was compared. Zero-inflated model with negative binomial distribution was once again producing the best fit with hurdle models showing relatively good fit as well, results being visible in table 4.6 and log-likelihood values for the final candidate zero-inflated and hurdle model variations in table 4.7.

**Table 4.6.** 10-fold cross validation mean of MSE values for model and distribution comparison

Model	10-fold CV mean of MSE	Link
Poisson	2.097647	Inverse
Quasi-Poisson	2.097647	Inverse
Negative binomial	2.099468	Inverse
Zero-Inflated (Poisson distribution)	2.096082	Logit/Log
Zero-Inflated (Negative binomial distribution)	2.096033	Logit/Log
Hurdle (Poisson distribution)	2.096083	Logit/Log
Hurdle (Negative binomial distribution)	2.096103	Logit/Log

\*Zero-Inflated model had lowest mean of MSE when calculated with 10-fold crossvalidation.

**Table 4.7.** Log-likelihood comparison for final candidates of zero-inflated and hurdle models

Model	Log-likelihood	Link
Zero-Inflated (Poisson distribution)	-9190	Logit/Log
Zero-Inflated (Negative binomial distribution)	-9172	Logit/Log
Hurdle (Poisson distribution)	-9190	Logit/Log
Hurdle (Negative binomial distribution)	-9172	Logit/Log

\*Both zero-inflated and hurdle model with negative binomial distribution resulted in highest log-likelihood values.

## 5 Results and conclusions

**Table 5.1.** Explanations of variable names.

Explanatory variable	Explanation	Data type
hb_leisure_long	Accessibility (excluding car accessibility)	continuous
T_OMISTAA_AUTON	Car ownership	binary
income_group_s t a t	Income group	categorical
JOB_STATUS	Employment status	categorical
CITY_RURAL_SPARSE	Merged YKR area group	categorical

\*The default reference group is a full time working person living in city YKR class merged area earning 20000€-40000€/year and owning a car.

**Table 5.2.** Zero-inflated model and Hurdle model risk ratios of zero process.

Explanatory variable	Zero inflated model	Hurdle model
(Intercept)	1.43	1.03
Accessibility (excluding car accessibility)	0.93	0.94
Not owning a car	0.78	0.77
Income group 0€+	0.90	0.88
Income group 40000€+	1.47	1.49
Income group 60000€+	1.48	1.48
Income group 80000€+	1.65	1.72
Income group 100000€+	2.07	2.07
Part time worker	1.29	1.28
Occasional worker	1.36	1.41
Not working	0.97	1.00
Rural merged YKR area group	0.68	0.71
Sparse rural merged YKR area group	0.86	0.92

\*The table shows risk ratios for not producing zero trips with reference group being a full time working person living in city YKR class merged area earning 20000€-40000€/year and owning a car.

In this chapter the two final model alternatives get examined further after the goodness of fit tests. The two chosen models were the zero-inflated model with negative binomial distribution and hurdle stop-go model with negative binomial distribution as well. The models were run in R with the sustainable accessibility (excluding car accessibility), car ownership, income group, working status and merged YKR living area used as explanatory variables. These variables are presented in table 5.1. As discussed earlier in section 4.1, some of the explanatory variables have been merged into smaller amount of groups in them. Unemployed, not in the

working force(eg. studying), and both laid off from work categories were merged into one value called "Not working". Number of cars variable has been transformed into binary variable of owning or not owning a car and the 7 YKR areas have been merged into 3 larger categories consisting of city, rural and sparse rural values.

The two final model summaries are zero inflated negative binomial model in (APPENDIX A) and stop-go hurdle negative binomial model (APPENDIX B) from R including the model, parameters and p-values for respective explanatory variables and groups in them.

As both of the chosen models include both zero producing and count producing processes, in the following tables the parameters from the models have been scaled into odds for the zero part and into percentual change in count amounts. The reference group has been chosen to be a persons earning between 20000€-40000€ in a year, owning a car, living in a city area and working full time. The decisions of the reference group were made according to the median earning level in Finland as well as with assumptions of being able to choose a group representing an average case as well as possible. For instance the income group being between 20000€-40000€ could be seen to have a fair number of both car owners and non car owners in it as well as different living area classifications for the persons as well as different employment status cases.

**Table 5.3.** Zero inflated model and Hurdle model probabilities of making at least one trip in the zero process of the final models.

Explanatory variable	Zero inflated model	Hurdle model
(Intercept)	58.9%	50.7%
Accessibility (excluding car accessibility)	48.2%	48.4%
Not owning a car	43.8%	43.7%
Income group 0€+	47.4%	47%
Income group 40000€+	59.5%	59.8%
Income group 60000€+	59.6%	59.7%
Income group 80000€+	62.3%	63.5%
Income group 100000€+	67.4%	67.4%
Part time worker	56.2%	56.2%
Occasional worker	57.6%	58.6%
Not working	49.2%	50%
Rural merged YKR area group	40.5%	41.6%
Sparse rural merged YKR area group	46.2%	47.8%

\*The table shows probabilities for not producing zero trips with reference group being a full time working person living in city YKR class merged area earning 20000€-40000€/year and owning a car.

In table 5.2, the intercept is interpreted as the odds for probability of making at least one trip against the probability of not making a trip, being a division between the two probabilities. The changes in odds of making at least one trip against not doing one between the groups are increased by the times of the coefficient of the

explanatory variable or its sub-group while the other factors remain the ones chosen in the reference group.

Table 5.3 includes the probabilities of making at least one trip in the zero process of the two final statistical models. The intercept is showing the probability concerning the reference group and the other explanatory variable group values show the probability of making at least one trip when the respective explanatory group value changes while the other reference explanatory variable values remain the same. From the model that performed the best, the zero inflated model, we can see that not owning a car decreases the probability of making at least one trip significantly by around 15% whereas the income group starts to have an increase effect of almost 4% in the income group of those earning over 80 000 euros and 8.5% increase when income group is increased to those with earning of over 100 000 euros in a year. On the other hand, also the effect of belonging to the lowest income group of earning less than 20 000 euros per year decreases the probability of making at least one trip by 11.5%.

**Table 5.4.** Expected number of trips of zero-inflated model and hurdle model and effective change in it by explanatory variable/group by the amount of multiplication of the coefficient.

Explanatory variable	Zero inflated model	Hurdle model
(Intercept)	1.88	1.91
Accessibility (excluding car accessibility)	1.00	1.00
Not owning a car	0.94	0.93
Income group 0€+	0.93	0.93
Income group 40000€+	1.10	1.10
Income group 60000€+	1.08	1.08
Income group 80000€+	1.20	1.20
Income group 100000€+	1.14	1.14
Part time worker	1.05	1.05
Occasional worker	1.16	1.16
Not working	1.08	1.08
Rural merged YKR area group	1.04	1.04
Sparse rural merged YKR area group	1.16	1.15

\*The table shows expected value for trips count with reference group being a full time working person living in city YKR class merged area earning 20000€-40000€/year and owning a car. The coefficients for explanatory variables show the multiplication of probability regarding the explanatory variable or its group.

Regarding this in the probability of making long distance leisure trips there is small variation between different working groups but not working(those unemployed, laid off from work and outside of working force) decreases the probability by almost 10%. Even larger difference can be found for those who live in the rural areas, where the decrease in the reference group for making a trip goes down by over 18% and for those living in sparse rural areas the decrease is almost 13%.

The interpretation for the count part of the models is slightly different. In tables 5.4 and 5.5 the intercept states the expected number of trips for the reference group as it is the average number for the group. The change in the coefficients in the tables then describe the percentual change in the expected number of long distance leisure trips with a length over 100 kilometres. The trips count is increased or decreased by multiplication of the coefficient in table 5.4 regarding the change in variable or the group in the variable. For continuous accessibility this refers to a one-unit change correspondingly.

**Table 5.5.** Expected number of trips for zero-inflated model and hurdle model and percentage changes in them by explanatory variables/groups. These results have been derived from the coefficients of the effects of the explanatory variables shown in table 5.4.

Explanatory variable	Zero inflated model	Hurdle model
(Intercept)	1.88	1.91
Accessibility (excluding car accessibility)	+0.49%	+0.38%
Not owning a car	-6.13%	-6.56%
Income group 0€+	-6.53%	-7.02%
Income group 40000€+	+9.64%	+9.63%
Income group 60000€+	+7.76%	+7.80%
Income group 80000€+	+20.28%	+20.20%
Income group 100000€+	+13.91%	+13.99%
Part time worker	+4.99%	+5.31%
Occasional worker	+15.79%	+15.96%
Not working	+7.98%	+8.08%
Rural merged YKR area group	+4.45%	+4.33%
Sparse rural merged YKR area group	+15.60%	+15.46%

\*The table shows the expected number of trips with reference group being a full time working person living in city YKR class merged area earning 20000€-40000€/year and owning a car. Percentual change is reported for different values of other explanatory variables or their groups.

As can be seen from table 5.5, the count process of the zero-inflated model the change in expected amount of trips for those in the reference group without a car decreases by over 6% as is the case for those in the lowest income group with earning of less than 20 000 euros in a year. Belonging to a higher income group increased the trip count by 12.9% on average. Findings also differ from the previously reported zero process by the amount of trips within different worker groups, where it was found that part time working increases number of trips by 5% and occasional working by over 15% inside the reference group compared to full time workers. Areal differences are also found where for those living in sparse rural areas the expected amount of trips rises by almost 16% and for those living in other rural areas by over 4% compared to those living in city areas within the reference group.

## **5.1 Implications and conclusions**

After modelling the data statistically it is time to conclude implications regarding the original research question of how does the effect of accessibility affect trip rates. From the resulting tables we can see that especially income level has a significant contribution to the number of leisure trips of over 100 kilometres conducted for both making such trips at all and also for the expected number of them. However, working part time or occasionally also seems to increase the amount compared to the reference group of full time workers. These findings would indicate that the economic resources contribute largely but from the second finding it could also be stated that the amount of time available seems to have a significant effect as well, as the part time and occasional workers could be thought to have more timely resources in hand compared to the full time workers.

Another interesting finding is that the area group of sparse rural areas seems to make the most trips of the three area groups as was seen in the explorative analysis and also in the chosen statistical models it is indicated that the expected count is indeed the highest even while the odds of making at least one trip are slightly lower in the zero producing processes of the models compared to the city areas. This could be explained by the distances from the sparse rural areas, as it could be assumed that in the more populated areas there are a plenty of options closer by for leisure trips without the need to travel such a long distance on usual basis. On the other hand from the sparse rural areas, many destinations are often located further away.

After testing several different models, distributions and link functions it seems as if accessibility would not have effect on the the target group trips of this study, long distance leisure trips of length over 100 kilometres. The other factors such as income level, car ownership and employment status seem to have a higher effect on trip production. However, some effect was seen when inspecting the effect of aerial trip accessibility in the zero inflated model as a lone explanator but even in that case the variable ended up not being significant. The effect of income, car ownership and employment status are not exactly surprising results. What also needs to be taken into account in this study is that the accessibility map and YKR classification map differ fairly as YKR is a pure areal classification with population levels whereas accessibility is defined so that the areas with the best accessibility are placed increasingly around closer to central located areas in Finland as the limit for the trip distance is set to 100 kilometres. Those areas with the highest accessibility do not exactly correspond to e.g the capital areas of Finland of the YKR classification.

## **5.2 Final words**

This thesis was one approach to try to study the effect of accessibility and other underlying factors to trip production with different available accessibility measures tested. The insignificance of accessibility in the results might seem counterintuitive at first but when looking at the previously discussed underlying implications and conclusions it could be thought that for instance the areal factors might contribute to the effect of the accessibility measure in this work as well. As stated by Ortuzar

(2011), it has often been the case that in aggregate urban modelling applications the results have not been as expected with the estimated parameters for accessibility variable having resulted in being non-significant or of the wrong sign. A specific challenge in this work is the definition of trips over 100 kilometres which already limits the direction of a trip alone and may lead to problematic distortions in models. It is however worth noting that concerning aerial accessibility, it was close to being a significant factor regarding making a trip at all.

For future studies it could be recommended to further look at the measure of accessibility and perhaps find ways to include more contributing variables to it. Another suggestion to build on this study could be examining different types of trips as an alternative to long distance leisure trips. To conclude, for instance the distance limit of 100 kilometres could be altered or the travel time could be used as a limit instead and their connection with accessibility would be an interesting topic to research and explore further.

# References

- Daly, A. and S. Miller (2006). “Advances in modelling traffic generation”. In: *European Transport Conference (ETC)*.
- Emmert-Streib, F. (2023). “Computational Diagnostics of Data Data.ML-390 Week 4”. In: *Lecture notes, Tampere University*, pp. 60–73.
- Isotalo, J. (2023). “Count Data Models”. In: *Lecture notes, Tampere University*, pp. 4–10.
- Jahanshahi, K. et al. (2009). “Understanding travel behaviour and factors affecting trip rates”. In: *Association for European Transport*.
- (2017). “How can we model travel frequency? A critical review of current practice”. In: *Conference Paper 49*, pp. 29–33.
- Kaupunki-maaseutu-luokitus (YKR)* (2018). URL: <https://ckan.ymparisto.fi/dataset/kaupunki-maaseutu-luokitus-ykr>.
- Krasic, D. and L. Novacko (2015). “The Impact of Public Transport Network Accessibility on Trip Generation Model”. In: *Traffic Planning Preliminary Communication*.
- Kristoffersson, I., S. Berglund, and S. Algers (2020). “Estimation of a large-scale tour generation model taking travellers’ daily tour patterns into account”. In: *Transportation Planning and Technology*.
- Kristoffersson, I., A. Daly, and R. Algers (2017). “Modelling the attraction of shopping centres”. In: *IDEAS Working Paper Series from RePEc*.
- Miller, E. (2020). “Measuring Accessibility: Methods and Issues: Discussion Paper”. In: *Discussion papers (International Transport Forum)*, pp. 1–22.
- Ortuzar, J. and L. Willumsen (2011). *Modelling Transport*. 4th ed. Chichester, West Sussex, United Kingdom: John Wiley & Sons.
- Pastinen, V. et al. (2016). “Henkilöliikennetutkimus 2016: Suomalaisten liikkuminen”. In: *Finnish Transport and Communications Agency Traficom*.
- (2020). “National Model System for Transport — A Study of the Prerequisites and Options for Developing a Model System”. In: *Finnish Transport and Communications Agency Traficom*.

# APPENDIX A: Zero inflated negative binomial model

```

1  summary(modelZI.negbin_final)
2
3  Call:
4  zeroinfl(formula = n_trips ~ hb_leisure_long + T_OMISTAA_AUTON + income_group_s
5          JOB_STATUS + CITY_RURAL_SPARSE, data = trip_counts_leisure, dist = "negbin"
6
7  Pearson residuals:
8      Min      1Q  Median      3Q      Max
9  -0.8165 -0.5645 -0.4683  0.5247 11.4480
10
11 Count model coefficients (negbin with log link):
12
13      Estimate Std. Error z value Pr(>z)
14 (Intercept)      0.628755  0.676999  0.929  0.35302
15 hb_leisure_long      0.004866  0.043006  0.113  0.90992
16 T_OMISTAA_AUTON0     -0.063244  0.061696 -1.025  0.30532
17 income_group_stat40000 + 0.092042  0.048378  1.903  0.05710 .
18 income_group_stat0 +    -0.067569  0.067637 -0.999  0.31780
19 income_group_stat60000 + 0.074709  0.053065  1.408  0.15917
20 income_group_stat80000 + 0.184678  0.061349  3.010  0.00261 **
21 income_group_stat100000 + 0.130273  0.061575  2.116  0.03437 *
22 income_group_statNo info -0.026184  0.070715 -0.370  0.71118
23 JOB_STATUSNo info      0.208843  0.208608  1.001  0.31677
24 JOB_STATUSNot working  0.076797  0.037814  2.031  0.04226 *
25 JOB_STATUSOC          0.146584  0.074550  1.966  0.04927 *
26 JOB_STATUSPT          0.048656  0.059603  0.816  0.41431
27 CITY_RURAL_SPARSERural 0.043523  0.039984  1.089  0.27637
28 CITY_RURAL_SPARSEsparse ru 0.144973  0.070461  2.057  0.03964 *
29 Log(theta)            2.578447  0.202120 12.757 < 2e-16 ***
30
31 Zero-inflation model coefficients (binomial with logit link):
32
33      Estimate Std. Error z value Pr(>z)
34 (Intercept)     -0.35878  1.16424 -0.308  0.75795
35 hb_leisure_long  0.07081  0.07404  0.956  0.33887
36 T_OMISTAA_AUTON0  0.24655  0.09304  2.650  0.00805 **
37 income_group_stat40000 + -0.38507  0.07950 -4.843 1.28e-06 ***
38 income_group_stat0 +  0.10427  0.10211  1.021  0.30717
39 income_group_stat60000 + -0.38909  0.08865 -4.389 1.14e-05 ***
40 income_group_stat80000 + -0.50204  0.10819 -4.641 3.48e-06 ***
41 income_group_stat100000 + -0.72802  0.11243 -6.475 9.47e-11 ***
42 income_group_statNo info  0.16650  0.10808  1.541  0.12342

```

```

41  JOB_STATUSNo info          0.46002    0.33016    1.393    0.16352
42  JOB_STATUSNot working     0.03229    0.06460    0.500    0.61717
43  JOB_STATUSOC              -0.30769    0.13329   -2.308    0.02098 *
44  JOB_STATUSPT              -0.25098    0.10405   -2.412    0.01586 *
45  CITY_RURAL_SPASERural      0.38283    0.06698    5.715    1.10e-08 ***
46  CITY_RURAL_SPASESparse ru  0.15158    0.11673    1.299    0.19409
47  ---
48  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
49
50  Theta = 13.1767
51  Number of iterations in BFGS optimization: 45
52  Log-likelihood: -9172 on 31 Df

```

## APPENDIX B: Stop-go hurdle negative binomial model

```

1  summary(hurdle_model_negbin_final)
2
3  Call:
4  hurdle(formula = n_trips ~ hb_leisure_long + T_OMISTAA_AUTON + income_group_sta
5         JOB_STATUS + CITY_RURAL_SPARSE, data = trip_counts_leisure, dist = "negbin"
6
7  Pearson residuals:
8      Min      1Q  Median      3Q      Max
9  -0.8181 -0.5646 -0.4681  0.5227 11.4743
10
11 Count model coefficients (truncated negbin with log link):
12             Estimate Std. Error z value Pr(>z)
13 (Intercept)      0.646972   0.676389   0.957  0.33882
14 hb_leisure_long    0.003765   0.042985   0.088  0.93021
15 T_OMISTAA_AUTON0 -0.067847   0.061448  -1.104  0.26953
16 income_group_stat40000 + 0.091946   0.048152   1.909  0.05620 .
17 income_group_stat0 + -0.072794   0.067461  -1.079  0.28056
18 income_group_stat60000 + 0.075117   0.053004   1.417  0.15643
19 income_group_stat80000 + 0.183984   0.061273   3.003  0.00268 **
20 income_group_stat100000 + 0.130937   0.061504   2.129  0.03326 *
21 income_group_statNo info -0.028673   0.070748  -0.405  0.68526
22 JOB_STATUSNo info    0.212345   0.207408   1.024  0.30593
23 JOB_STATUSNot working 0.077738   0.037949   2.048  0.04051 *
24 JOB_STATUSOC        0.148091   0.074108   1.998  0.04568 *
25 JOB_STATUSPT        0.051709   0.059141   0.874  0.38193
26 CITY_RURAL_SPARSERural 0.042401   0.039941   1.062  0.28842
27 CITY_RURAL_SPARSESparse ru 0.143720   0.069380   2.071  0.03831 *
28 Log(theta)         2.583799   0.202598  12.753 < 2e-16 ***
29 Zero hurdle model coefficients (binomial with logit link):
30             Estimate Std. Error z value Pr(>z)
31 (Intercept)      2.693e-02  1.065e+00   0.025  0.97983
32 hb_leisure_long  -6.468e-02  6.777e-02  -0.954  0.33989
33 T_OMISTAA_AUTON0 -2.553e-01  8.397e-02  -3.041  0.00236 **
34 income_group_stat40000 + 3.971e-01  7.216e-02   5.503  3.73e-08 ***
35 income_group_stat0 + -1.222e-01  9.294e-02  -1.315  0.18857
36 income_group_stat60000 + 3.938e-01  8.024e-02   4.908  9.22e-07 ***
37 income_group_stat80000 + 5.432e-01  9.856e-02   5.511  3.56e-08 ***
38 income_group_stat100000 + 7.270e-01  9.989e-02   7.279  3.38e-13 ***
39 income_group_statNo info -1.667e-01  9.923e-02  -1.680  0.09301 .
40 JOB_STATUSNo info  -3.712e-01  3.146e-01  -1.180  0.23799

```

```

41  JOB_STATUSNot working      3.913e-05  5.914e-02  0.001  0.99947
42  JOB_STATUSOC              3.439e-01  1.211e-01  2.839  0.00452 **
43  JOB_STATUSPT              2.496e-01  9.278e-02  2.690  0.00714 **
44  CITY_RURAL_SPARSERural    -3.407e-01  6.115e-02  -5.572  2.52e-08 ***
45  CITY_RURAL_SPARSESparse ru -8.663e-02  1.065e-01  -0.813  0.41600
46  ---
47  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
48
49  Theta: count = 13.2474
50  Number of iterations in BFGS optimization: 28
51  Log-likelihood: -9172 on 31 Df

```